WARSAW UNIVERSITY OF LIFE SCIENCES
SGGW

costnet

COSTNET18 CONFERENCE

BOOK OF ABSTRACTS

WARSAW, 26 - 28 SEPTEMBER 2018

Authors

WORKING ALONE, TOGETHER: FOR GOOD OR ILL? MULTIDISCIPLINARITY IN PRIMARY HEALTH CARE ORGANISATIONS

Nour Alrabie

Toulouse School of Management,
France

In France, Primary Health Care Organisations are known as Multidisciplinary Health Homes. This concept, over his ten years of existence, has resulted in a proliferation of this organisational form with more than one thousand health homes. Established all over the country, it brings liberal multidisciplinary healthcare professionals together to share a workspace and a health project. These liberal professionals continue to work on their own side by side but they are supposed to be collaborating and adopting a patient centred approach to offer an optimum care particularly for complex pathologies. On the base of four health homes, our study is willing to draw up these emerging networks and to detect the strength of established ties among different professionals according to their profession and frequency of communication. This would allow us to better understand if skill-mix (i.e. which professions to be brought together) makes sense by itself. If it does, we would try to find out what optimum skill-mix to be for professional integration. If not, we would try to identify the explanatory factors of strong and weak ties through a qualitative field residency.

Key words: Self-employment, Co-location, Collaboration, Skill-mix, Emerging networks

MEAN FIELD ANALYSIS OF PERSONALIZED PAGERANK WITH IMPLICATIONS FOR LOCAL GRAPH CLUSTERING

Konstantin Avrachenkov

Inria Sophia Antipolis,
France

Arun Kadavankandy

CentraléSupelec,
France

Nelly Litvak

University of Twente,
Netherlands

We analyse a mean-field model of Personalized PageRank on the Erdos-Renyi random graph containing a denser planted Erdos-Renyi subgraph. We investigate the regimes when the values of Personalized Page-Rank concentrate around the mean-field value and when they do not. We also study the optimization of the damping factor, the only parameter in Personalized PageRank. Our theoretical results help to understand the applicability of Personalized PageRank and its limitations for local graph clustering. This is a joint work with Arun Kadavankandy (CentraleSupelec) and Nelly Litvak (University of Twente).

Key words: Personalized PageRank, Mean Field, Local Graph Clustering

NETWORK ANALYSIS OF BIKE SHARING DATA

Vladimir Batagelj

Institute of Mathematics, Physics and Mechanics
Slovenia,
IAM Koper,
Slovenia

Many cities around the world provide a bike sharing service. Some of them (San Francisco Bay Area, New York, Chicago, Boston, Philadelphia, Washington D.C., Melbourne, Paris, London, etc.) offered as open data the data sets about trip histories. They usually consist of the trip start and end (day, time, station) and member's or bike's id. Sometimes they are augmented by additional data about stations and changes of their status. For some data sets also the weather data were collected. On the basis of these data, interesting insights into the dynamics of bike sharing systems can be obtained. We propose different symbolic networks based on trip distributions that summarize selected aspects of a given bike sharing system. We present some analyses of selected bike sharing systems based on the constructed networks.

Key words: trip distributions, dynamic modelling, network analysis, realtime data, symbolic data analysis

DYNAMIC CHAIN GRAPH MODELS FOR MIXED TIME SERIES DATA

Pariya Behrouzi

Wageningen University,
Netherlands

This paper introduces sparse dynamic chain graph models for network inference in high dimensional non-Gaussian time series data. The proposed method parametrized by a precision matrix that encodes the intra time-slice conditional independence among variables at a fixed time point, and an autoregressive coefficient that contains dynamic conditional independences interactions among time series components across consecutive time steps. The proposed model is a Gaussian copula vector autoregressive model, which is used to model sparse interactions in a high-dimensional setting. Estimation is achieved via a penalized EM algorithm. In this paper, we use an efficient coordinate descent algorithm to optimize the penalized log-likelihood with the smoothly clipped absolute deviation penalty. We demonstrate our approach on simulated and genomic datasets. The method is implemented in an R package tsnetwork.

Key words: Dynamic networks, Chain graph models, time-series data, Gaussian copula

DOES COOPERATION BETWEEN HOSPITALS INCREASE THE QUALITY OF CARE?

Paolo Berta

Department of Mathematics,
Brunel University London,
UK,
Department of Statistics and Quantitative Methods,
University Bicocca-Milan,
Italy

Francesco Moscone
Veronica Vinciotti

Department of Mathematics,
Brunel University London,
UK

Motivated by reasons such as convenience or altruism, hospitals often engage in mutual cooperation. In this paper, we evaluate whether cooperation between hospitals has an impact on the hospitals' quality. We take the network of patients' transfers between hospitals as a proxy of their level of cooperation, and we focus in particular on hospital discharges for over 900,000 patients admitted to 145 hospitals in the Lombardy region (Italy) in 2014. Around 1.7% of these were transferred to other hospitals after admission. Detecting whether cooperation has an impact on quality is confounded by the fact that the decision made by hospitals to transfer patients may be partly motivated by the quality of the destination hospital. For this reason, as a first step, we use a social relation model to identify exogenous determinants of patients' flows, such as the distance between hospitals. As a second step, a Poisson mixed effect model is used to detect the impact of the network of predicted transfers on the overall mortality of the origin and destination hospitals. The findings of this analysis showed that hospital cooperation does increase quality and thus that it should be better supported by policy makers in order to stimulate the improvement of the healthcare systems.

Key words: social relations model, community detection method, healthcare quality, hospital cooperation

MEASURING RANK ROBUSTNESS IN SCORED PROTEIN INTERACTION NETWORKS

Lyuba Bozhilova

University of Oxford,
UK

Protein interaction networks (PINs) are models of cellular architecture in which nodes represent proteins, and edges represent the binary physical or functional interactions between them. Often these interactions are assigned confidence scores, the purpose of which is to quantify the strength of any available experimental evidence. PINs are then constructed by applying a score cut-off and only considering the interactions which meet this ad-hoc data quality threshold.This means that multiple PINs can be built to model the same set of biological processes of interest by varying the score threshold. In this talk we argue that if a node centrality, such as degree or betweenness, is to be useful for extracting biological signal from PINs, it should ideally induce similar node rankings across a range of thresholds.We propose three measures - rank continuity, identifiability, and instability - for such threshold robustness. We apply these to a set of twenty-five metrics of which we identify four: number of edges in the step-1 ego network, and the leave-one-out difference in average redundancy, average number of edges in the step-1 ego network, and natural connectivity, as robust across medium-high confidence thresholds. Our measures show good agreement across PINs from different species and data sources. However, analysis of synthetically generated scored networks shows that robustness results are context-specific, and depend both on network topology and on how scores are placed across network edges.

Key words: protein interaction networks, rank, robustness

AGENT-BASED MODELING OF THE GLOBAL PHONOGRAPHIC MARKET.

Andrzej Buda

Interdisciplinary Research Institute,
Poland

We have investigated the product life-cycles of more than 17 000 hit singles performed on the 12 biggest national phonographic markets in Europe. We have considered weekly singles charts from the last 60 years in each country. We consider the spread of hit singles popularity (chart topping) as an epidemiological process performed on various European countries. Popular hit singles are contagious from one country to another. Thus, the stages of popularity spread might be defined by time delays between the initial hit single release and reaching the highest position on consecutive national singles charts. According to these empirical results, we consider the global phonographic market as a network of national phonographic markets. Each local phonographic market has Barabasi-Albert structure and includes a node with the highest connectivity defined as 'capital'. Local phonographic markets are connected to each other by capitals only. We have applied Agent-Based Modeling in global phonographic market to investigate popularity spread that starts from random initial node and the final trajectory becomes a superposition of product life-cycles in local markets that might affect each other by their capitals. In our research we consider the global contemporary Internet era and historical situation where the global phonographic market consists of local subnetworks that have been divided into cultural clusters. Our results are consistent with empirical data, where only a minority of songs becomes a hit single globally and the most important parameter is attractiveness of a product. The success of a song is determined by initial node and cultural clusters as well. There is no significant dependence on size of a market. However, when initial node is peripherial or belongs to big markets, it may extend a global product life-cycle significantly. Our Big Data analysis may be applicable within typical marketing strategies or boundary conditions including heavy promotion of unattractive product (it extends product life cycles extremely, but may affect a few markets from the same cluster only).PACS numbers: 87.23.Ge, 89.75.-k, 89.65.-s, 87.10.-e, 87.19.xb

Key words: agent based modeling, big data, social networks, phonographic market

LEARNINGS FROM LARGE SCALE EXPERIMENTS ON VIRAL MARKETING AND VIRAL SPREADING

Geoffrey Canright

Telenor Research,
Norway

In this talk I will present some results from roughly 6 years of experiments on viral marketing. The aim is to present many interesting empirical results, but then necessarily only at a high level. These results will include: viral spreading that occurred 'by accident', rather than by design; how having the possibility to give an offer to friends influences the likelihood of adoption; the extremely uneven distribution of adoption over the social network; how adoption and spreading are affected when marketing to pairs of individuals, rather than to isolated individuals; and how word-of-mouth spreading and Internet-mediated spreading each display their own distinct signatures-which we can tease out from the data.

Key words: viral marketing, social networks, mobile phone data

OBJECTIVE BAYES MODEL SELECTION OF GAUSSIAN INTERVENTIONAL ESSENTIAL GRAPHS

Federico Castelletti

Universita Cattolica del Sacro Cuore (Milano),
Italy

Graphical models based on Directed Acyclic Graphs (DAGs) are a very common tool in many scientific areas for the investigation of dependencies among variables. Typically, the objective is to infer models from the data or measuring dependence relationships between variables. The set of all (marginal and) conditional independencies encoded by a DAG determines its Markov property.However, it is well known that we cannot distinguish between DAGs encoding the same set of conditional independencies (Markov equivalent DAGs) using observational data.Markov equivalent DAGs are then collected in equivalence classes each one represented by an Essential Graph (EG), also called Completed Partially Directed Graph (CPDAG).When the interest is in model selection it is then convenient to explore the EG space, rather than the whole DAG space, even if the number of EGs increases super-exponentially with the number of vertices. An exhaustive enumeration of all EGs is not feasible and so structural learning in the EG space has been confined to small dimensional problems.However, to avoid such limit, several methods based on Markov chains have been proposed in recent years.In many applications (such as genomics) we have both observational and interventional data produced after an exogenous perturbation of some variables or from randomized intervention experiments. The concept of intervention is strictly related to the causal interpretation of a DAG. Interventions destroy the original causal dependency on the intervened variables and modify the Markov property of a DAG. This results in a finer partition of DAGs into equivalence classes, each one represented by an Interventional Essential Graph (I-EG). Hence, model selection of DAGs in the presence of observational and interventional data can be performed over the I-EG space, thus improving the identifiability of the true data generating model.In this work we deal with the problem of Gaussian DAG model selection from a Bayesian perspective. In particular, we adopt an objective Bayes approach based on the notion of fractional Bayes factor. We then obtain a closed formula to compute the marginal likelihood of an I-EG given a collection of observational and interventional data.Next, we construct a Markov chain to explore the I-EG space possibly accounting for sparsity constraints. Hence, we propose an MCMC algorithm to approximate the posterior distribution of I-EGs and provide a quantification of inferential uncertainty by measuring some features of interest, such as probabilities of edge inclusion.We apply our methodology, that we name Objective Bayesian Interventional Essential graph Search (OBIES) to simulation settings and to the analysis of protein-signaling data, where a collection of observations measured under different experimental conditions can be interpreted as interventional data.

Key words: Graphical models, Interventional data, Causal inference

THE INTERRELATION BETWEEN PERSONAL NETWORKS AND CRITICAL EVENTS IN THE LIFE COURSE PERSEPCTIVE

Vida Cesnuityte

Mykolas Romeris University,
Lithuania

The research aim is to measure the interrelation between the structure and size of the personal network in face of critical events of the individual life course. The hypothesis is that personal network interrelated with particular stage of the individual life course but, also, with critical events in the individual's life. The analysis is based on representative quantitative survey data collected within the project „Trajectories of family models and social networks: intergenerational perspective. Fieldwork carried out between November, 2011 - February, 2012. Sample size of 2000 respondents represent four birth cohorts of Lithuanian habitants: 1950-1955, 1960-1965, 1970-1975 and 1980-1985. The following situations of vulnerability experienced by the respondents personally or by their family members were explored: unemployment of family members, long-term illness, caring for a sick family member, abortion, death, suicide, encounters with legislation, addiction, depression, domestic violence, family conflict, divorce, infidelity of spouse or partner, poor housing, financial difficulties, emigration. The Event History Analysis methods applied in the analysis.

Key words: personal network, individual life course, critical events, Event History Analysis

VLOGGING IN ROMANIA: AN EXPLORATORY RESEARCH OF VLOGS PREFERENCES AMONG TEENAGERS FROM BIHOR COUNTY

Smaranda Cioban
Raluca Buhas
Adrian Hatos

University of Oradea,
Romania

The study presents a classification of a group of teenagers in relation to the content they access on Youtube. It analyzes the results of an online survey conducted by students from the University of Oradea under the coordination of professor Adrian Hatos and Raluca Buha? about the Youtube preferences of teenagers from Bihor county. The 196 respondents that took part at this research were teenagers from Bihor county selected by students from their acquintances. Out of the 196 participants, only two declared that they do not watch vlog channels. Although the survey did not explicitly asked participants to express the reasons why they prefer a certain content, we provided possible answers to this question by exploring the vlogger network that resulted from the way participants ranked their favorite vloggers. In this sense, we calculated the relationship between modularity classes and sociodemographic characteristics. The relationship was calculated through logistic regression with IBM SPSS 20. Although the analyzed network is very small (56 vloggers were considered relevant out of the 336 identified ; the analyzed network consists of the connection between the 56 vloggers and the 196 participations based on the way they ranked the favorite vloggers.

Key words: youtube, vlogging, Social Networks, modularity, new communication, digital sociology

THE PERSONALIZATION-PRIVACY PARADOX IN ATTENTION ECONOMY

Julien Cloarec

Toulouse School of Management,
France

"The more people use their addictive-by-design social media, the more attention social-media companies can sell to advertisersand the more data about the users behaviour they can collect for themselves." (The Economist, 2017) "Facebook, Twitter, Instagram, Google [are] caught in a zero-sum race for our finite attention. [...] No other media used a precise, personalized profile of everything we've said, shared, clicked, and watched to influence our behavior at this scale." (Time Well Spent, co-founded by Tristan Harris, former Design Ethicist at Google)

Personalization is the way marketers learn the consumers' preferences, adapt their offers, and evaluate the effectiveness (Wedel and Kannan, 2016, JM). These three stages refer to close-loop marketing, which can be fully automated through adaptive personalization. In the end, this will lead to automated attention analysis (Wedel and Kannan, 2016, JM). However, one should not forget that personalization and privacy are entwined. The major implication of the trade-off between these two constructs is the personalizationprivacy paradox, which refers to the tension between personalization and privacy, which follows from marketers exploiting consumers data to offer personalized product information (Sutanto et al., 2013, MISQ). While this trade-off redefine online business models (Casadesus-Masanell and Hervas-Drane, 2015, MGS), it lacks comprehensive theoretical foundations (Sutanto et al., 2013, MISQ). The current management literature raises several research opportunities : capturing the richness of firms' advertising (Yadav and Pavlou, 2014, JM), understanding consumer and privacy intersections (Martin and Murphy, 2016, JAMS), engaging customers despite distractions (Kannan and Li, 2017, IJRM), directing attention to personalized recommendations (Li and Kahanna, 2015, JAIS), and dealing with shorter attention spans (Chung, Wedel and Rust, 2016, JAMS). This can be achieved by combining economic models (i.e., personalization-privacy trade-offs in macro-environments) and marketing concepts (i.e., attention) (Yadav and Pavlou, 2014, JM).

"Information consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention" (Simon, 1971, p. 40). Falkinger (2007, JET) defines an attention economy as "a family of senders, which employ costly signals to attract the attention of receivers characterized by their attention capacity" (p. 268). While, at first, personalization can keep attention focused (Benlian, 2015, JMIS), capturing consumers' attention leads to a wasteful competition for attention (Falkinger, 2008, EJ). In particular, even though humans can divide their attention (Benlian, 2015, JMIS), technology fragments attention (Ayyagari et al., 2011, MISQ).

Multi-sided platforms, such as social networking sites, are competing for our attention and the implications of the personalization-privacy paradox are complex. Personalization involves attention costs (Casadesus- Masanell and Hervas-Drane, 2015, MGS) because it captures consumers' selective attention while increasing both intrusiveness (Bleier and Eisenbeiss, 2015, MKS) and privacy concerns (Gal-Or, Gal-Or and Penmetsa, 2018, ISR). In particular, when personalization is too obtrusive, it creates perception of manipulation (Goldfarb and Tucker, 2011, MKS). Capturing attention, which is a limited cognitive resource, increases the consumers' level of effort. This has major moderating effects on the consequences of privacy concerns (Dinev, McConnell, and Smith, 2015, ISR).

Markets for consumer information are different from usual economic markets (Chellappa and Shivendu, 2007, JMIS). The competition caused by personalization is localized to individual consumers (Lee, Ahn and Bang, 2011, MISQ), which might be detrimental for companies (e.g., when personalized prices are too low). Fortunately, privacy protection can be a competitive benefit for firms (Lee, Ahn and Bang, 2011, MISQ), aven though competing for attention often involves the creation of a secondary market of consumer information (Casadesus-Masanell and Hervas-Drane, 2015, MGS).

Key words: Personalization-Privacy Paradox, Attention Economy, Competition

BAYESIAN LEARNING OF MARKOV EQUIVALENCE CLASSES OF DAGS WITH OBSERVATIONAL AND INTERVENTIONAL DATA.

Guido Consonni

Universita Cattolica SC,
Italy

Graphical models based on Directed Acyclic Graphs (DAGs) represent a powerful tool for investigating dependencies among variables. It is well known that one cannot distinguish between DAGs encoding the same set of conditional independencies (Markov equivalent DAGs) using only observational data. However, the space of all DAGs can be partitioned into Markov equivalence classes, each being represented by a unique Essential Graph (EG), also called Completed Partially Directed Graph (CPDAG).In some fields, in particular genomics, one can have both observational and interventional data, the latter being produced after an exogenous perturbation of some variables in the system, or from randomized intervention experiments. Interventions destroy the original causal structure, and modify the Markov property of the underlying DAG, leading to a finer partition of DAGs into equivalence classes, each one being represented by an Interventional Essential Graph (I-EG) (Hauser and Buehlmann).In this talk we consider Bayesian model selection of EGs under the assumption that the variables are jointly Gaussian. In particular, we adopt an objective Bayes approach, based on the notion of fractional Bayes factor, and obtain a closed form expression for the marginal likelihood of an EG. Next we construct a Markov chain to explore the EG space under a sparsity constraint, and propose an MCMC algorithm to approximate the posterior distribution over the space of EGs. Our methodology, which we name Objective Bayes Essential graph Search (OBES), allows to evaluate the inferential uncertainty associated to any features of interest, for instance the posterior probability of edge inclusion. An extension of OBES to deal simultaneously with observational and interventional data is also presented: this involves suitable modifications of the likelihood and prior, as well as of the MCMC algorithm. We conclude by presenting results for simulated and real experiments (protein-signaling data).This is joint work with Federico Castelletti, Stefano Peluso and Marco Della Vedova (Universita' Cattolica del Sacro Cuore).

Key words: Graphical models, Model selection, Objective Bayes

A BAYESIAN HIERARCHICAL REGRESSION MODEL FOR LONGITUDINAL CLONAL COMPARISON IN GENE THERAPY STUDIES

Luca Del Core

Johann Bernoulli Institute for Mathematics and Computer Sciences,
University of Groningen,
Netherlands,
San Raffaele Telethon Institute for Gene Therapy (SR-Tiget), Milan,
Italy

Andrea Calabria
Daniela Cesana
Eugenio Montini

San Raffaele Telethon Institute for Gene Therapy (SR-Tiget), Milan,
Italy

Marco Grzegorczyk

Johann Bernoulli Institute for Mathematics and Computer Sciences,
University of Groningen,
Netherlands

Hematopoietic stem cell Gene Therapy (GT) using integrating viral vectors has become effective medical treatment to cure patients affected by monogenic diseases, demonstrating the safety and long term efficacy of the treatment. The molecular analysis of stem cell clones and its progeny identified by vector integration sites (IS) allows the assessment of viral vector safety, monitoring over time the clonality, that is the heterogeneity of the clonal population, and clonal expansion. In this emerging context, no statistical methods exist for the comparative analysis of the clonality among different groups of vectors. With this project, we aim at analysing the clonal population over time to better characterise a genotoxic profile.

We used experimental data from a mouse study aimed at evaluating clonal expansions in tumour prone mice and control mice using genotoxic and non-genotoxic vectors. IS data have been collected over time (monthly) from whole blood and in four lineages (myeloid and lymphoid compartments). Population clonal diversity and clonal expansion have been measured using the Shannon Entropy as reported in many GT applications. We analysed to which extent the different factors (treatments, cell lineages, mouse genotypes) affect the entropy. Due to the hierarchical shape of the data, we implemented a Bayesian hierarchical regression model, in which the entropy (as proxy for diversity/expansion) takes the role of the target. Since the full conditional distributions of the involved parameters are analytically available, a Gibbs sampling is implemented in R for MCMC simulations, which result to be mixed and convergent. Our results suggest that some factors may have an effect on the clonal diversity/expansion and will be further investigated.

Key words: Bayesian Statistics, Hierarchical Regression Models, Gene Therapy, Preclinical Studies, Clonal Diversity, Clonal Expansion

BRAIN NETWORKS AS A PREDICTOR OF AGING ALONG LIFESPAN

Jesus M. Cortes

Biocruces Health Research Institute,
Spain

Physiological aging affects brain structure and function impacting morphology, connectivity and performance. Here, I will talk on recent results in our laboratory showing how the brain networks of an individual predict its level of aging. In particular, by building functional (resting-state) and structural (tractography) brain networks from neuroimages of healthy participants (N=155) ranging from 10 to 80 years, we introduced the new metrics of brain-connectome age (BCA), defined as the age estimator resulting from the structure-function connectome. We next compared BCA to the chronological age (ChA) to find two-fold results: First, aging widely affects the connectivity of multiple brain structures, such as anterior cingulate, medial prefrontal cortices, basal ganglia, thalamus, insula, cingulum, hippocampus, parahippocampus, occipital cortex, fusiform, precuneus and temporal pole. Second, the connectivity between basal ganglia and thalamus to frontal areas, a.k.a. the fronto-striato-thalamic circuit, makes the major contribution to age estimation. In conclusion, our results highlight the key role played by the fronto-striato-thalamic circuit in the process of healthy aging. Notably, the same methodology can be generally applied to identify the structural-functional brain network patterns correlating to other biomarkers than ChA.

Key words: brain networks, maximum likelihood estimator, physiological aging

SYMMETRIC CORE-COHESIVE BLOCKMODEL IN AN INTERACTIONAL PRE-SCHOOL NETWORKS

Marjan Cugmas
Aleš Žiberna
Anuška Ferligoj

Faculty of Social Sciences,
University of Ljubljana,
Slovenia

The symmetric core-cohesive blockmodel is proposed. A blockmodel is a network where the units are clusters of equivalent (according to the structure of the links) units from the studied network. The term block, in the context of blockmodels, refers to a submatrix showing the links between two clusters. The symmetric core-cohesive blockmodel consists of three or more clusters. The units from each cluster are internally well linked while those from different clusters are ideally not linked to each other. The exception are the units from so called core cluster. These units have a mutual links to all the units in the network. The other clusters are called cohesive clusters.The presentation addresses the question of whether the symmetric core-cohesive blockmodel can be found in interactional pre-school networks. To this aim, the data collected among pre-school children in Head Start preschools (United States, data collected in 2005 and 2006) are analysed. The Monte Carlo simulations are used to confirm that the selected social mechanisms (e.g., popularity, transitivity, mutuality, assortativity), defined on the local network level, can lead to the symmetric core-cohesive blockmodel global structure.

Key words: blockmodeling, interactional networks, local mechanisms

GENE COEXPRESSION NETWORK FOR THE STUDY OF RHIZOBIUM LEGUMINOSARUM

Javier Pardo Diaz
Mariano Beguerisse
Philip Poole
Charlotte Deane
Gesine Reinert

University of Oxford,
UK

Rhizobium leguminosarum is a bacterium that fixes atmospheric nitrogen when associated with plants of the legume family (eg. peas, beans, lentils...), improving legume growth. Nitrogen fixation is governed by a complex interplay of genes and data give rise to gene coexpression networks. Network model construction and techniques such as community detection can help understand this nitrogen fixation process and help find the genes involved in it. R. leguminosarum is an interesting organism to study due to the very large changes in its metabolism from the "free-living" bacteria to the "plant-associated" bacteria. These differences also affect the transcriptome and can be captured by gene coexpression data. We use rich but noisy gene expression data from triplicates of microarrays from three different groups of conditions: free-living bacteria, bacteria from the soil where the legumes were growing and bacteria from inside the plant to build our network. We measured the correlation between the expression of genes under these different growth conditions. We then imposed a threshold to select only the strongest relationships. We constructed a network in which nodes are the genes and connections represent high positive correlations. We evaluated the network against published data from three different databases: KEGG, STRING and OperonDB. The results obtained suggest that the network will be useful for the identification of biological pathways. We aim next to use community detection to identify groups of genes involved in the same biological processes.

Key words: Gene Coexpression Network, Community Detection, Bioinformatics

DISCRETE-TIME MODEL FOR A MOTION OF SUBSTANCE IN A CHANNEL OF A NETWORK WITH APPLICATION TO CHANNELS OF HUMAN MIGRATION

Zlatinka Dimitrova
Nikolay K. Vitanov

Bulgarian Academy of Sciences,
Bulgaria

We discuss a discrete-time model for motion of substance in a channel of a network. For the case of stationary motion of the substance and for the case of time-independent values of the parameters of the model we obtain a new class of statistical distributions that describe the distribution of the substance along the nodes of the channel. The case of interaction between a kind of substance specific for a node of the network and another kind of substance that is leaked from the channel is studied in presence of possibility for conversion between the two substances. Several scenarios connected to the dynamics of the two kinds of substances are described. The studied models: (i) model of motion of substance through a channel of a network, and (ii) model of interaction between two kinds of substances in a network node connected to the channel, are discussed from the point of view of human migration dynamics and interaction between the population of migrants and the native population of a country.

Key words: networks, channels of networks, migration, network flows, statistical distributions of substance in channel of a network

ANOMALY DETECTION IN NETWORKS USING SPECTRAL METHODS AND NETWORK COMPARISON APPROACHES

Mihai Cucuringu

Alan Turing Institute,
UK,
University of Oxford,
UK

Andrew Elliott

Alan Turing Institute,
UK

Milton Martinez Luaces
Paul Reidy

Accenture Plc

Gesine Reinert

Alan Turing Institute,
Department of Statistics,
University of Oxford
UK

Detecting financial fraud is a global challenge with financial fraud losses estimated to at least GBP 768 million in the UK in 2015. This paper is motivated by the task of detecting anomalies in networks of financial transactions, with accounts as nodes and a directed weighted edge between two nodes if one of the nodes transfers money to the other one. The weight of the edge denotes the transaction amount.

Anomalies in networks may, for example, constitute a long path of large transaction amounts, or a ring of large payments; cliques of accounts which send money to each other are also another possible anomaly. In this paper we introduce a method which is able to detect previously unspecified anomalies in networks. The method is based on a combination of graph comparison with spectral analysis techniques. Altogether, graph comparison,spectral analysis and a set of local statistics give rise to a total of 140 features. We then use a simple feature sum method, as well as a random forest method, in order to classify nodes as normal or anomalous.

We first test the method on synthetic networks which we generated, and second on an additional synthetic network model, the generating mechanism of which was unknown during the main research stage to the team developing the method. The first set of synthetic networks was split in a training set of 70 percent of the networks, and a test set of 30 percent of the networks. The resulting classifier was then tested on the second synthetic network as well.

We compare our method with Oddball, a widely used method for anomaly detection in networks, as well as to random classification. While Oddball outperforms random classification, both our feature sum method and our random forest method outperform Oddball on the first synthetic model. On the test set of the first model, the random forest outperforms feature sum, whereas on the second synthetic data set, initially feature sum tends to pick up more anomalies than random forest, with this behaviour reversing for later anomalies. In both of our methods, the top 2 percent of nodes contained on average over 90 percent of the planted anomalies.

CITATION NETWORK ANALYSIS OF THE LITERATURE ON CLUSTERING NETWORKS

Anuška Ferligoj

University of Ljubljana,
Slovenia

Vladimir Batagelj

Institute of Mathematics, Physics and Mechanics
Slovenia,
IAM Koper,
Slovenia

Patrick Doreian

Department of Sociology,
University of Pittsburgh,
USA,
Faculty of Social Sciences
University of Ljubljana
Slovenia

There is a large literature on clustering networks. Bibliographic data for the network clustering literature including both community detection and blockmodeling works through to February 22, 2017 was collected. The primary data source was the Web of Science. From the obtained data we created a citation network among works. In addition, we included data on authors, journals and keywords to generate some two-mode networks. Lists of the most prominent journals where works in the network clustering literature appeared were created. The CPM path was identified. It revealed a clear transition from the social network part of the literature to the community detection part. The key-route paths revealed the same transition but with more works and a more nuanced view of it. Ten link islands, as clusters, were identified. Detailed discussion of the obtained results were provided with a clear distinction between the community detection and social networks literatures.

EXTRACTING THE VALUE OF BIG DATA

Denise De Gaetano

Malta College of Arts,
Science and Technology,
Malta

Despite the large volumes of data, companies still struggle to access, manage and extract the information that their day-to-day processes generate. The growth of IT systems has provided these same companies, the ability to capture this potentially valuable data, within a number of applications, databases and organizations. Apart from a strong IT infrastructure, changes in the board and management, within a company also need to be carried out. This allows a number of departments to work in coordination and help ensure success of the utilization of the data at hand. Understanding what is required out of the data, is the first step to generate the best results from the company and customer data. The strategy is to understand how the information can enable an improvement in the business.

Key words: big data, real-time analysis, aviation

NON-HOMOGENEOUS DYNAMIC BAYESIAN NETWORK MODELS

Marco Grzegorczyk

Johann Bernoulli Institute for Mathematics and Computer Sciences,
University of Groningen,
Netherlands

In systems biology, non-homogeneous dynamic Bayesian networks (NH-DBNs) have become a popular modelling tool for learning the topologies of cellular networks from temporal measurements. In the presentation two novel NH-DBNs for network structure learning will be presented. (1) When the data correspond to one long non-homogeneous time series, a multiple changepoint process can be used to divide the data into disjoint segments with segment-specific network parameters. In the standard uncoupled NH-DBN those parameters are learned separately for each segment, although parameters might stay identical (or similar) over time. To allow for information-exchange among segments, NH-DBNs with coupled network parameters have been proposed. When the parameters stay similar over time, coupled NH-DBNs lead to better performances, but the coupled models have no mechanisms to uncouple the parameters. Parameter coupling can be counter-productive when the segment-specific parameters are dissimilar. In real-world applications, not rarely, there is a mixture of both parameter types: parameters that evolve gradually (smoothly) and parameters that change drastically over time. For this scenario we propose a new NH-DBN with partially edge-wise coupled network parameters. The new model operates edge-wise and infers for each individual edge whether the corresponding parameter stays similar and should thus be coupled or whether the parameter changes from segment to segment and should better stay uncoupled. (2) The second model is based on a similar idea, but has been developed for the complementary scenario that a set of short time series has been measured under different experimental conditions. In this setting, usually some network parameters are condition-specific, while other parameters stay constant across conditions. For this situation, we propose a new partially NH-DBN, which infers for each individual edge whether the corresponding parameter is constant among conditions or condition-specific. While the NH-DBN model with partially edge-wise coupled parameters (1) infers the right trade-off between an uncoupled and a coupled non-homogeneous dynamic Bayesian network model, the partially NH-DBN model (2) infers the right trade off between a homogeneous and a non-homogeneous dynamic Bayesian network model. Our empirical results for (1-2) show that both models can lead to significantly improved network structure predictions.

NON TRANSPARENCY IN LARGE NETWORKS

Gholamreza Jafari

Physics Department,
Shahid Behesti University,
Iran,
The Center for Network Science,
Central European University,
Hungary

In studying the network dynamics of complex systems such as the world-wide web, genes regulatory networks, and social networks it regularly happens that we do not have access to the complete information about them. Not only the information is incomplete, but also some of the information may be wrong. So, how can we understand the general or particular behavior of such networks in these scenarios?! In this type of giant networks, a new and very interesting type of uncertainty arises when there is absolutely no guarantee whether a node or a link is visible at all for some parties. There is a network of colleagues, a network of high school classmates, a network of friends in Facebook, a network of relatives and so many other networks forming a very big puzzle of different types of relationships. This puzzle is like a multilayer or multidimensional network which we call it, the Dark Network, a network of networks, containing the whole information of socio-technical or biological systems. Almost no one can have access to the Dark Network, and even if some does, it is a not a kind of problems to be solved computationally! What supercomputer can handle this amount of data? This problem can even get more interesting when the present knowledge between the parties are not the same! What type of effort can we make to settle the dispute between two agents or parties when they have different pictures of something in common? Notice that every two contradicting perspectives about such situations can be correct, depending on what side you are looking at the problem! We are actually developing physics-based algorithms which can point to hidden connections between spatially disparate nodes of the Dark Network. This includes finding clusters and communities emerged from such a many-many-body system, especially when the expectation of sub-networks about the whole picture are contradicting each other.

Key words: Social networks, Non transparency, Large networks

MODEL OF FUTURE PROPAGATION OF AFRICAN SWINE FEVER IN EASTERN AND CENTRAL EUROPE

Andrzej Jarynowski

Interdisciplinary Research Institute in Wroclaw,
Poland

ASF virus is a domestic and wild pig infection, which is probably the most important current economic threat in Europe causing already billions EUR losses. It spreads from East to West of Europe with speed of around 200km per year. We introduced an early warning analysis for ASF propagation in Eastern Europe. On the basis of older outbreaks we want to predict where and when outbreaks will take place in the future.More than 4000 infection events that took place since February 2014 have been analysed according to time, longitude and latitude.We choose a pseudo gravity propagation model by taking into account additionally to infection notifications: domestic pig abundance, disease vectors (wild boar) abundance, pork production, and human failure to restrictions. We use the effective distance methodology as well as Monte Carlo simulations and inferred parameters for both approaches to predict future propagation patterns.

Key words: African Swine Fever, veterinary epidemiology, epizoology, infecious disease modeling

GRAPHICAL MARKOV MODELS WITH AN APPLICATION ON TRAFFIC ACCIDENT DATA

Cem Kadılar

Hacettepe University,
Turkey

Undirected graphical models are widely used for modeling, visualization, inference, and exploratory analysis of multivariate data with wide-ranging applications. Graphical models are models based on graphs in which nodes represent random variables, and the edges represent conditional independence assumptions. Graphical Markov models are multivariate statistical models which are currently under vigorous development and which combine two simple but most powerful notions, generating processes in single and joint response variables and conditional independences captured by graphs. The development of graphical Markov started with work by Wermuth (1976, 1980). The study of these models is an active research area, with many questions still open. In this study, graphical Markov models are described. Then, interpretations are illustrated with an application using traffic accident data set of Turkey.

Key words: graphical markov chain, air pollution, precipitation

PARTIALLY NON-HOMOGENEOUS DYNAMIC BAYESIAN NETWORKS BASED ON BAYESIAN REGRESSION MODELS WITH PARTITIONED DESIGN MATRICES,GAUSSIAN PROCESS

Mahdi Shafiee Kamalabad

University of Groningen,
Netherlands

Non-homogeneous dynamic Bayesian networks (NH-DBNs) are a popular modelling tool for learning cellular networks from time series data. In systems biology, time series are often measured under different experimental conditions, and not rarely only some network interaction parameters depend on the condition while the other parameters stay constant across conditions. For this situation, we propose a novel partially NH-DBN, based on Bayesian hierarchical regression models with partitioned design matrices. With regard to our application to timecourse data from mammalian target of rapamycin complex 1 (mTORC1) signalling, we also propose a Gaussian process based method to solve the problem of non-equidistant time series measurements. On synthetic network data and on yeast gene expression data the new model leads to improved network reconstruction accuracies. We then use the new model to reconstruct the topology of the mTORC1 network. The inferred network topology shows features that are consistent with the biological literature.

Key words: Non-homogeneous dynamic Bayesian networks, Bayesian hierarchical regression models, partitioned design matrices,network reconstruction accuracy

FUNCTIONAL APPROXIMATIONS VIA STEIN'S METHOD WITH APPLICATIONS TO U-STATISTICS AND RANDOM GRAPHS

Mikołaj Kasprzak

University of Oxford,
UK

We provide a strategy for comparing distributions of stochastic processes using a generalisation of Stein's method with an emphasis on the exchangeable-pair approach. Our method is applied to bounding the distance between scaled non-degenerate U-statistics and Brownian Motion, between a class of scaled degenerate U-statistics and piece-wise constant Gaussian Processes and between some statistics connected to a graph-valued process and a limit continuous Gaussian Process.

Key words: Stein's method, U-statistics, random graphs

DEEP SPECTRAL GRAPH MATCHING NETWORK

Yosi Keller

Bar Ilan University,
Israel

In this work we present a Deep Learning based approach for visual correspondence estimation, by deriving a deep spectral graph matching network. We formulate the state-of-the-art unsupervised Spectral Graph Matching (SGM) approach, as part of an end-to-end supervised deep learning network. Thus allowing to utilize backpropagation to learn optimal image features, as well as algorithm parameters. For that we present a transformation layer that converts the learned image feature, within a pair of images, to an anity matrix used to solve the matching problem via a new metric loss function. The proposed scheme is shown to compare favorably with contemporary state-of-the-art matching schemes when applied to annotated data obtained from the PASCAL, ILSVRC, KITTI and CUB-2011 datasets

Key words: graph matching, deep learning

ITERATIVE ESTIMATION APPROACH FOR ERGMS WITH NODAL RANDOM EFFECTS

Sevag Kevork

Ludwig-Maximilians-Universität München (LMU),
Germany

The presence of unobserved heterogeneity in Exponential Random Graph Models (ERGM) is an obvious concern. We extend the well-known Exponential Random Graph Model (ERGM) by including random effects to account for unobserved heterogeneity in the network. This leads to an ERGM with random structure on the coefficients. Estimation is carried out by combining approximate penalized pseudo-likelihood estimation for the random effects with maximum likelihood estimation for the remaining parameters in the model. This allows to fit nodal heterogeneity effects even for large scale networks.

Key words: Exponential random graph models, Random effects, Network analysis

BRANCH DECOMPOSITIONS FOR ANALYZING STOCHASTIC PROCESSES ON NETWORKS

Konstantin Klemm

IFISC,
Spain

Many models of social interaction on a network are based on concepts from Statistical Physics, such as percolation and spin kinetics under pairwise interactions. Sociophysics models thus inherit the hardness of analyzing the dynamics on one specific (quenched) network in exact terms. With frustration in spin interactions, for instance, the computation of ground states is NP-hard, thus unlikely feasible in polynomial time. In practice, the cavity method may provide good approximations by treating networks as tree-like. Here we find, however, that exact computation is possible in many medium-size networks due to existence of a branch decomposition of small width: the edge set of the network may be bipartitioned recursively with only few nodes shared by the two sides of the partition at each level.

Key words: branch decomposition, stochastic process, exact computation

INCORPORATING PRIOR INFORMATION AND BORROWING INFORMATION IN HIGH-DIMENSIONAL SPARSE REGRESSION USING THE HORSESHOE AND VARIATIONAL BAYES

Gino B. Kpogbezan

University of Leiden,
Netherlands

We introduce a sparse high-dimensional regression approach that can incorporate prior information on the regression parameters and can borrow information across a set of similar datasets. Prior information may for instance come from previous studies or genomic databases, and information borrowed across a set of genes or genomic networks. The approach is based on prior modelling of the regression parameters using the horseshoe prior, with a prior on the sparsity index that depends on external information. Multiple datasets are integrated by applying an empirical Bayes strategy on hyperparameters.For computational efficiency we approximate the posterior distribution using a variational Bayes method. The proposed framework is useful for analysing large-scale data sets with complex dependence structures. We illustrate this by applications to the reconstruction of gene regulatory networks and to eQTL mapping.

Key words: high-dimensional Bayesian inference, empirical Bayes, horseshoe prior, variational Bayes approximation, prior knowledge

THE TRADE NETWORK OF FIREARMS AND AMMUNITION: A CENSORED REGRESSION MODEL WITH ENDOGENOUS NETWORK EFFECTS

Michael Lebacher

Ludwig-Maximilians-Universität München,
Germany

In the paper we analyse data on the international trade of firearms and ammunition provided by the Norwegian Initiative on Small Arms Transfers (NISAT). Our focus is on modelling transfers of weapons using the well established gravity model of trade as a baseline. Taking a network based view on the transfers, we do not only rely on exogenous covariates but also estimate linear endogenous network effects. Therefore we include a measure for reciprocity as well as sender- and receiver-effects. As an analytical tool we are using a spatial autocorrelation (SAR) model with multiple weight matrices. The intractable observed likelihood of the censored regression model is maximized employing the Monte Carlo EM algorithm. Our main results are strong and stable endogenous network effects as well as evidence for substantial path dependence. An analysis of the estimated distributions allows for singling out those transfers that have a higher probability to engage in arms trade than observed in the dataset.

Key words: Arms trade network, Censored regression model, Monte Carlo EM algorithm, Maximum likelihood, Network effects, Spatial autocorrelation model

BAYESIAN EXPONENTIAL RANDOM GRAPH MODELS FOR POPULATIONS OF NETWORKS

Brieuc Lehmann

MRC Biostatistics Unit,
University of Cambridge,
UK

While exponential random graph models (ERGMs) were originally developed in the context of social networks, recently there has been significant interest in applying ERGMs to neuroimaging data in both structural and functional connectivity studies. As opposed to social network studies, neuroimaging datasets generally consist of multiple networks across several individuals. This presents an additional layer of complexity - studies to date have circumvented this issue either by combining multiple networks into a single network, or by fitting an ERGM to each network separately. Ideally, however, one would like to pool information across networks to better estimate the model parameters. Motivated by this, we develop establish a novel hierarchical framework for ERGMs. In this presentation, we will also address the challenge posed by the intractability of the likelihood, illustrating our results with both real and simulated data.

Key words: Exponential random graph models, neuroimaging

DIRECT AND INDIRECT EFFECT IN RECURSIVE REGRESSION GRAPH MODELS

Monia Lupparelli

Department of Statistical Sciences,
University of Bologna,
Italy

In linear regression modelling the distortion of effects after marginalizing over variables of the conditioning set has been widely studied in several contexts. For Gaussian variables, the relationship between marginal and partial regression coefficients is well-established and the issue is often addressed as a result of W. G. Cochran. Possible generalizations beyond the linear Gaussian case have been developed, nevertheless the case of discrete variables is still challenging, in particular in medical and social science settings. A multivariate regression framework is proposed for binary data with regression coefficients given by the logarithm of relative risks and a multivariate Relative Risk formula is derived to define the relationship between marginal and conditional relative risks. The method is illustrated through the analysis of the morphine data in order to assess the effect of preoperative oral morphine administration on the postoperative pain relief.

Key words: binary data, graphical models, morphine data, path analysis, relative risk

EXTRACTING KEYWORDS FROM LANGUAGE COMPLEX NETWORKS

Sanda Martincic-Ipsic

University of Rijeka,
Croatia

The task of keyword extraction is to automatically identify a set of terms that best describe the document. We proposed a novel Selectivity-Based Keyword Extraction (SBKE) method, which extracts keywords from the source text represented as a complex network. The node selectivity value is calculated from a weighted network as the average weight distributed on the links of a single node and is used in the procedure of keyword candidate ranking and extraction. The selectivity slightly outperforms an extraction based on the standard centrality measures. Therefore, the selectivity and its modification generalized selectivity as the node centrality measures are included in the SBKE method. Selectivity-based extraction does not require linguistic knowledge as it is derived purely from statistical and structural information of the network and it can be easily ported to new languages and used in a multilingual scenario. The true potential of the proposed SBKE method is in its generality, portability and low computation costs, which positions it as a strong candidate for preparing collections which lack human annotations for keyword extraction.

Key words: keyword extraction, language complex network, selectivity, generalized selectivity

HIGH-DIMENSIONAL BAYESIAN INFERENCE OF GRAPHICAL MODELS WITH APPLICATION TO BRAIN CONNECTIVITY

Reza Mohammadi

University of Amsterdam,
Netherlands

In graphical models, Bayesian frameworks provide a straightforward tool, explicitly incorporating underlying graph uncertainty. In principle, the Bayesian approaches are based on averaging the posterior distributions of the quantity of interest, weighted by their posterior graph probabilities. However, the Bayesian inference has not been used in practice for high-dimensional graphical models, because computing the posterior graph probabilities is hard and the number of possible graph models is very large. In this talk, we discuss the computational problems related to Bayesian structure learning and we offer several solutions to cope with the high-dimensionality problems. We apply our method to high-dimensional fMRI data from brain connectivity studies to show its empirical usefulness. In addition, we have implemented our method in the R packages BDgraph and ssgraph which are available online.

Key words: Graphical models, High-dimensional Bayesian inference, Model averaging, Brain connectivity

GROUP-REGULARIZED LOGISTIC ELASTIC NET REGRESSION: IMPROVED OMICS-BASED CLASSIFICATION

Magnus Münch

Leiden University,
Netherlands

Classification problems are common in omics research. Such problems arise, for example, in the design of diagnostic tests, and the prediction of treatment response. Often external information on the omics features is available. Examples of such information sources are: (a) results on the same genes obtained in a previous study (e.g., p-values), (b) information from a publicly available database that summarizes the prior information on the molecular features involved (e.g., the Cancer Gene Census), (c) omic annotation (e.g., the location of a gene on the chromosome) and (d) previously shown importance of features in related problems. Although such information can rarely be directly included in the statistical analysis, it is often useful and thus has the potential to enhance classification performance.

We propose to include external information by a group-regularized (logistic) elastic net regression algorithm. The groups of features are based on the external information, such that each group of features receives its own penalty parameter. The method makes use of the Bayesian formulation of logistic elastic net regression to estimate both the model and penalty parameters in an approximate empirical-variational Bayes framework. By estimating the group-specific penalty parameters from the data, we avoid a priori (i.e., subjective) specification of feature importance.

Simulation results show that in settings where the grouping of the features is informative, group-specific penalization of the features does indeed enhance classification performance. These findings are confirmed in an application of the method in a cancer omics study.

AIR POLLUTION ANALYSIS WITH HIGHLY CORRELATED WITH METEOROLOGICAL VARIABLES USING GRAPHICAL METHODS

Gamze Ozel

Hacettepe University,
Turkey

Prediction of natural events such as hydrological phenomenon, drought, flood, ground-surface water amount, pollution of water sources, and issues related to air pollution requires an efficient model of precipitation. Precipitation is one of the most indispensable climatological variables. In this study, spatial pattern of Aegean Region in Turkey was shown using precipitation data from 36 meteorological stations obtained by graphical models.

Key words: graphical markov chain, air pollution, precipitation

OBJECTIVE BAYES ANALYSIS FOR GRAPHICAL VECTOR AUTOREGRESSIVE MODELS

Lucia Paci

Universita Cattolica SC,
Italy

Vector Autoregressive (VAR) models are widely used to analyze and forecast multivariate time series. However, the large number of parameters of VAR models can lead to unstable inference and inaccurate forecasts, particularly with many variables. For this reason, restrictions supported by the data are usually required. We propose an objective Bayes factor approach based on graphical VAR models for learning contemporaneous dependencies among variables. We show that, if the covariance matrix at each time is Markov with respect to a decomposable graph, then the likelihood of a graphical VAR can be factorized as an ordinary decomposable graphical model. Additionally, using a fractional Bayes factor approach, we are able to obtain the marginal likelihood in closed form and perform graphical model selection with limited computational burden.

Key words: Bayesian model selection, decomposable graphical model, fractional Bayes factor, multivariate time series

WHICH CENTRALITY MEASURES IN PEER LEARNER NETWORKS ARE THE BEST PREDICTORS OF SECOND LANGUAGE ACQUISITION?

Michał B. Paradowski

Institute of Applied Linguistics,
University of Warsaw,
Poland

We present the results of two studies investigating the influence of peer interaction dynamics and social graph topology among foreign students on measurable outcomes of second language (L2) acquisition. To comprehensively characterise the structure of the learner networks, established metrics were used such as node degree, closeness, betweenness and other centrality measures as well as local clustering coefficients, using generalisations of these metrics to weighted graphs, which allows for both the number and strength of interactions to contribute to the metrics. Additionally, we used community detection algorithms.The topological data were then overlaid over independently collected socio-biographical data and learning outcomes in order to carry out multi-faceted analyses of the mutual influence of individual on social factors and vice versa. To determine how these were associated with performance, we reduced a set of over 100 independent variables with the use of the dimensional reduction, and performed multiple linear regression.Analyses of the various standard centrality measures vis-a-vis performance in a pilot study reveal that the best predictors of (both self-reported and independently measured) progress are closeness and degree centrality, while betweenness and PageRank fail to correlate. This tendency is observed in both objective and subjectively assessed progress in learning. This suggests that for language acquisition via social interaction, it is the structural properties of the network that matter more than processes such as information flow. We will also discuss several other significant findings.Social network analysis provides new insight into the link between social relations and language acquisition, showing how social network configuration and peer interaction dynamics are stronger predictors of L2 performance than individual factors such as attitude or motivation, and offers a novel methodology for investigating the phenomena.

Key words: social network analysis (SNA), centrality measures, cluster detection, community detection, second language acquisition

A NETWORK FRAMEWORK FOR ALGORITHMIC DESIGN OF EXPERIMENTS

Ben Parker

University of Southampton,
UK

We demonstrate that considering experiments in a network framework allows us to exploit automorphisms of the graph to reduce the number of evaluations of candidate designs for those experiments, and thus find optimal designs faster. We show that the use of automorphisms for reducing the number of evaluations required of an optimality criterion function is effective on designs where experimental units have a network structure.Moreover, we show that we can take block designs with no apparent network structure, such as one-way blocked experiments, row-column experiments, and crossover designs, and add block nodes to induce a network structure. Considering automorphisms can thus reduce the amount of time it takes to find optimal designs for a wide class of experiments.

RECONSTRUCTING NETWORKS WITH UNKNOWN AND HETEROGENEOUS ERRORS

Tiago Peixoto

University of Bath,
UK

The vast majority of network datasets contains errors and omissions, although this is rarely incorporated in traditional network analysis. Recently, an increasing effort has been made to fill this methodological gap by developing network reconstruction approaches based on Bayesian inference. These approaches, however, rely on assumptions of uniform error rates and on direct estimations of the existence of each edge via repeated measurements, something that is currently unavailable for the majority of network data. Here we develop a Bayesian reconstruction approach that lifts these limitations by not only allowing for heterogeneous errors, but also for individual edge measurements without direct error estimates. Our approach works by coupling the inference approach with structured generative network models, which enable the correlations between edges to be used as reliable error estimates. Although our approach is general, we focus on the stochastic block model as the basic generative process, from which efficient nonparametric inference can be performed, and yields a principled method to infer hierarchical community structure from noisy data. We demonstrate the efficacy of our approach with a variety of empirical and artificial networks.

Key words: network reconstruction, community structure, network inference

ANALYZING SOCIAL NETWORKS VIA MARGINAL MODELS WITH INDIVIDUAL-SPECIFIC EFFECTS

Stefano Peluso

Universita Cattolica,
Italy

A new modeling framework for social networks arising from a sequence of relational events is proposed. We directly model the joint distribution of the binary variables indicating if each single actor is involved or not in an event. The adopted parametriza- tion is based on first- and second-order effects, formulated as in marginal models for categorical data. In particular, second-order effects are log-odds ratios with meaningful interpretation from the social perspective in terms of tendency to cooperate, in contrast to first-order effects interpreted in terms of tendency of each single actor to participate in an event. These effects are parametrized on the basis of the event times, so that suitable trajectories of individual behaviors may be represented. Inference is based on a composite likelihood function, maximized by an algorithm with numerical complexity proportional to the square of the number of units in the network. A classication composite likelihood is used to cluster the actors, simplifying the interpretation of the data structure. The proposed approach is illustrated on a dataset of scientific articles published in four top statistical journals from 2003 to 2012.
cluster analysis, composite likelihood, latent trajectories, log-odds ratios, social networks

GENERALIZATION OF AZZALINI'S METHOD WITH APPLICATION IN MEDICINE SCIENCE

Božidar Popović

Faculty of Science and Mathematics,
University of Montenegro,
Montenegro

In the real life there are many situations where symmetric distributions are not appropriate for fitting. In that way, Azzalini's method (Azzalini 1985) is widely used to get asymmetric or skewed distribution. Shortly this technique can be described as follows: a random variable Z possesses the skew normal distribution with parameter α if its probability density function (pdf) is $f(z;\alpha) = 2\Phi(\alpha z)\phi(z)$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are cumulative density function (cdf) and pdf, respectively, of the standard normal random variable. In terms of the conditional distribution Azzalini's method consists of the following: let X_i $i = 1, 2$ are independent random variables with pdfs $f_i(x_i)$ and cdfs $F_i(x_i)$. Then the conditional pdf of $X = X_1$ given $\alpha X_1 > X_2$ is

$$f_X(x) = \frac{f_1(x)F_2(\alpha x)}{P\{\alpha X_1 > X_2\}}, \tag{1}$$

where $P\{\alpha X_1 > X_2\} = \int_0^\infty \left[\int_0^{\alpha x_1} f_2(x_2)dx_2\right] f_1(x_1)dx_1 = E_{X_1}(F_2(\alpha x))$. The Eq.(1) can be interpreted as weighted distribution with weight $\omega = F_2(\alpha x)$. Described method has been used to construct new skewed distributions from a given symmetric distribution, for example, skew-t (Gupta et al. 2002), skew-Cauchy (Arnold and Beaver 2000), skew-Laplace (Aryal and Nadarajah 2005) and skew-logistic (Nadarajah 2009). The aim of this note is to generalize Azzalini's method such that it is applicable in case when X_i $i = 1, 2$ are dependent random variables. Dependence can be modelled using copula function. In case of dependence of the random variables X_i $i = 1, 2$ the equation (1) becomes

$$f_X(x) = \frac{f_1(x) \int_{-\infty}^{\alpha x} c(F_1(x_1), F_2(x_2); \theta) f_2(x_2)dx_2}{P\{\alpha X_1 > X_2\}},$$

where $P\{\alpha X_1 > X_2\} = E_{X_1}\left[\int_{-\infty}^{\alpha X_1} c(F_1(X_1), F_2(X_2); \theta) f_2(x_2)dx_2\right]$ and $c(\cdot)$ is the copula function. Some statistical properties will be derived together with some particular cases. These results are applied on two biochemical data sets: C- reactive protein and insulin.

Key words: Generalization of Azzalini's method, Application in medical science

CONSTRUCTION OF PROTEIN-PROTEIN INTERACTION NETWORKS UNDER STEADY-STATE ACTIVATIONS

Vilda Purutcuoglu

Department of Statistics,
Middle East Technical University (METU),
Turkey

The construction of protein-protein interaction (PPI) networks can be described by various models from Boolean networks to diffusion models. Among these alternatives, the representation of steady-state activation of the systems is the most common one since majority of the data are available for this type of modelling. In this study, we apply one of the well-known methods in this branch, called the Gaussian graphical model, by suggesting the open form of the model parameters without using the penalized likelihood approaches. We derive the modified maximum likelihood estimators of the model parameters. Furthermore, we assume the family of long-tailed symmetric (LTS) distributions, rather than strict normality assumption, for the presentation of the states. By this way, we can cover a wider range of densities from Cauchy, student-t to normal. Moreover, we suggest the application of alternative model selection criteria, namely, ICOMP, CAIC, CAICF, in place of RIC, EBIC and STARS, which are the common model selection rules used with GGM to construct PPI networks. The proposal criteria have been already implemented in different types of regression models, whereas, they have not yet performed in the estimation of PPI networks. In this study, we derive these expressions for GGM and graphical model under LTS distributions. Finally, we evaluate the performance of our suggested approaches in terms of accuracy and computational time by applying in real datasets and Monte Carlo studies.

Key words: Protein interaction networks, frequentist inference, model selection criterion

OVERLAPPING MIXTURE MODELS FOR NETWORK DATA (MANET) WITH COVARIATES ADJUSTMENT

Saverio Ranciati
Giuliano Galimberti

Department of Statistical Sciences,
University of Bologna,
Italy

Ernst C. Wit

Johann Bernoulli Institute for Mathematics and Computer Sciences,
University of Groningen,
Netherlands

Veronica Vinciotti

Department of Mathematics,
Brunel University,
UK

Network data often come in the form of actor-event information, where two types of nodes comprise the very fabric of the network. Examples of such networks are: people voting in an election, users liking/disliking media content, or, more generally, individuals - actors - attending events. Interest lies in discovering communities among these actors, based on their patterns of attendance to the considered events. To achieve this goal, we propose an extension of the model introduced in [Ranciati et al. 2017]: our contribution injects covariates into the model, leveraging on parsimony for the parameters and giving insights about the influence of such characteristics on the attendances. We assess the performance of our approach in a simulated environment.

Key words: Bayesian inference, bimodal network, MCMC, probit regression

RISK AND RUIN IN AN INSURANCE MARKET WITH BIPARTITE NETWORK STRUCTURE

Gesine Reinert

Department of Statistics,
University of Oxford,
UK

Anita Behme

Technische Universität Dresden,
Germany

Claudia Klüppelberg
Oliver Kley

Technische Universität München,
Germany

In this talk we introduce a bipartite network model for an insurance market. Using Pareto-tailed losses and multivariate regular variation we obtain asymptotic results for systemic conditional risk measures based on the Value-at-Risk and the Conditional Tail Expectation. These results allow us to assess the influence of an individual institution on the systemic or market risk and vice versa through a collection of conditional systemic risk measures. Further we give results for ruin probabilities in this setting, which involves introducing a network Pollaczek-Khintchine constant. This talk is based on joint papers with Anita Behme, Oliver Kley and Claudia Klüppelberg.

Key words: bipartite network, insurance market, systemic risk

MODEL-BASED CLUSTERING FOR POPULATIONS OF NETWORKS

Mirko Signorelli

Department of Biomedical Data Sciences,
Leiden University Medical Center,
Netherlands

Ernst C. Wit

Johann Bernoulli Institute for Mathematics and Computer Sciences,
University of Groningen,
Netherlands

We propose a model-based clustering method for populations of networks that describes the joint distribution of a sequence of networks in a parsimonious manner, and can be used to identify subpopulations of networks that share certain topological properties of interest. We discuss how maximum likelihood estimation can be performed with the EM algorithm and study the performance of the proposed method on simulated data. We conclude with an example application to a sequence of face-to-face interaction networks measured in an office environment. Joint work with Ernst Wit. Preprint available on arXiv: arXiv:1806.00225.

BAYESIAN AND SPLINE BASED APPROACHES FOR GRAPHON ESTIMATION

Benjamin Sischka

Ludwig-Maximilians-Universität München (LMU),
Germany

In our method we propose to use (linear) B-splines for graphon estimation. This allows to easily accomodate constraints in the estimation routine so that the estimated graphon fulfills the canonical representation, meaning its univariate margin is monotonic. The estimate allows to apply Bayesian ideas in that we can explore the degree distribution and the ordering of the nodes with respect to their degree. Given that the degree of a node is random, we can investigate the variability and uncertainty using MCMC techniques.

Key words: graphon, B-spline regression, Bayesian method, MCMC, Gibbs sampling

TRIADIC CLOSURE PATTERN ANALYSIS IN TEMPORAL NETWORK MODELS

Ruta Užupytė

Vilnius University,
Lithuania

Temporal networks can be used to model a huge diversity of complex systems that evolve in time. In this case dynamics refers to a not static structure of connections as the links between objects dynamically change over time, while the set of network nodes remains fixed. Our research presents a possible framework for the analysis of longitudinally observed networks. The proposed approach is based on the assumption of statistical dependence between observations according to a continuous time Markov process and method of moments. Moreover, we are assuming that the edge dynamics involve a term that quantifies a triadic closure effect - friends of friends tend to become friends.

Key words: Temporal networks, Network dynamics, Triadic closure

IMPROVING PROTOCOLS FOR CELL DIFFERENTIATION THROUGH REINFORCEMENT LEARNING

Lorenz Wernisch

University of Cambridge
UK

Great progress has been made in generating mature cells and tissues from human pluripotent stem cells (hPSCs) for clinical applications, for example, in producing platelets for blood transfusion from megakaryocytes (MKs). Forward programming, the induction of specific transcription factors related to cell development such as GATA1, has been successfully applied to produce MKs from hPSCs. However, they retain undesirable characteristics of stem cells and their ability to produce platelets is limited. In preparation for experiments optimising the experimental protocol we explore how far reinforcement learning can guide such experiments in real time. Over a time course of several days cell markers indicating the maturity of cells can be monitored and the protocol adapted. In simulations we test reinforcement algorithms based on Gaussian process state space models and current knowledge of relevant gene regulatory networks for their ability to suggest suitable experimental interventions.

COMMFINEDWALKER: CONTROLLING FOR INSPECTION BIAS IN THE ANNOTATION OF PROTEIN INTERACTION NETWORKS

James Wilsenach

University of Oxford,
UK

Community detection on Protein Interaction Networks (PINs) can help biomedical scientists to identify novel functional components within the proteome. Filtering communities by their functional homogeneity to obtain functionally cohesive communities has been shown to bias community selection towards regions of the protein network that are better understood and well annotated with functional data. CommWalker is a recently developed tool designed to mitigate some of this bias by repeatedly sampling a communitys regional functional context. The method filters communities by comparing a communitys functional profile to its local annotation background. CommWalker is able to pick out larger numbers of communities from less well annotated parts of the network. As CommWalker uses random walks, there is no guarantee that the area explored by it remains local. We present CommFinedWalker, an augmented approach in which walks are confined to a specified regional context. This approach is able to detect even more communities from potentially under-researched network regions. This was validated by analysing the ability of CommFinedWalker to capture co-expressionally enriched communities including those from annotationally poor network areas.

Key words: Protein Interaction Networks, Gene Ontology, Annotated Networks, Inspection Bias, Bias Correction, Functional Module Detection, Community Detection

CAN FRIENDS AND FAMILY REPLACE THE CHILDCARE SERVICES? PRELIMINARY ANALYSIS FOR POLAND

Olga Zajkowska

Faculty of Applied Informatics and Mathematics,
Warsaw University of Life Sciences,
Poland

Similarly to other CEE countries Poland has lowest low fertility rates. At the same time there is not enough child care services and far to little services for elderly care. Based on the GGS data we discuss if lack of formal institutions can be substituted by informal ones- extended family, friends or neighbors. We come up with rather non optimistic conclusions and some policy implications.

Key words: fertility, parental leaves, family support, family ties

HOW TO IMPLEMENT THE NEAREST NEIGHBOURS ALGORITHM TO ESTIMATE MISSING TIES DUE TO ACTOR NON-RESPONSE?

Anja Žnidaršič

Faculty of Organizational Sciences,
University of Maribor,
Slovenia

Patrick Doreian

Department of Sociology,
University of Pittsburgh,
USA,
Faculty of Social Sciences
University of Ljubljana
Slovenia

Anuška Ferligoj

Faculty of Social Sciences,
University of Ljubljana,
Slovenia,
NRU HSE Moscow,
Russia

Networks are often used to represent a real-world dynamics of relations among units. One of the classifications of networks could be made based on the measurement level of ties among units. The network could be binary, where only a presence of a tie is recorded, in the valued network the magnitude of tie value is taken into account, while in signed networks the distinction between positive and negative relationships is important. Social network data are usually gathered by surveys and therefore prone to measurement errors. Missing network data could appear in several forms, here we will focus on missing ties due to non-response or simply item non-response. Item non-response appears when an actor participate in a survey, but forget or deny to respond on particular items in the research and a consequence is a row with few missing entries in a matrix representation of a network.

One of the best treatments in case of actor non-response is the k-nearest neighbours approach which searches for k closest actors according to their incoming ties and then calculates the outgoing ties of the non-respondent as a median of outgoing ties of selected nearest neighbours. Here we will try to determine if that treatment performs well also in the case of item non-response and what is the number of appropriate nearest neighbours to impute missing ties. Simulation results based on several real networks and different blockmodeling approaches will be presented.

Key words: missing network data, item non-response, k-nearest neighbours, imputations

TOTALLY POSITIVITY, GRAPHICAL MODELS, AND CONVEX OPTIMIZATION

Piotr Zwiernik

Universitat Pompeu Fabra,
Spain

Probability distributions that are multivariate totally positive of order 2 (MTP2) appeared in the theory of positive dependence and in statistical physics through the celebrated FKG inequality. The MTP2 property is stable under marginalization, conditioning and it appears naturally in various probabilistic graphical models with hidden variables. Models of exponential families with the MTP2 property admit a unique maximum likelihood estimator. In the Gaussian case, the MLE exists also in high-dimensional settings, when $p >> n$, and it leads to sparse solutions. The main aim of this talk is to given an idea of what the MTP2 condition is as well as to show how total positivity becomes useful in graphical modelling.

Key words: Gaussian graphical models, latent variables, total positivity