CMSS Social Network Data: Collection and Analysis Workshop

October 21-23, 2013

SPEAKER TITLES/ABSTRACTS

Bruce Desmarais
University of Massachusetts

"Partitioned Multinetwork Embeddings"

We introduce a joint model of network content and context designed for exploratory analysis of email networks via visualization of topic-specific communication patterns. Our model is an admixture model for text and network attributes that uses multinomial distributions over words as admixture components for explaining email text and latent Euclidean positions of actors as admixture components for explaining email recipients. This model allows us to infer topics of communication, a partition of the overall network into topic-specific subnetworks, and two-dimensional visualizations of those subnetworks. We validate the appropriateness of our model by achieving state-of-the-art performance on a prediction task and semantic coherence comparable to that of latent Dirichlet allocation. We demonstrate the capability of our model for descriptive, explanatory, and exploratory analysis by investigating the inferred topic-specific communication patterns of a new email data set, the New Hanover County email corpus.

Elena Erosheva
University of Washington

"Asking Questions about Numbers: Practical Considerations in RDS Degree Measurement"

A network-based type of sampling technique and the corresponding set of estimates, known as Respondent-Driven Sampling (RDS), is the current method of choice for many researchers studying hard-to-reach populations. RDS exploits social networks by starting with a small set of individuals and allowing the respondents at each wave to recruit the next wave of the sample from their contacts. However, it is often the case that little is known about networks in target populations. An overarching aim of formative RDS research is to assess appropriateness and feasibility of RDS studies in a specific population, including whether important assumptions of RDS estimators about the population-specific network structure are satisfied, before carrying out full-scale RDS studies. Measurement of respondents' network degree is a cornerstone in both formative RDS research and RDS studies. For example, Volz-Heckathron estimator relies on weighting by respondents' degree. In this talk, I outline a number of practical considerations related to measurement of respondents' degree in RDS studies.

Jacob Foster
UCLA

"Cultural Enrichment: Linking Structure to Culture in Network Analysis"

Most network analyses focus on the complex web of relations between actors of a single type. Authors are linked by co-authorship; papers, by citation. In many cases, however, data are readily available concerning not just the structure but the content of those relationships. What topics are co-authors actually writing about? Do literatures linked by citation talk about the same thing?

In this talk, I will argue that attending to content and culture can enrich network analysis. I will focus on recent work using information theory to model communicative efficiency in scholarship, revealing cultural holes that cut across patterns of citation. I will briefly touch on two other examples: first, showing how the performance of teams of biomedical scientists is shaped by the "teams" of chemicals and topics they study; and second, exploring how authors, chemicals, diseases, and methods use each other to forge new connections.

Krista Gile
University of Massachusetts

"Inference from Link-Tracing Network Samples"

It is often the case that a population of interest is connected by a network of relations, and that it is beneficial to exploit this network in the sampling process. This typically involves a form of link-tracing sampling, in which subsequent sample units are selected from among the network neighbors of earlier samples.

Although the various link-tracing sampling designs have much in common, the foundational assumptions and approaches of existing inferential strategies vary widely. Inference is also affected by the selection procedure for the initial sample, specifics of the link-tracing process, and other information available about the population.

In this talk, we present a conceptual review of classical and recent approaches to inference from link-tracing network samples, highlighting the foundational assumptions required by the methods and their implications for inference. We review the current state of research and outstanding issues.

Eric Kolaczyk
Boston University

"Estimating Network Degree Distributions from Sampled Networks: An Inverse Problem"

Networks are a popular tool for representing elements in a system and their interconnectedness. Many observed networks can be viewed as only samples of some true underlying network. We study the problem of how to estimate the degree distribution of a true underlying network from its sampled network, under various common network sampling designs. We show that it can be formulated as an inverse problem that is, in many cases, ill-posed. Accordingly, we offer a

penalized least-squares approach to solving this problem, with the option of additional constraints. The resulting estimator is a linear combination of singular vectors of a matrix, relating the expectation of our sampled degree distribution to the true underlying degree distribution, which is defined entirely in terms of the sampling plan. Choice of the penalization parameter is made through a Monte Carlo version of Stein's unbiased risk estimation. We present the results of a simulation study, characterizing the performance of our proposed method, and we illustrate its use in the context of monitoring large-scale social media networks.

Tyler McCormick
University of Washington

"Latent Space Models for Multiview Network Data"

Most social relationships consist of multiple dimensions of interaction, with actors forming multiple types of relationships with the same alters (an actor will not trust all of his/her acquaintances, for example). We present statistical models for data consisting of multiple relationships measured on the same set of actors. Our approach builds on work on latent space models for networks where the propensity for two individuals to form ties is conditionally independent given the distance between the individuals in an unobserved social space. Drawing on recent work on the multivariate Bernoulli distribution, we propose a model that parsimoniously represents dependence between relationship types while also maintaining enough flexibility to allow individuals to serve different roles in different relationship types. The approach also naturally yields pairwise representation through compression.

Brendan Murphy
University College Dublin

"Mixed-Membership of Experts Stochastic Blockmodel"

Social network analysis is the study of how links between a set of actors are formed. Typically, it is believed that links are formed in a structured manner, which may be due to, for example, political or material incentives, and which often may not be directly observable. The stochastic blockmodel represents this structure using latent groups which exhibit different connective properties, so that conditional on the group membership of two actors, the probability of a link being formed between them is represented by a connectivity matrix. The mixed membership stochastic blockmodel (MMSBM) extends this model to allow actors membership to different groups, depending on the interaction in question, providing further flexibility.

Attribute information can also play an important role in explaining network formation. Network models which do not explicitly incorporate covariate information require the analyst to compare fitted network models to additional attributes in a post-hoc manner. We introduce the mixed membership of experts stochastic blockmodel, an extension to the MMSB which incorporates covariate actor information into the existing model. The method is illustrated with application to the Lazega Lawyers friendship dataset. Model and variable selection methods are also discussed.

Authors: Arthur White and Thomas Brendan Murphy, School of Mathematical Sciences,
University College Dublin, Dublin 4, Ireland.

Karl Rohe
University of Wisconsin

“Local Clustering and the Blessing of Transitivity”

In massive networks, standard "global" algorithms and models are inappropriate for several reasons. Instead, several researchers have demonstrated the benefits of local algorithms. This talk aims to give these methods a framework for statistical inference.

First, if a stochastic blockmodel is both sparse and transitive (two persistent feature of empirical networks), then it contains blocks that do not grow asymptotically (i.e. local clusters).

Second, we propose a simple local algorithm that leverages the triangles in the graph. To study this local algorithm, we propose the "local stochastic blockmodel" that makes drastically reduced assumptions on the global structure of the graph; for local inference, this model only makes local assumptions. Our proposed algorithm finds the local block w.h.p. under an asymptotic regime that allows both bounded block size and bounded or growing node degrees.

Aleksandra Slavkovic
Pennsylvania State University

“Differentially Private Graphical Degree Sequences and Synthetic Graphs”

Increasing volumes of personal and sensitive data are collected and archived by health networks, government agencies, search engines, social networking websites, and other organizations. The social networks, in particular, are a prominent source of data for researchers in economics, epidemiology, sociology and many other disciplines and have sparked a flurry of research in statistical methodology for network analysis. While the social benefits of analyzing these data are significant, their release can be devastating to the privacy of individuals and organizations. In this talk, we give a brief overview of challenges associated with protecting confidential data, and the problem of releasing summary statistics of graphs needed to build statistical models for networks while preserving privacy of individual relations. We present an algorithm for releasing graphical degree sequences of simple undirected graphs under the framework of differential privacy. The algorithm is designed to provide utility for statistical inference in random graph models whose sufficient statistics are functions of degree sequences. Specifically, we focus on the tasks of existence of maximum likelihood estimates, parameter estimation and goodness-of-fit testing for the beta model of random graphs. We show the usefulness of our algorithm by evaluating it empirically on simulated and real-life datasets. As the released degree sequence is graphical, our algorithm can also be used to release synthetic graphs under the beta model.
(Joint work with Vishesh Karwa)

Cosma Shalizi
Carnegie Mellon University

"When Can We Learn Network Models from Samples?"

Statistical models of network structure are models for the entire network, but the data is typically just a sampled sub-network. Parameters for the whole network, which are what we care about, are estimated by fitting the model on the sub-network. This assumes that the model is "consistent

under sampling" (forms a projective family). For the widely-used exponential random graph models (ERGMs), this trivial-looking condition is violated by many popular and scientifically appealing models; satisfying it drastically limits ERGMs' expressive power. These results are special cases of more general ones about exponential families of dependent variables, which we also prove. As a consolation prize, we offer easily checked conditions for the consistency of maximum likelihood estimation in ERGMs, and discuss some possible constructive responses. (Joint work with Alessandro Rinaldo; paper at http://arxiv.org/abs/1111.3054)

Blair Sullivan
North Carolina State University

“Is Intermediate-Scale Structure Tree-like in Social Networks?”

Network science is a rapidly growing interdisciplinary field with methods and applications drawn from across the natural, social, and information sciences. A significant challenge in analyzing large complex networks has been understanding the “intermediate-scale structure”—those properties not captured by metrics which are very local (e.g., clustering coefficient) or very global (e.g., degree distribution). It is often this structure which governs the dynamic evolution of the network and the behavior of diffusion-like processes on it. Although there is a large body of empirical evidence suggesting that complex networks are often “tree-like” at intermediate to large size-scales (e.g. hierarchical structures in biology, hyperbolic routing in the Internet, and core-periphery behavior in social networks), it remains a challenge to quantify and take algorithmic advantage of this structure in data analysis.

In this talk, we describe recent empirical and theoretical results aimed at integrating techniques from structural graph theory (tree decompositions and minors), core-periphery heuristics (k-core decompositions) and metric tree embeddings (Gromov hyperbolicity) into scalable and robust tools for extracting meaningful tree-like structure from large informatics networks. We present computational results showing the successes and failings of existing measures of tree-likeness on real world data, and discuss recent progress integrating structural information into downstream frameworks for local inference. Finally, as time permits, we will discuss applications and several open problems.

A.C. Thomas
Carnegie Mellon University

“Protocols for Randomized Experiments to Identify Network Contagion”

Identifying the existence and magnitude of “social contagion”, or the spread of an individual trait along ties in a social network, is a challenging task due in part to the tendency of individuals with similar characteristics to connect, also known as homophily. While randomized experiments on individuals of a network would seem to be the ideal method for establishing contagion, there are still considerable methodological issues stemming from structural considerations, including the likelihood of inclusion in the sample group and the implicit dependence between units due to latent homophily. We construct a protocol that correctly adjusts for these factors in a number of experimental situations. (joint with Michael Finegold)

Johan Ugander
Cornell University

“Graph Cluster Randomization: Design and Analysis for Experiments in Networks”

A/B testing is a standard approach for evaluating the effect of online experiments; the goal is to estimate the 'average treatment effect' of a new feature or condition by exposing a sample of the overall population to it. A drawback with A/B testing is that it is poorly suited for experiments involving social interference, when the treatment of individuals spills over to neighboring individuals along an underlying social network. In this work, we propose a novel methodology for using graph clustering to analyze average treatment effects under social interference. To begin, we characterize graph-theoretic conditions under which individuals can be considered to be 'network exposed' to an experiment. We then show how clustered graph randomization admits an efficient exact algorithm to compute the probabilities for each vertex being network exposed under several of these exposure conditions, allowing for a straightforward effect estimator using inverse probability weights. This estimator is also unbiased assuming that the exposure model has been properly specified. Given this framework for graph cluster randomization, we analyze the variance of the estimator as a property of cluster design, and the bias/variance trade-offs of the estimator under simulated exposure model misspecifications. Our analysis of the variance includes a novel clustering algorithm for which the variance is at most linear in the degrees of the graph, an important challenge. Our analysis of misspecifications highlights when clustering appears to be strongly favorable: when the network has a sufficiently clustered structure, and when social interference is sufficiently strong. We further illustrate our method with real experiments, including a large experiment on Facebook on Thanksgiving Day 2012.

This is joint work with Brian Karrer, Dean Eckles, Lars Backstrom (Facebook) and Jon Kleinberg (Cornell).

Rebecca Willett
University of Wisconsin-Madison

“Tracking Influence in Dynamic Social Networks”

Cascading chains of interactions are a salient feature of many real-world social networks. One particularly well-studied example is social reciprocity between two people, but more distributed series of interactions are also possible: kindnesses are "paid forward", gang violence begets retaliations, nation-state conflicts are accompanied by proxy wars, and chain emails are regularly forwarded. This talk addresses the challenge of tracking how the actions within a social network stimulate or influence future actions. We adopt an online learning framework well-suited to streaming data, using a multivariate Hawkes model to encapsulate autoregressive features of observed events within the social network. Recent work on online learning in dynamic environments is leveraged not only to exploit the dynamics within the social network, but also to track that network structure as it evolves. Regret bounds and experimental results demonstrate that the proposed method (with no prior knowledge of the network) performs nearly as well as would be possible with full knowledge of the network. Joint work with Eric Hall.