Statistical Methods for the Analysis of Network Data

June 15th to 17th, 2009

ORGANIZATION

SCHEDULE

Monday, June 15th

08:00-09:00 Registration Opens

09:00-09:30 Welcome to Statistical Methods for the Analysis of Network Data

09:30-10:15 "Maximum Likelihood Estimation for Social Network Dynamics"

Tom A.B. Snijders, Michael Schweinberger and Johan Koskinen

10:15-11:00 "Strategies for Online Inference of Network Mixture"

Hugo Zanghi, Franck Picard, Vincent Miele and Christophe Ambroise

11:00-11:30 Tea/Coffee

11:30-12:15 "A State-Space Mixed Membership Blockmodel for Dynamic Network Tomography"

Wenjie Fu, Le Song and Eric Xing

12:15-13:00 "Hierarchical Relational Models for Document Networks"

Jonathan Chang and David Blei

13:00-14:00 Poster Session & Lunch

14:00-14:45 "Network-based auto-probit modeling with application to protein function prediction"

Eric D. Kolaczyk

14:45-15:30 "Estimating Time-Varying Networks"

Mladen Kolar, Le Song, Amr Ahmed and Eric P. Xing

15:30-16:00 Tea/Coffee

16:00-16:45 "Modeling Networks from Sampled Data"

Mark Handcock and Krista Gile

16:45-18:00 Poster Session & Reception

Tuesday, June 16th

09:00-09:35 " Estimation from Network-Based Respondent-Driven Sampling"

Krista J. Gile

09:35-10:10 "Network Sampling And Sampled Networks"

Edoardo Airoldi, Joseph Blitzstein and Ben Olding

10:10-10:45 "Effects of network size on exponential family random graph models and their inference based on egocentrically sampled networks."

Pavel N. Krivitsky, Mark S. Handcock, Martina Morris

10:45-11:15 Tea/Coffee

11:15-11:50 " Latent Structure Models for Social Networks using Aggregated Relational Data"

Tyler H. McCormick and Tian Zheng

11:50-12:25 "A Mixture of Experts Latent Postion Cluster Model"

Isobel Claire Gormley and Thomas Brendan Murphy

12:25-13:00 "Variational Bayesian Inference for the Latent Position Cluster Model"

Michael Salter-Townshend and Thomas Brendan Murphy

13:00-14:00 Lunch

14:00-14:35 Hierarchical Extensions of Exponentially-Parameterized Random Graph Models

Michael Schweinberger and Mark S. Handcock

14:35-15:10 Bayesian Inference for the p^* model

Alberto Caimo and Nial Friel

15:10-15:45 The Exchangable Graph Model

Edoardo M. Airoldi

Late Afternoon Tourist activity.

20:00- Dinner in Fallon & Byrne

Wednesday, June 17th

09:00-09:35 “Characterising Gene Co-expression Networks Via Graphical Models”

Gabriel Coelho Gonçalves de Abreu, Rodrigo Labouriau, David Edwards

09:35-10:10 “Quantifying and comparing complexity of cellular networks: structure beyond degree statistics”

Alessia Annibale and Anthony Coolen

10:10-10:45 “Node and link roles in protein-protein interaction networks”

Sumeet Agarwal, Charlotte Deane, Nick Jones and Mason Porter

10:45-11:15 Tea/Coffee

11:15-11:50 “Approximate Bayesian Computation under model uncertainty with an application to network data”

Oliver Ratmann, Sylvia Richardson, Carsten Wiuf

11:50-12:25 “Analogical Reasoning in Information Retrieval”

Ricardo Silva, Edoardo Airoldi, Katherine Heller, Zoubin Ghahramani

12:25-13:00 “Statistical Analysis of Evolving Networks”

Stephen E. Fienberg

13:00-14:00 Lunch

Afternoon Informal tourist event.

Posters

“The Effect of Random Social Interactions and Dynamic Network Structures on Product Adoption”

Declan Mungovan, Enda Howley and Jim Duggan

“Using Distinct Aspects of Social Network Analysis to Improve the Customer Loyalty Process”

Carlos Andre Reis Pinheiro and Markus Helfert

“Translating evidence into practice: a shared priority in public health?”

H. McAneney, J.F. McCann L. Prior, J. Wilde and F. Kee

“An inferential approach for one-sample gene network comparisons: assessing dissimilarity via a resampled local covering metric”

Phillip Yates and Nitai Mukhopadhyay

“Analysis of a web-based network of educators”

April Galyardt, Turadg Aleahmad, Stephen Fienberg and Brian Junker

“Inferences for Differential Transmission Rates of an Infectious Disease on a Network”

Rachel Schutt

“Uncertainties in network analysis due to the thresholding problem”

Andrew Thomas

“Sensitivity of Spectral Clustering. Similarities between Spectral Clustering and Multidimensional Scaling”

Zhivko Stoyanov

Program Committee

Edoardo Airoldi, Harvard University
Stephen Fienberg, Carnegie Mellon University
Nial Friel, University College Dublin
Claire Gormley, University College Dublin
Mark Handcock, University of Washington
Peter Hoff, University of Washington
Brendan Murphy, University College Dublin
Cosma Shalizi, Carnegie Mellon University
Tom Snijders, University of Oxford and University of Groningen

Local Organizing Committee

Nial Friel, University College Dublin
Claire Gormley, University College Dublin
Brendan Murphy, University College Dublin

Contents

1 Maximum Likelihood Estimation for Social Network Dynamics

Tom A.B. Snijders, Michael Schweinberger and Johan Koskinen A model for network panel data is discussed, based on the assumption that the observed data are discrete observations of a continuous-time Markov process on the space of all digraphs on a given node set, in which changes in tie variables are independent conditional on the current digraph. The model for tie changes is parametric and designed for applications to social network analysis, where the network dynamics can be interpreted as being generated by choices made by the social actors who are represented by the nodes of the digraph. An algorithm for calculating the Maximum Likelihood estimator is presented, based on data augmentation and stochastic approximation. An application to an evolving friendship network is given and a small simulation study is presented which suggests that the Maximum Likelihood estimator is more efficient than an earlier proposed Method of Moments estimator for small data sets.

2 Strategies for Online Inference of Network Mixture

Hugo Zanghi, Franck Picard, Vincent Miele and Christophe Ambroise

The statistical analysis of complex networks is a challenging task, given that appropriate statistical models and efficient computational procedures are required in order for structures to be learned. One line of research has aimed at developing mixture models for random graphs, and this strategy has been successful in revealing structures in social and biological networks. The principle of these models is to assume that the distribution of the edge values follows a parametric distribution, conditionally on a latent structure which is used to detect connectivity patterns. However, these methods suffer from relatively slow estimation procedures, since dependencies are complex and do not necessarily allow for computational simplifications. In this paper we adapt online estimation strategies, originally developed for the EM algorithm, to the case of models for which the probability of the missing data conditionally on the available observations is not tractable. Our work focuses on two methods, the first based on the SAEM algorithm, and the second on variational methods. We perform a simulation to compare these two algorithms with existing approaches, and we use the method to decipher the structure of the US political websites network. We show that our online EM-based algorithms offer a good trade-off between precision and speed, when estimating parameters for mixture distributions in the context of random graphs.

3 A State-Space Mixed Membership Blockmodel for Dynamic Network Tomography

Wenjie Fu, Le Song and Eric Xing

School of Computer Science, Carnegie Mellon University

In a dynamic social or biological environment, the interactions between the underlying actors can undergo large and systematic changes. The latent roles or membership of the actors as determined by these dynamic links will also exhibit rich temporal phenomena, assuming a distinct role at one point while leaning more towards a second role at an another point. To capture this dynamic mixed membership in rewiring networks, we propose a state space mixed membership stochastic blockmodel which embeds an actor into a latent space and track its mixed membership in the latent space across time. We derived efficient approximate learning and inference algorithms for our model, and applied the learned models to analyze a social network between monks, and a rewiring gene interaction network of Drosophila melanogaster collected during its full life cycle. In both cases, our model reveals interesting patterns of the dynamic roles of the actors.

4 Hierarchical Relational Models for Document Networks

Jonathan Chang and David Blei

Princeton University

We develop the relational topic model (RTM), a hierarchical model of both network structure and node attributes. We focus on document networks, where the attributes of each document are its words, i.e., discrete observations taken from a fixed vocabulary. For each pair of documents, the RTM models their link as a binary random variable that is conditioned on their contents. The model can be used to summarize a network of documents, predict links between them, and predict words within them. We derive efficient inference and estimation algorithms based on variational methods and evaluate the predictive performance of the RTM for large networks of scientific abstracts, web documents, and geographically tagged news.

5 Network-based auto-probit modeling with application to protein function prediction

Eric D. Kolaczyk

Boston University

Predicting the functional roles of proteins based on various genome-wide data, such as protein-protein association networks, has become a canonical problem in computational biology. Approaching this task as a binary classification problem, we develop a network-based extension of the spatial auto-probit model. In particular, we develop a hierarchical Bayesian probit-based framework for modeling binary network-indexed processes, with a latent multivariate conditional autoregressive (CAR) Gaussian process. The latter allows for the easy incorporation of protein-protein association network topologies – either binary or weighted – in modeling protein functional similarity. We use this framework to predict protein functions, for functions defined as terms in the Gene Ontology (GO) database, a popular rigorous vocabulary for biological functionality. Furthermore, we show how a natural extension of this framework can be used to model and correct for the high percentage of false negative labels in training data derived from GO, a serious short-coming endemic to biological databases of this type that nevertheless appears to have received little attention in the literature to date. Our method performance is evaluated and compared with standard algorithms on weighted yeast protein-protein association networks, extracted from a recently developed integrative database called STRING. Results show that our basic method is competitive with or better than these other methods, and that the extended method – incorporating the uncertainty in negative labels among the training data – can yield truly substantial improvements in predictive accuracy.

6 Estimating Time-Varying Networks

Mladen Kolar, Le Song, Amr Ahmed and Eric P. Xing

School of Computer Science, Carnegie Mellon University

Stochastic networks are a plausible representation of the relational information among entities in dynamic systems such as living cells or social communities. While there is a rich literature in estimating a static or temporally invariant network from observation data, little has been done towards estimating time-varying networks from time series of entity attributes. In this paper, we present two new machine learning methods for estimating time-varying networks, which both build on a temporally smoothed l_1-regularized logistic regression formalism that can be cast as standard convex-optimization problem and solved efficiently using generic solvers scalable to large networks. We report promising results on recovering simulated time-varying networks. For real datasets, we reverse engineer the latent sequence of temporally rewiring political network between Senators from the US senate voting records and the latent evolving gene network which contains more than 4000 genes from the life cycle of *Drosophila melanogaster* from microarray time course.

7 Modeling Social Networks from Sampled Data

Mark S. Handcock and Krista J. Gile

University of Washington

Network models are widely used to represent relational information among interacting units and the structural implications of these relations. Recently, social network studies have focused a great deal of attention on random graph models of networks whose nodes represent individual social actors and whose edges represent a specified relationship between the actors.

Most inference for social network models assumes that the presence or absence of all possible links is observed, that the information is completely reliable, and that there are no measurement (e.g. recording) errors. This is clearly not true in practice, as much network data is collected though sample surveys. In addition even if a census of a population is attempted, individuals and links between individuals are missed (i.e., do not appear in the recorded data).

In this paper we develop the conceptual and computational theory for inference based on sampled network information. We first review forms of network sampling designs used in practice. We consider inference from the likelihood framework, and develop a typology of network data that reflects their treatment within this frame. We then develop inference for social network models based on information from adaptive network designs.

We motivate and illustrate these ideas by analyzing the effect of link-tracing sampling designs on a collaboration network.

8 Estimation from Network-Based Respondent-Driven Sampling

Krista J. Gile

Respondent-Driven Sampling is a widely used variant of link-tracing sampling on a network, designed to allow for estimation in hard-to-reach populations. It is commonly used to estimate the prevalence of diseases such as HIV in high-risk populations such as injection drug users. Beginning with a researcher-selected convenience sample, each person sampled is given a small number of uniquely identified coupons to distribute to other members of the target population, making them eligible for enrolment in the study. This strategy is highly effective at collecting large diverse samples from many hard-to-reach populations.

Current estimation relies on sampling weights estimated by treating the sampling process as a random walk on a graph, where the graph is the social network of relations among members of the target population. These estimates are based on strong assumptions allowing the sample to be treated as a probability sample. In particular, the current estimator assumes a with-replacement sample or small sample fraction, while in practice samples are without-replacement, and often include a large fraction of the population. A large sample fraction, combined with different mean nodal degrees for infected and uninfected population members, induces substantial bias in the estimates. We introduce a new estimator which accounts for the without-replacement nature of the sampling process, and removes this bias. We then briefly introduce a further extension which uses a parametric model for the underlying social network to reduce the bias induced by the initial convenience sample.

9 Network Sampling And Sampled Networks

Edoardo Airoldi, Joseph Blitzstein and Ben Olding

Harvard University

The recent explosion of interest in network data has required the development of new models such as exponential random graph models, stochastic blockmodels, and latent space models. Yet it is typically impossible to observe the full network, which necessitates mechanisms for sampling within the network.

Various network sampling schemes have been proposed and implemented, often based on link-tracing (sampling some nodes, then some neighbors, then some neighbors neighbors, etc.). Most of this work has been from a design-based perspective, where the network itself is not modeled stochastically and all uncertainty stems from the sampling itself. But when will the generative mode mesh well with the sampling scheme? Handcock and Gile [3] give conditions under which ignorability holds, in which case inference can be based on the face value likelihood without worrying about the sampling. These conditions often do not hold in practice. To explore the impact of sampling schemes on the likelihood, estimation, and inference we simulate various sampling schemes together with models for the network itself, focusing on widely used exponential random graph (ERGM) models. Our main sampling schemes of interest are snowball sampling [2] (which we define more broadly to include common link-tracing schemes such as respondent-driven sampling [4] and neighborhood sampling), simple random sampling (of nodes or of edges), and shortest path sampling [1] (also known as traceroute sampling, and widely used in the study of the Internet). For fair comparisons, we control for the number of nodes in each sample.

To study potential bias-variance tradeoffs obtained from the various possible sampling schemes, we simulate random ERGM graphs and regress statistics of interest on the corresponding statistics for subsampled graphs. For the case of snowball sampling, we can tune the breadth (number of neighbors taken at each stage) and assess whether there is a tradeoff with depth (how far into the network the sample pervades) with a fixed sample size. In Figure 1 is shown an example with the 3-star statistic is shown with the number of starting points fixed at 3, and various values of the breadth k. Here it turns out that there is no tradeoff: larger width turns out to improve both the variance and the bias.

References

[1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. Proceedings of the thirty-seventh annual ACM symposium on Theory of computing, pages 694703, 2005.

[2] L. Goodman. Snowball sampling. The Annals of Mathematical Statistics, pages 148170, 1961.

[3] M. S. Handcock and K. Gile. Modeling social networks with sampled or missing data. Annals of Applied Statistics, 2009.

[4] D. D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems, 44(2):174199, 1997.

[5] M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. PNAS, 102(12):42214224, 2005.

10 Effects of network size on exponential family random graph models and their inference based on egocentrically sampled networks

Pavel N. Krivitsky, Mark S. Handcock, Martina Morris

Exponential family random graph models (ERGMs) provide a principled way to model and simulate features common in human social networks, such as propensities for homophily and friend-of-a-friend triad closure. We show that, by default, ERGMs preserve density as network size increases. This density invariance is not realistic for most social networks, and we suggest a simple modification based on the concept of an offset which instead preserves the mean degree and accommodates changes in network composition asymptotically.

We derive an approach to fitting ERGMs based on an egocentrically sampled network data set using a sampling and method of moments framework, and use it to demonstrate applicability of this approach to ERGMs with dyadic dependence.

11 Latent Structure Models for Social Networks using Aggregated Relational Data

Tyler H. McCormick and Tian Zheng

Department of Statistics, Columbia University

We propose a method to estimate structural properties, such as clustering, in networks using aggregated relational data. Questions of the form How many Xs do you know? collect aggregated relational data and are easily incorporated into standard surveys. We propose a latent space framework 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 doesnt require that the entire network be observed. Using this framework, we derive evidence of social structure in personal acquaintances networks, estimate individual and population degree distributions, and demonstrate evidence of non-random social mixing. 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.

12 A Mixture of Experts Latent Postion Cluster Model

Isobel Claire Gormley and Thomas Brendan Murphy

School of Mathematical Sciences, University College Dublin.

Network data detail the interactions between a set of nodes or actors. For example social network data represents the interactions between a group of social entities whereas gene regulatory network data details gene interactions. The latent space model for network data (Hoff et al., (2002)) locates each node in a latent space and models the probability of an interaction or link between two nodes as a function of their locations. The latent position cluster model (Handcock et al., (2007)) extends the latent space model to appropriately deal with network data in which clusters of nodes exist. In the latent position cluster model it is assumed that the location of a node in the latent space is drawn from a finite mixture model, each component of which represents a cluster of nodes.

A mixture of experts model (Jacobs et al., 1991) builds on the structure of a mixture model by taking account of both observations and associated covariates when modelling a heterogeneous population. Here, a mixture of experts extension of the latent position cluster model is developed so that covariate information can be included. The mixture of experts framework allows covariates to enter the latent position cluster model in a number of ways, yielding different model interpretations. For example, the covariates may enter the model through the mixing proportions of the finite mixture model, and/or through the link probabilities. Model selection techniques are used to select the manner in which covariates enter the mixture of experts latent position cluster model.

Estimates of the model parameters are derived in a Bayesian framework using a Markov Chain Monte Carlo algorithm. In the Metropolis-Hastings steps of the algorithm the target distributions pose sampling difficulties suitable proposal distributions which shadow the target distributions are de- rived using ideas from surrogate transfer optimization (Lange et al., (2000)). The developed methodology is demonstrated through a number of illustrative network data examples.

References

Handcock, M. S., Raftery, A. E. and Tantrum, J. M., Model-based clustering for social networks. Journal of the Royal Statistical Society, Series A. 170 (2007) 301-354.

Hoff, P. D., Raftery, A. E. and Handcock, M. S., Latent Space Approaches to Social Network Analysis. Journal of the American Statistical Association. 97 (2002) 1090-1098.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E., Adaptive mixture of local experts. Neural Computation. 3 (1991) 79-87.

Lange, K., Hunter, D. R. and Yang, I., Optimization transfer using surrogate objective functions. Journal of Computational and Graphical Statistics. 9 (2000) 1-59.

13 Variational Bayesian Inference for the Latent Position Cluster Model

Michael Salter-Townshend and Thomas Brendan Murphy

School of Mathematical Sciences, University College Dublin.

Networks are used to represent data on interactions between actors or nodes. They are employed to model a diverse range of statistical problems, from disease epidemics to social human networks.

Recent work on Bayesian analysis of the links between these actors has focussed on embedding the actors in a latent "social space". Links between actors are more likely given a closer relative position in this social space. Such a model allows explicit modelling of the clustering that is exhibited in many network datasets (see Handcock et al 2007).

However, inference via MCMC is cumbersome and scaling of these methods to large networks with many interacting nodes is a challenge. Variational Bayesian methods offer one solution to this problem. An approximate, closed form posterior is formed, with unknown variational parameters. These parameters are tuned by minimisation of the Kullback-Leibler divergence between the approximate variational posterior and the true posterior, known only up to proportionality.

The computational overhead is far less than sampling based methods and this allows for richer models to be developed. However, the variational posterior may not be a good approximation to the true posterior and the quality of the approximation is difficult to assess.

Handcock, M.S., Raftery, A.E., and Tantrum J.M. (2007) 'Model-based Clustering for Social Networks'. *Journal of the Royal Statistical Society: Series A* **170**, 2, 1–22.

14 Hierarchical Extensions of Exponentially-Parameterized Random Graph Models

Michael Schweinberger and Mark S. Handcock

Relational data, corresponding to relationships between actors and frequently represented by graphs, have been widely modeled by Exponentially-parameterized Random Graph Models (ERGMs). An important practical stumbling block for modeling observed data by ERGMs is the so-called near-degeneracy problem: ERGMs tend to place most probability mass on a small number of graphs; the graphs with non-negligible probability mass do not resemble observed data; and ERGMs are unstable in the sense that small changes in parameter values are associated with large changes in the probability mass function. The near-degeneracy problem of ERGMs has important consequences for simulating data from ERGMs and inferring from observed data to ERGMs. We propose a hierarchical approach to ERGMs with a view to alleviating the near-degeneracy problem. In the hierarchical framework, we elaborate on non-parametric Bayesian ERGMs capturing latent local structure. These models allow to uncover unobserved local structure, represent a wide range of structural features of relational data, and are simple and interpretable. We demonstrate the potential of these models by applying them to empirical data.

15 Bayesian Inference for the p^* model

Alberto Caimo and Nial Friel

School of Mathematical Sciences, University College Dublin.

This talk is concerned with Bayesian inference for the exponential random graph model or p* model. This model is widely used and studied, however it is extremely difficult to handle from a statistical viewpoint, since the normalising constant, which depends on parameters of the model, is intractable for all but trivially small networks. In this talk we show how inference for this model can be carried out using the exchange algorithm of Murray et al (2006), which circumvents the need to calculate the normalising constant. Our talk also illustrates how this approach gives improves performance with respect to the Monte Carlo maximum likelihood method of Geyer and Thompson (1992). Moreover the performance of this approach appears to be more robust to the problem of near degeneracy.

16 The Exchangable Graph Model

Edoardo M. Airoldi

Harvard University

Mapping connectivity patterns in a graph onto the space of binary strings via the simplest possible model would enable comparisons between different statistical models of pairwise measurements. Such a strategy would also leads to calculations for assessing the statistical significance associated with the observed overlap between two cliques in a graph. The *exchangeable graph model* maps a graph over N nodes to a set of node-specific binary strings, to support these two analyses in practice (Airoldi, 2006).

The data generating process instantiating an exchangeable graph model for a matrix of binary observations on pairs of N nodes is specified as follows,

1. For each node in $n \in \mathcal{N}$

   1.1. Sample node-specific binary strings $\vec{b}_n \sim Uniform$ (vertex set of K-hypercube),

2. For node pair $n, m \in \mathcal{N} \times \mathcal{N}$

   2.3. Sample the binary physical binding event $x_{nm} \sim Bernoulli\ (\ q(\vec{b}_n, \vec{b}_m)\ )$,

where $\vec{b}_{1:N}$ are binary strings K-bit long, and q is function that projects binary strings into the $[0, 1]$ interval. This generating process leads to weakly dependent edges; the edges are conditionally independent given their binary string representations, technically they are *exchangeable.* In this sense, an *exchangeable graph model* provides the minimal step-up in complexity from the random graph model (Erdös and Rényi, 1959; Gilbert, 1959).

Briefly, the number of bits captures the complexity of a graph. For instance, for $K < N$ the model provides a parsimonious representation of the graph. For directed graphs the function q is asymmetric in the arguments. The sparsity of the bit strings can be controlled with a hierarchical construction based on a distribution on the unit hypercube (Airoldi, 2009). In an exchangeable graph model there are two main sources of variability: (i) the probability of an edge decreases with the number of bits K, as more complexity reduces the chances of an edge, and (ii) the probability of an edge increases with $1/\alpha$, as concentrating density in the corners of the unit K-hypercube improves the chances of an edge. While this model does not quite fit the definition of non-homogeneous models of Bollobás et al. (2007), it is tractable enough to allow the analysis of the giant component, albeit approximately, by leveraging the branching process strategy similar to the one developed by Durrett (2006). As in Durrett's analysis, the giant component emerges because a number of smaller components must intersect with high probability. In addition, the giant component has a peculiar structure in exchangeable graph models; connected components are themselves connected to form the giant component a soon as bit-strings that match on two bits appear with high probability. For an illustration see figure 1, where nodes that *bridge* two connected components are evident in the left panel. In the Figure, there are no nodes that bridge three components, as having bit-strings that match on three bits is an unlikely in this parameter setting.

In practice, given a graph we can infer the corresponding set of binary strings from data. The likelihood that correspond to an exchangeable graph model is simple to write,

$$\ell(Y|\alpha) = \int d\ \vec{b}_{1:N}\ (\ \prod_{n,m} Pr\ (Y_{n,m}|\vec{b}_n, \vec{b}_m, q) \prod_n Pr\ (\vec{b}_n|\alpha)\ ),$$

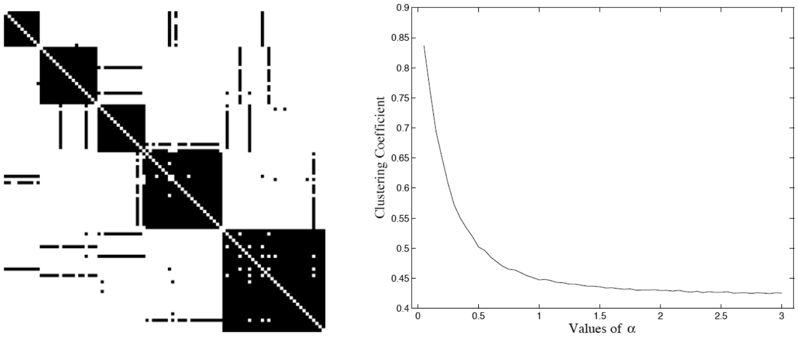

Figure 1: *Left.* An example adjacency matrix that correspond to a fully connected component among 100 nodes. *Right.* The clustering coefficient as a function of α on a sequence of graphs with 100 nodes.

and we can apply sampling or variational inference techniques (Airoldi, 2007).

The exchangeable graph model allows to assess the complexity of an observed graph leveraging notions in information theory. For instance, we can use MDL (i.e. the minimum description length principle) to decide how many bits we need to explain the observed connectivity patterns in a graph, with high probability. We can also quantify how much *information* is retained at different bit-lenghts, and plot the corresponding *information profile* for $K < N$, and an *entropy histogram* for any given value of K.

The exchangeable graph model allows comparison of any set of statistical models that are proposed to summarize an observed graph. As an illustration, consider an observed graph G and two alternative models A and B. Rather than comparing how well models A and B recover the degree distribution of G, or any other set of graph statistics, and independently of whether it makes sense of not to directly compare the two likelihoods of A and B (in fact, these models need not have a likelihood), we can proceed as follows.

1. Given a graph G, fit models $A(\Theta_a)$ and $B(\Theta_b)$ to obtain an estimate of their parameters.

2. Sample M graphs at random from the support of $A(\Theta_a^{Est})$ and $B(\Theta_b^{Est})$.

3. Compute the distributions of summary statistics based on notion from information theory, such as information profile and entropy histogram, corresponding to the $2M$ graphs sampled from A and B.

4. Compare models in terms of the distribution on the statistics above, such as the complexity of the two models' supports, the similarity between the complexity of G and the models' complexity, and so on.

Last, the exchangeable graph model allows to evaluate the distribution of the number of bit-strings with I matching bits, for any integer $I < K$. From a theoretical perspective, this distribution leads to expectations on the number of nodes that bridge I communities, where the members of each community have only one out of I matching bits. In practice, we may want to specify K in advance so that each bit corresponds to a well defined property. For instance, in applications to biology nodes may correspond to proteins and the K bits encode presence/absence of specific protein domains. The distribution on the number of I matchings leads to p-values that summarize how unexpected it is to observed binding events among a set of proteins that share a certain combination of domains.

Overall, the exchangeable graph model introduces weak dependence among the edges of a random graph in a controlled fashion, which ultimately leads to a range of more structured connectivity patterns and enables model comparison strategies rooted in notions from information theory. The focus here is not on modeling per-se. In fact, the model is kept as simple as possible. Rather, the exchangeable graph model provides a bridge between graph connectivity and node attributes to support graph model comparison and significance analysis of communities overlap.

References.

E. M. Airoldi. *Bayesian mixed membership models of complex and evolving networks.* PhD thesis, School of Computer Science, Carnegie Mellon University, December 2006.

E. M. Airoldi. Getting started in probabilistic graphical models. *PLoS Computational Biology*, 3 (12):e252, 2007.

E. M. Airoldi. A family of distributions on the unit hypercube. Technical report, Harvard university, Department of Statistics, March 2009.

B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

R. Durrett. *Random Graph Dynamics.* Cambridge University Press, 2006.

P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 5:290–297, 1959.

E. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30:1141–1144, 1959.

17 Characterising Gene Co-expression Networks Via Graphical Models

Gabriel Coelho Gonçalves de Abreu, Rodrigo Labouriau, David Edwards

Research group for Bioinformatics, Genetics, and Statistics, Department of Genetics and Biotechnology, Aarhus University

The use of network theory has gained great attention in systems biology. Typical applications are co-expression, regulatory, evolution, and protein- protein interaction networks. The analysis of gene co-expression networks is capable to assess the relationship between pairs of genes, considering the presence and inuence of other genes. In this way, one can identify genes occupying a central position in the network. Such genes (or hubs) could inu- ence the entire stability of the network. Therefore, the use of co-expression networks can give insight about the role played by certain genes in the expression profile of an organism.

We use the setup of graphical models to assess relationships between genes. The network of interest is defined as an undirected graph given by $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V} = \{v_1, \ldots, v_p\}$ is the set of p vertices (the genes), and $\mathbf{E} = \{(u, v)|u, v \in \mathbf{V} \text{ and } u \neq v\}$ is the set of edges, in which multiple edges are not allowed. Two vertices are connected by an edge if, and only if, their expression levels are conditionally correlated given the expression levels of other genes in the network. In this way, we eliminate all the spurious asso- ciations, due to makovian properties of these graphical models.

We present an efficient algorithm that finds a forest-graph with mini- mum BIC (Bayesian Information Criteria) among all possible forest-graphs. Under regularity conditions, this graph is the best possible representation of the network in terms of forest-graphs. In the same way, we obtain optimal representations in terms of triangulated graphs. An R package implementing the algorithm, and which uses a reasonable amount of computer resources, is under development.

18 Quantifying and comparing complexity of cellular networks: structure beyond degree statistics

Alessia Annibale and Anthony Coolen

Dept of Mathematics, King's College London

We show that for any observed cellular network there exists a well-defined canonical random graph ensemble that produces graphs with structural characteristics identical to (and controlled solely by) the degree statistics and the degree correlations of the given network. We construct this ensemble and study its mathematical properties. We then demonstrate how the ensemble generates various powerful quantitative tools for analysing cellular signalling networks at a macroscopic level (dependent on structural measures only, not on network size), such as (i) precise measures of network complexity, (ii) precise measures for quantifying structural distances between networks, and (iii) numerical algorithms for generating 'null models' with prescribed macroscopic structural properties.

19 Node and link roles in protein-protein interaction networks

Sumeet Agarwal, Charlotte Deane, Nick Jones, Mason Porter

University of Oxford, Oxford, United Kingdom

A key question in modern biology is how the complexity of protein-protein interaction networks relates to biological functionality. One way of understanding the set of proteins and their interactions (the interactome) is to look at them as a network of nodes connected by links. By studying the structure of this network, we may hope to learn something about the interactomes organisation. Here we attempt to look at different approaches for using network models to assign structural and functional roles to proteins and protein interactions. It has been proposed that highly connected nodes, or hubs, in the interactome fall into two classes, date and party [1], and that these play a key role in the modular organisation of the yeast interactome. This classification was made on the basis of the extent to which hubs are co-expressed with their interaction partners, but was then used to impute to them specific topological roles. We attempt to use purely topological statistics to examine the extent to which these hubs really fall into the roles thus attributed. We use a community detection approach based on maximising modularity [2] to partition the inter- action network into functionally coherent modules. We then assign roles to proteins based on how their interactions are distributed within their own module and across other modules [3]. Based on a study of multiple yeast and human datasets, our results suggest that there is little evidence for a clear date/party distinction, but rather nodes in the protein interaction network seem to perform a variety of roles falling along a continuum, and there is no strong correlation between these roles and co-expression. We also examine alternative approaches to studying topological roles. So far, most work has focused on node-centric measures; here we attempt using a betweenness metric [4, 5] to quantify the centrality of links rather than nodes. We show that this measure relates to protein functional similarity as assessed by annotation overlap in the Gene Ontology [6], and may also be relevant to understanding how the interactome works as a system.

References:

[1] Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha J. Walhout, Michael E. Cusick, Frederick P. Roth, and Marc Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):8893, July 2004.

[2] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

[3] Roger Guimerà and Lúis A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895900, 2005.

[4] Linton C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):3541, 1977.

[5] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, 99(12):78217826, 2002.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):2529, May 2000.

20 Approximate Bayesian Computation under model uncertainty with an application to network data

Oliver Ratmann, Sylvia Richardson, Carsten Wiuf

In many areas of computational biology, the likelihood $f(x_0|\theta, M_i)$ of a scientific model M_i is now intractable, either because of the large amounts of relevant data x_0 or because interesting models are highly complex. This hampers scientific progress in terms of iterative data acquisition, parameter inference, model checking and model refinement within a Bayesian framework (1), because the workhorses of Bayesian inference, Markov Chain Monte Carlo and Sequential Monte Carlo are not any longer available.

One approach is based on the idea that simulating data x from $f(\cdot|\theta, M_i)$ is often much faster than computing $f(x_0|\theta, M_i)$ directly, so that a pseudo-likelihood approach is based on accepting those θ for which x is sufficiently close to x_0 . Approximate Bayesian Computation (2) embeds this idea within the Bayesian framework, and efficient MCMC and SMC algorithms have been developed (4; 5). By inclusion of unknown error terms between the model and the data, we show that ABC is intimately linked to Bayesian predictive diagnostics, set up a Bayesian analogue to predictive diagnostics for conservative model criticism when $f(x_0|\theta, M_i)$ is intractable, provide a deeper understanding of the ABC approximation, and give clues as to when this approximation may be biased by inferring posterior errors (3). We show how these ideas are implemented into existing MCMC methods at no or little additional computational cost.

The recent increase in biological data has placed a new focus on the complex networks embedded in biological systems, and one important goal is to create biologically realistic models of network formation, evolution and function (6). We have examined qualitative models of network evolution based on lateral gene transfer, duplication-divergence and link turnover under a number of different assumptions on missing data, and show that models of duplication-divergence (4) match the observed data best. Our work supports the great promises associated with ABC for the analysis of huge datasets or complex models, and shows that ABC fully embraces the process of statistical reasoning when $f(x_0|\theta, M_i)$ in intractable as it affords parameter inference as well as model criticism.

References

[1] George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791799, 1976.

[2] Paul Marjoram and Simon Tavare. Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet*, 7(10):759770, 2006.

[3] O. Ratmann, C. Andrieu, T. Hinkley, C. Wiuf, and S. Richardson. Model criticism with likelihood-free inference, with an example from evolutionary systems biology. *Proceedings of the National Academy of Sciences*, cond. acc., 2009.

[4] O. Ratmann, O. Jørgensen, T. Hinkley, M. P.H. Stumpf, S. Richardson, and C. Wiuf. Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H.pylori and P.falciparum. *PLoS Computational Biology*, 3(2007):e230, 11 2007.

[5] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 2008.

[6] C. Wiuf and O. Ratmann. Statistical and Evolutionary Analysis of Biological Networks, chapter Evolutionary analysis of protein interaction networks. Imperial College Press., 2009.

21 Analogical Reasoning in Information Retrieval

Ricardo Silva, Edoardo Airoldi, Katherine Heller, Zoubin Ghahramani

University College London, Harvard University, University College London, University of Cambridge

We illustrate a new methodology for the exploratory analysis of observations on pairs of sampling units (which include graphs and link matrices) motivated by the principles of analogical reasoning. We present an application which automatically ranks analogies between pairs of proteins that are known to interact (i.e., physically bind). For example, given a pair of interacting proteins, $P_1 : P_2$, which of two other interacting pairs, $P_3 : P_4$ or $P_5 : P_6$, interacts in a way that is most similar to $P_1 : P_2$? In other words, is the interaction $P_3 : P_4$ more analogous to $P_1 : P_2$ than $P_5 : P_6$ is? The goal of such exploratory analysis is to find new subclasses of interactions that might be relevant for further study: e.g., $P_1 : P_2$ might belong to a class of interactions that is not yet fully formalized, and scientists exploring the interaction between $P_1 : P_2$ might want to find other interactions which behave in an analogous way. We present a Bayesian formulation of this question and illustrate its potential application for exploring new taxonomies of protein-protein interactions.

This problem is non-trivial because direct similarity between objects is not an appropriate way of measuring analogies. For instance, the analogy between an electron around the nucleus of an atom and a planet around the Sun is hardly justified by isolated, non-relational, comparisons of an electron to a planet, and an atomic nucleus to the Sun. In our domain, each link in a link matrix is a pair of proteins $P_i : P_j$, the relation being a physical binding between the components.

Our application is built upon a measure of analogical similarity derived by Silva et al. (2007). In particular, it quantifies the similarity of two pairs of proteins ($P_1 : P_2, P_3 : P_4$) via a discriminative approach. It starts from the assumption that there is a (unknown) classification function $f_1(\cdot, \cdot)$ that classifies $P_1 : P_2$ as interacting (as opposed to a class of proteins that do not interact), and a (unknown) classification function f_2 for $P_3 : P_4$. We show how this task can be reduced to a novel variation of the Bayesian sets method (Ghahramani and Heller, 2005) for relational data with a discriminative model that effectively bridges observations on pairs of sampling units (i.e., edges in a graph) with high-dimensional vectors of unit-specific covariates (i.e., node-specific attributes).

References

Z. Ghahramani and K. Heller. Bayesian sets. 18th NIPS, 2005.

R. Silva, K. Heller, and Z. Ghahramani. Analogical reasoning with relational Bayesian sets. 11th International Conference on Artificial Intel ligence and Statistics, AISTATS, 2007.

22 From Here to Eternity: Developing Dynamic Network Models

Stephen E. Fienberg

Department of Statistics and Machine Learning Department, Carnegie Mellon University

Much of the recent literature on the modeling of network data has focused on snapshots of networks, often accumulated over periods of time. More interesting are dynamic network models but these are often simplistic or focus on selected network characteristics. In this presentation we attempt to provide a common framework for the modeling of networks evolving over time and we discuss how different strategies that appear in the literature fit within the framework.

23 The Effect of Random Social Interactions and Dynamic Network Structures on Product Adoption

Declan Mungovan, Enda Howley and Jim Duggan

Department of Information Technology, National University of Ireland, Galway.

Agent Based Modelling (ABM) techniques have been used in recent times to study the behaviour of complex systems. Agent based simulations are capable of generating populations of heterogeneous, self- interested agents that interact with one another. Emergent behaviour in a system may then be understood as a result of these individual interactions. Focusing on a simulated market duopoly we identify key feedbacks that capture individual consumer behaviour. Recent research in the domain of product dissemination has focused on the role that complex networks play on customer adoption of new goods. A "winner-take-all" outcome has been observed in situations where an agent's propensity to choose a product is influenced by the choices of others on their social network. The hypothesis states that a firm or technology that gets ahead tends to increase its market share. In this scenario an agent receives a utility from using a product that others have also adopted. Social networks in this domain to date has focused on a static network of acquaintance and have not considered the possibility of random interactions with the rest of the population. In our model agents interact with members of their own social network plus a second random network that is composed of a subset of the rest of the population. We formalise a method that chooses a random agent using a weighting method based on an individuals distance on the network we call preferential random choice. This means that friends-of-friends are more likely to randomly interact with one another than agents with a higher degree of separation. Using this method we aim to investigate the effect that random interactions have on the dissemination of product when an agent is primarily influenced by his existing acquaintances. We also investigate the effect of product adoption when agents can periodically change their acquaintances. Furthermore, previous ABM of product adoption has set each agent as being an adopter of either of two products with the final configuration being a "winner-take-all" or shared market arrangement. Here we allow agents to drop either product altogether if this pays the greatest utility. This will allow for a greater exploration of possible outcomes from our simulations. We aim to bridge the gap between the worlds of static complex networks of agents and fully connected environments where every agent interacts with everyone else.

24 Using Distinct Aspects of Social Network Analysis to Improve the Customer Loyalty Process

Carlos Andre Reis Pinheiro, Markus Helfert

School of Computing, Dublin City University,

Telecommunications market is characterized by an increased competitive environment. In order to maintain the customer base stable, telecom companies must provide valuable offers, suitable products and a range of pricing discounts. Most often, these actions are based on the customers behavior and additionally on their values in a corporate perspective.

Establishing likelihood assigned to the possibility of churn is not enough. Defining the customers value according personal information or billing behavior is not quite suitable. In order to optimize the customer loyalty process it is necessary to discern about the differences between the customers and highlight the characteristics more relevant for the company.

The main feature of any community is the relationship events between their members. From novel options for communications and relationships, with flexibility and mobility, those communities gained new boundaries, creating new means of relationships and establishing several social networks. The cohesion of these communities makes the peoples influence more relevant to the companies.

Social network analysis can reveal the possible correlations among the churns events inside communities, proving the stronger influence and impact when these events are triggered by a core node within the social network. Similarly, if the event is triggered by a boundary node the impact over the others members should be weaker. The influence means the number of customers which should follow the initial churns event in a chain processs perspective.

Social network analysis in telecommunications can help companies to recognize the customers behavior predicting the strength of links between the customers and the impact of the events among them. In this particular scenery it is more important to retain an influencer customer than a financial valuable one. In fact, the length of influence is important, thus representing the span of the triggered chain process.

Traditional supervised learned models based on neural artificial networks can establish a distinguish knowledge about the customers behavior. This type of model is fitted to specific classification goals as to predict churn events. In order to assess the correlation among the churn events it is important to analyze the events in a perspective of a chain. Some of the relevant analysis is figure out the relation among the events, the correlation assigned to events and the customers attributes, and mostly the relationship between the strength of the links and the churn events monitored.

The knowledge about the length of the customers influence can be used to define a new customers value and therefore allow companies to establish and perform more focused loyalty campaigns. This new perspective change substantially the traditional way that companies manage their customers, evaluating them based on their influence instead their usage or most common the billing account. Particularly in the telecommunications market this new approach can be quite relevant due the natural social networks hidden inside the data.

In order to identify the most valuable customers, we propose to apply social network analysis to discover their relations and thus their influence factors. A distinguish differentiation among the customers can be raised by this technique, identifying the customers importance and hence the best type of retention action which should be performed. The influence factor can reveal the customers which are able to trigger churn events in a chain perspective. This kind of knowledge is more relevant in a telecommunications scenery than traditional attributes, allowing companies to perceive the links, the nodes, and mainly the strength of both of them inside their networks.

Comparing a set of customers we calculate an influence factor, representing the value assigned to each individual customers plus the sum of their satellites nodes values. The value of the satellites nodes is weighted due the strength of the links between the nodes, which means the strength of these customers relationships. In this way, the customers value is established based on their linkages with other customers, the weights, the frequency and the recency of their connections, plus some personal and demographic attributes. The value depends on more the importance of the customer within the social network than his isolated score. This is quite different considering to the other approaches where the customers value is mostly based on isolated and individual attributes. In our approach, the customers value is based on the characteristics of the relationships more than the isolated ones, making these values more assigned to the influence factors than the personal behaviors. The customers value can be calculated based on the relative values assigned to the linkages and the value related to the other customers. In a chain event such as churn it is important to understand all aspects related to the network.

Using this kind of knowledge companies will be able to keep not just the higher value nodes but also the satellite nodes within the virtual social networks. In practical terms, companies will be able to retain the customers and their relations, which mean, the other customers who maintain some kind of connection among them.

25 Translating evidence into practice: a shared priority in public health?

H. McAneney, J.F. McCann L. Prior, J. Wilde and F. Kee

Over the last five years within the UK, the Research Councils, the Department of Health and major charities such as the Wellcome Trust, have begun to address the need to build capacity in public health research and to ensure better mechanisms for translating evidence into practice. Following reports such as Public Health Sciences: Challenges and Opportunities, major new ventures such as the National Prevention Research Initiative, the creation of Public Health Research Centres of Excellence, and the new public health stream of the National Institute for Health Research, appear to have forged a common purpose to support better research for better health. This study has capitalized on the occasion of the launch of one such Centre to describe the social networks of its stakeholders and investigate the nature and extent of the relationships between them.

Using results obtained from 98 respondents from 44 organizations and research clusters we have been able to assess the expectations, goals, and network connections of the respondents. Analysis of data on participant expectations and personal goals suggest that the academic members of the network were more likely to expect the work of the Centre to produce new knowledge as compared to non-academics, but less likely to expect the Centre to generate health interventions and influence health policy. Academics were also less strongly oriented than non-academics to knowledge transfer as a personal goal, though more confident that research findings would be diffused beyond the immediate network. A social network analysis of our data suggests that a central core of around 5 nodes is crucial to overall configuration of the regional public health network in Northern Ireland, and that whilst the overall network structure is fairly robust, the connections, between some component parts of the network - such as academics and the third sector - are unidirectional.

26 An inferential approach for one-sample gene network comparisons: assessing dissimilarity via a resampled local covering metric

Phillip Yates and Nitai Mukhopadhyay

Dept. of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA

The analysis of weighted co-expression gene sets is gaining momentum in bioinformatics circles. In addition to substantial research directed towards inferring co-expression networks on the basis of experimental data inferential methods are being developed to compare gene transcription networks across one or more phenotypes. Common gene set hypothesis testing procedures are mostly confined to comparing the average gene/node transcription levels between one or more groups and make limited use of additional network features. Ignoring the gene set architecture disregards relevant network topological comparisons and can result in familiar n¡p overparameterized test issues.

In this paper we propose to compare a sampled co-expression network with a target network where the measure of separation is determined with a local covering metric. The proposed metric is an additive measure over each of the genes/nodes in the network. Comparable to other local dependency structures, e.g., spatial interaction or AR(1) models, the dissimilarity measure at each node is determined using the network properties of nearby neighbors. Since the cover primarily relies on edges, weights, and in- and out-degrees the specification of a more complex network parameterization is avoided. In order to draw statistical inferences we use a resampling approach. Our method, which admittedly discounts large-distance or clustered co-expression effects, allows for both an overall network test and an examination of individual gene/node effects. In addition to testing for network relational differences our dissimilarity statistic can easily be extended to incorporate the mean/variance transcription comparisons of existing gene set methods. We evaluate our proposed metric using both simulated data (courtesy of the R library statnet) and the GSEA diabetes microarray data originally presented in Mootha et al. (2003). We will present our results along with some of the difficulties encountered in forming a nearby-neighbor resampling-based test statistic without the benefit of a full probability model.

27 Analysis of a Web-Based Network of Educators

April Galyardt, Turadg Aleahmad, Stephen Fienberg and Brian Junker

This study examines Classroom 2.0, a web community dedicated to helping teachers incorporate web 2.0 collaborative technologies into their classrooms. Members of the network have six modes of interaction available, but the majority of activity occurs in only two areas and forms two distinct subnetworks within the community. Content analysis using Latent Dirichlet Allocation suggests that one subnetwork is based on social connections and the other subnetwork is built around information sharing. The Holland and Leinhardt p1 model and a sample of ego-networks are used to compare the structure of the two subnetworks with previous findings that social networks and information-based networks have different underlying structure.

28 Inferences for Differential Transmission Rates of an Infectious Disease on a Network

Rachel Schutt

Columbia University

We model an infectious disease as a stochastic process on a social network, where differential transmission rates depend upon the characteristics of the nodes. By observing the infection status of nodes at two observation points in time, as well as the infection status of neighbors, and by treating infection times as missing data, we are able to infer differential transmission rates. Differential transmission rates can be used for predicting the size of an outbreak, and making intervention decisions that could be based on "quarantining" certain relationships particularly susceptible to the spread of disease. Further, it's possible that transmission rates could be used to infer network structure. We evaluate both discrete and continuous modeling schemes using data simulated from networks with a wide variety of properties.

29 Uncertainties in network analysis due to the thresholding problem

Andrew C. Thomas

Harvard University

In order to conduct analyses of systems where connections between individuals take on a range of values, a common technique is to dichotomize the data according to their positions with respect to a threshold value. However, there are two issues to consider: how the results of the analysis depend on the choice of threshold, and what role the presence of noise has on a system with respect to a fixed threshold value. I demonstrate consequences of each of these problems with respect to a set of commonly used generative network models.

30 Sensitivity of Spectral Clustering. Similarities between Spectral Clustering and Multidimensional Scaling

Zhivko Stoyanov

Bath University

A common practice in analysing the structure and clustering of big data sets is to represent them in a lower dimensional space, usually R^2 or R^3. The idea of that representation is that the features of interest of the data are preserved, while it is also much easier to view. One can then make conclusions by simply looking at the data, or further apply some well-known method for analysing or clustering. However, as it often happens in practice, data sets are subject to noise. So an interesting question to ask is: "How does the noise present in the data affect its lower-dimensional representation?" We consider two methods of lower-dimensional representation, Spectral Clustering and Multidimensional Scaling (MDS), and describe a possible way of analysing the sensitivity of Spectral Clustering when data is subject to random noise. We also attempt to find similarities between Spectral Clustering and MDS.

National Science Foundation

WHERE DISCOVERIES BEGIN

sfi

science foundation ireland

fondúireacht eolaíochta éireann

Irish Statistical Association

www.istat.ie

Clique

UCD DUBLIN

ims