College of Letters and Science
Department of Statistics
UC Davis

Statistical Sciences Symposium, 2015

Network Data : Information and Sciences

FRIDAY / SATURDAY

April **10-11**

ROOM AGR, WALTER A. BUEHLER ALUMNI CENTER
University of California, Davis

Hosted by the Department of Statistics at the University of California, Davis. Co-sponsored by the Institute of Mathematical Statistics, the MPS Dean's Office of UC Davis, and the National Science Foundation via an RTG training grant.

KEYNOTE SPEAKERS:

Raissa D'Souza (Mechanical & Aeronautical Engineering, UC Davis)

John Doyle (Electrical Engineering & Bioengineering, Cal Tech

http://www.stat.ucdavis.edu/symposium2015

UCDAVIS

UC Davis Statistical Sciences Symposium 2015

Network Data : Information and Sciences

Location: Room AGR, Buehler Alumni Center, UC Davis

Friday (April 10)

12:30pm-01:00pm: ***Registration***

01:00pm-01:15pm: ***Welcome and Opening Remarks***
Hans-Georg Müller, Chair of the Department of Statistics

01:15pm-02:05pm: Raissa D'Souza (Mechanical and Aerospace Engineering, UC Davis)

02:05pm-02:40pm: Eric Xing (Computer Science, Carnegie Mellon U)

Break

03:10pm-03:45pm: Mark Goldman (Neuroscience, UC Davis)

03:45pm-04:20pm: Andrew Sornborger (Mathematics, UC Davis)

04:20pm-04:55pm: Fushing Hsieh (Statistics, UC Davis)

Break

06:00pm-9:00pm: ***Social Event (food and beverages will be served)***

Saturday (April 11)

8:15am-08:45am: ***Registration***

08:45am-09:35am: John Doyle (Electrical Engineering and Bioengineering, Caltech)

09:35am-10:10am: Martin Hilbert (Communication, UC Davis)

10:10am-10:45am: Zeev Maoz (Political Science, UC Davis)

Break

11:10am-11:45am: Ji Zhu (Statistics, U of Michigan)

11:45am-12:20pm: Michael Schweinberger (Statistics, Rice U)

Lunch

02:00pm-02:35pm: Vladimir Filkov (Computer Science, UC Davis)

02:35pm-03:10pm: Hao Chen (Statistics, UC Davis)

03:10pm-03:45pm: Yueyue Fan (Civil and Environmental Engineering, UC Davis)

Break

04:10pm-04:45pm: Brenda McCowan (Population Health and Reproduction, UC Davis)

04:45pm-05:20pm: Mark Lubell (Environmental Science and Policy, UC Davis)

05:20pm-05:30pm: ***Closing Remarks***

Hao Chen

Title: A new graph-based test for multivariate and object data

Abstract: Two-sample tests for multivariate and especially for non-Euclidean data are not fully explored. This paper presents a novel test based on a similarity graph constructed on the pooled observations from the two samples. It can be applied to multivariate data and non-Euclidean data as long as a dissimilarity measure on the sample space can be defined, which can usually be provided by domain experts. Existing tests based on a similarity graph lack power either for location or for scale alternatives. The new test utilizes a common pattern that was overlooked previously, and works for both types of alternatives. The test exhibits substantial power gains in simulation studies. Its asymptotic permutation null distribution is derived and shown to work well under finite samples, facilitating its application to large data sets. The new test is illustrated on two applications: The assessment of covariate balance in a matched observational study, and the comparison of network data under different conditions.

Raissa D'Souza

Title: Networks at the core of our complex, independent world

Abstract: Collections of networks are at the core of modern society, spanning technological, biological and social systems. Examples include transportation networks, electric power grids, online social networks, and gene-regulatory networks, to name just a few. Different classes of networks are subject to different constraints, such as geographical considerations, costs, flows, and timescales. Likewise, different networks were designed or evolved to fulfill different functions. Yet, are there underlying principles to discover? Furthermore, we know that each network on its own is a complex system shaped by the collective action of individual agents, and we are becoming increasingly aware that interactions between distinct networks can lead to unanticipated consequences, such as cascading failures and novel phase transitions. Is there a principled, scientific approach for analyzing consequences of interdependence?

Over the past decade a "science of networks" has been emerging which blends ideas from statistical physics, applied math, computer science and the social sciences. Here I will provide a brief overview of the field, highlighting in particular how mathematical models of random graphs allow us to understand aspects of real-world systems. These models give insights into phenomena such as percolation, self-organized tradeoffs, and synchronization. Of course applying the theoretical results to real-systems remains a challenge. The details depend on the data and the environmental context. This opening talk will set the stage for the dialogue to come over the next 1.5 days, as we hear from leading experts across the UC Davis campus and beyond on their studies of more domain specific aspects of networks.

John Doyle

Title: Universal laws and architectures: theory and lessons from hearts, bugs, brains, nets, grids, flows, and zombies

Abstract: The objectives of this talk are to 1) accessibly introduce a new theory of network architecture relevant to biology, medicine and technology, 2) illustrate the key ideas with familiar examples from neuroscience, including live demos using audience brains, and 3) highlight persistent errors and confusion in science regarding statistical issues that need your help. The most accessible background material is in a recent blog post, most importantly including links to online videos: rigorandrelevance.wordpress.com/author/doyleatcaltech.

My research is aimed at developing a more "unified" theory for complex networks motivated by and drawing lessons from neuroscience[4], cell biology [3], medical physiology [9], technology (internet, smartgrid, sustainable infrastructure)[1][8], and multiscale physics [2],[5],[6]. This theory involves several elements: hard limits, tradeoffs, and constraints on achievable robust performance ("laws"), the organizing principles that succeed or fail in achieving them ("architectures" and protocols), the resulting high variability data and "robust yet fragile" behavior observed in real systems and case studies (behavior, data, statistics), and the processes by which systems adapt and evolve (variation, selection, design). We will leverage a series of case studies with live demos from neuroscience, particularly vision and sensorimotor control, plus some hopefully familiar and simple insights from cell biology and modern computer and networking technology. One warning is that exposure can cause a rare and contagious form of synesthesia, but only if you pay careful attention. Time permitting, there are multiple contacts with zombies, their peculiar cultural ubiquity, and growing threat of zombie science (huge, grotesque errors that won't die). In addition to the above mentioned blog and videos, papers [1]and [4] (and references therein) are the most accessible and broad introduction while the other papers give more domain specific details.

Selected recent references:

[1] Alderson DL, Doyle JC (2010) Contrasting views of complexity and their implications for network-centric infrastructures. *IEEE Trans Systems Man Cybernetics—Part A: Syst Humans* 40:839-852.

[2] Sandberg H, Delvenne JC, Doyle JC. On Lossless Approximations, the Fluctuation-Dissipation Theorem, and Limitations of Measurements*, IEEE Trans Auto Control*, Feb 2011

[3] Chandra F, Buzi G, Doyle JC (2011) Glycolytic oscillations and limits on robust efficiency. *Science*, Vol 333, pp 187-192.

[4] Doyle JC**,** Csete ME(2011) Architecture, Constraints, and Behavior, *P Natl Acad Sci USA,* vol. 108, Sup 3 15624-15630

[5] Gayme DF, McKeon BJ, Bamieh B, Papachristodoulou P, Doyle JC (2011) Amplification and Nonlinear Mechanisms in Plane Couette Flo*w, Physics of Fluids*, V23, Issue 6, 065108

[6] Page, M. T., D. Alderson, and J. Doyle (2011), The magnitude distribution of earthquakes near Southern California faults, *J. Geophys. Res*., 116, B12309, doi:10.1029/2010JB007933.

[7] Namas R, Zamora R, An, G, Doyle, J et al, (2012) Sepsis: Something old, something new, and a systems view, *Journal Of Critical Care* Volume: 27 Issue: 3

[8] Chen, L; Ho, T; Chiang, M, Low S; Doyle J,(2012) Congestion Control for Multicast Flows With Network Coding, *IEEE Trans On Information Theory* Volume: 58 Issue: 9 Pages: 5908-5921

[9] Li, Cruz, Chien, Sojoudi, Recht, Stone, Csete, Bahmiller, Doyle (2014) Robust efficiency and actuator saturation explain healthy heart rate control and variability, PNAS 2014 111 (33) E3476-E3485; published ahead of print August 4, 2014

Yueyue Fan

Title: Stochastic origin-destination demand estimation in traffic networks: Information fusion via stochastic programming

Abstract: Travel demand in a traffic network, often expressed using an origin-destination (O-D) demand matrix, is one of the most critical input needed for transportation system planning and management. Due to the high cost involved in directly measuring O-D demand, there is a strong interest in estimating O-D demand based on directly measurable local information such as link traffic counts. In this talk, we will present a two-stage stochastic programming model for estimating the statistical properties of the O-D demand based on multiple observation sets of link counts. We will also discuss how this O-D estimation problem may be connected to general network identification problems with broader applications.

Vladimir Filkov

Title: Task Networks in Open Source Software Projects

Abstract: How do distributed groups of people self-organize to produce artifacts of high quality, like OSS, Wikipedia, etc.? Are earned trust and spontaneous cooperation essential determinants of such systems in the absence of rigid organizational structure? Over the past 3 years, we have adopted a task-oriented, social network perspective to study organization and productivity in OSS projects.
To do so we were relying on trace data of interleaved social and technical activities of developers, gathered from software repositories. We had to develop robust data analytic methods, rooted in network science and statistical modeling, that could handle heterogeneous data sets spanning many years in time, and having large amount of variance. Here, I will talk about the methods we developed and our results from three separate thrusts of this work: (1) on measuring contributions among OSS developers, (2) on collaboration in task networks, and (3) on the temporal congruence of OSS developer activities and the software call-graph structure.

Mark Goldman

Title: Inferring the features of network connectivity governing the dynamics of a brain memory circuit

Abstract: Networks of neurons in the brain are wired together through a myriad of synaptic connections that collectively determine network function. This poses a serious challenge to understanding brain network function, as experimentally measuring the strengths of large numbers of synaptic connections is prohibitively difficult. Furthermore,

because the dynamics of the network responses are typically much lower dimensional than the number of synaptic connection parameters, the possible network connectivities consistent with the observed neuronal responses can be highly degenerate. Here, we address the challenge of inferring the features of network connectivity that are critical to the performance of a brain network. We consider a network of neurons in the brain stem that mathematically integrates (in the sense of Calculus) its inputs and, in the absence of input, maintains a memory of the running total of its previous inputs. By combining data from anatomy, neuronal recordings, and network perturbations, we infer features of network connectivity essential to the network's memory function. This work suggests a simple framework for inferring features of large networks that are, and are not, critical to network function.

Martin Hilbert

Title: Curses and Blessings of a Data-Complete Science: Big Data and the Social Sciences

Abstract: Big Data has turned the social sciences from a traditionally data-poor science into arguably the most data-complete science to date – and this basically "overnight". With over 99% of all of human kinds' technologically mediated information in digital format, and a mobile penetration of 98% worldwide, the digitalization of human interaction produces an impressively detailed digital footprint of everything that's relevant for the social sciences. Each and every digital communication inevitably leaves a trace that can be analyzed to better understand and influence social conduct. This renders many traditional survey and data collection and production processes obsolete. While creating unprecedented opportunities for private actors and lots of low hanging fruits for academic research, it also creates challenges that call for a profound paradigm shift in our relation to data.

Fushing Hsieh

Title: Geometries in wine-tasting sensory bipartite networks.

Abstract: First we address the question: Is human aroma sensory symmetric? Based on a bipartite network derived from all 15 panels' aroma calibrations exams, we construct a coupling geometry that is framed by two coupled Ultrametric trees on targeted-aroma-axis and identified-aroma-axis. These two Ultrametric trees are compared with one tree based on symmetrized version and the popular Davis-Wine-Aroma wheels. Secondly we address the question: How to discover a panel's wine-tasting geometry? We compute a coupling geometry from each individual wine-tasting panel's bipartite network. The individual panel's ultrametric tree on 72 wines is then compared with an ultrametric tree of wine derived based on 13 biochemical characteristics. This comparison is based on partial coupling geometry and a network bootstrapping based energy distribution. The

primary goal of such a comparison is intended to help individual panel to discover his/her own potential biases on aroma recognitions.

The scheduled speaker **Mark Lubell** will be replaced by **Michael Levy**.

Title: Innovation, Cooperation, and the Structure of Agricultural Information Networks

Abstract: Social networks are a key medium through which farmers acquire information. The diffusion of innovations has been the dominant model in agricultural policy for nearly a century, but there are likely other important functions of social networks in farmer decision making. The benefits of many agricultural practices are conditional on the actions of others actors (e.g. many pest management and sustainability practices), so the adoption of practices can usefully be thought of as cooperation games. In such settings, there are incentives for individuals to monitor the practices of others. Here, we use exponential random graph models (ERGMs) to identify social network structures that help solve cooperation and innovation challenges. We argue that closed structures, epitomized by triadic closure, evolve to solve cooperation dilemmas while open structures, epitomized by network centralization, evolve to solve innovation problems. We test this in three vineyard management information networks. Closed triangles were present in all three networks at levels similarly above chance. This facilitates social monitoring and helps solve the cooperation dilemma aspects of agricultural management. One network was significantly more centralized than the others, and this network has had the greatest adoption of sustainable practices, suggesting that bridging and bonding social capital must be present together for sustainable agriculture to spread. Outreach professionals who are themselves growers, but not those who aren't, have extremely high betweenness centrality in all three networks, suggesting that they play a critical role spanning boundaries between communities.

Zeev Maoz

Title: Political Networks: Using Domestic and International Political Interactions in Longitudinal Network Processes

Abstract: This paper presents actual and potential use of political data in network research. It introduces a number of datasets on domestic political processes and international relations that have and can be used in longitudinal network research. Domestic datasets include bill co-sponsorships and roll call voting in congress or manifestos of political parties in multiparty democracies. International relations data include interactions between states such as alliances, trade, conflict, and membership in international organizations. These networks offer opportunities to study network evolution and co-evolution, the propagation of shocks and network re-organization following shocks, and multiplexes. I illustrate some of the studies employing these networks.

The scheduled speaker **Brenda McCowan** will be replaced by **Brianne Beisner** from Dept. of Population Health and Reproduction, School of Veterinary Medicine, UC Davis

Title: Characterizing Social Stability from Network Dynamics

Abstract: Stability is a system-level property that emerges from the synergistic interaction amongst the multiple component networks of the system, and among animal societies, a cohesive social group is one example of a stable social system. Network approaches are uniquely suited for identifying emergent, higher-order properties of a system. Here, we take a network approach to characterizing social stability by examining the power structure across multiple nonhuman primate social groups as well as the joint network relationship between the power structure and a second key network. We present two primary features that distinguish stable social groups from unstable ones, and couch our findings around the broader utility of using network dynamics to characterize the stability and/or health of any complex system.

Michael Schweinberger

Title: Exponential-family random graph models with local dependence

Abstract: Dependent phenomena, such as relational, spatial, and temporal phenomena, tend to be characterized by local dependence in the sense that units which are close in a well-defined sense are dependent. However, in contrast to spatial and temporal phenomena, relational phenomena tend to lack a natural neighborhood structure in the sense that it is unknown which units are close and thus dependent. An additional complication is that the number of observations is 1, which implies that the dependence structure cannot be recovered with high probability by using conventional high-dimensional graphical models. Therefore, researchers have assumed that the dependence structure has a known form. The best-known forms of dependence structure are inspired by the Ising model in statistical physics and Markov random fields in spatial statistics and are known as Markov random graphs. However, owing to the challenge of characterizing local dependence and constructing random graph models with local dependence, conventional exponential-family random graph models with Markov dependence induce strong dependence and are not amenable to statistical inference.

We take first steps to characterize local dependence in random graph models and show that local dependence endows random graph models with desirable properties which make them amenable to statistical inference. We show that random graph models with local dependence satisfy a natural domain consistency condition which every model should satisfy, but conventional exponential-family random graph models do not satisfy. In addition, we discuss concentration of measure results which suggest that random graph models with local dependence place much mass in the interior of the

sample space, in contrast to conventional exponential-family random graph models. We discuss how random graph models with local dependence can be constructed by exploiting either observed or unobserved neighborhood structure. In the absence of observed neighborhood structure, we take a Bayesian view and express the uncertainty about the neighborhood structure by specifying a prior on a set of suitable neighborhood structures. We present simulation results and applications to two real-world networks with ground truth.

Andrew Sornborger

Title: Mapping Functional Connectivity in Large-Scale Neuronal Networks

Abstract: Optical imaging of calcium indicators is an important adjunct to electrophysiology and is widely used to visualize neuronal activity. The ability to map functional connectivity and changes in functional connectivity is necessary for the study of the flow of activity in neuronal circuits. In this talk, I will discuss issues in the analysis of large neural imaging datasets and a procedure to rapidly compute reduced functional couplings between neuronal ensembles from calcium imaging data. This procedure provides a practical and user-independent means for summarizing the flow of activity between neuronal ensembles with the added benefit of distinguishing putative excitatory and inhibitory populations of neurons.

Eric Xing

Title: Reverse Engineering Evolving Networks Underlying Developing Systems

Abstract: Estimating rewiring gene regulatory networks over developing biological systems, such as proliferating cells, growing embryos, and differentiating cell lineages, is central to a deeper understanding of how cells evolve during development. However, one challenge in estimating such evolving networks is that their host cells are not only contiguously evolving, but also branching over time. For example, stem cells evolve into two more specialized daughter cells at each division, forming a tree of networks. Another example is in a laboratory setting: a biologist may apply several different drugs to a malignant cancer cell to analyze the changes each drug has produced in the treated cells. Each treated cell is not directly related to another treated cell, but rather to the malignant cancer cell that it was derived from. This posts an interesting question on multiplicity control.

We propose two interesting statistical frameworks, one builds on a generalization of kernel density estimator for regularized nonparametric estimation of inverse-covariance matrix from non-iid high-dimensional samples, and the other one builds on the L1 plus total variation penalized graphical logistic regression, to effectively estimate multiple evolving gene networks corresponding to cell types related by either a linear-sequence or a tree-genealogy, based on only a few samples from each cell type. Our methods take advantage of the similarity between related networks along the biological

lineage, while at the same time exposing differences between the networks. We demonstrate that our methods perform significantly better than existing methods via simulation, and enjoy strong statistical guarantees unlike other heuristic based approaches. We explore an application to a breast cancer analysis. Based on only a few microarray measurements, our algorithm is able to produce biologically valid results that provide insight into the progression and reversion of breast cancer. Finally, I will discuss a few additional complex scenarios for network estimation, where graphs are directional, have missing value, or are multi-attributes, and ideas for consistent structure estimation.

Ji Zhu

Title: Detecting Overlapping Communities in Networks Using Spectral Methods

Abstract: Community detection is a fundamental problem in network analysis. In practice, communities often overlap, which makes the problem more challenging. Here we propose a general, flexible, and interpretable generative model for overlapping communities, which can be viewed as generalizing several previous models in different ways. We develop an efficient spectral algorithm for estimating the community memberships, which deals with the overlaps by employing the K-medians algorithm rather than the usual K-means for clustering in the spectral domain. We show that the algorithm is asymptotically consistent when networks are not too sparse and the overlaps between communities not too large. Numerical experiments on both simulated networks and many real social networks demonstrate that our method performs well compared to a number of benchmark methods for overlapping community detection. This is joint work with Yuan Zhang and Elizaveta Levina.