samsi
NSF·Duke·NCSU·UNC

Data Science in the Social and Behavioral Sciences
Virtual Opening Workshop
January 5/11/12, 2021

SPEAKER TITLES/ABSTRACTS

Chris Bail
Duke University

"Using Bots and Linked Social Media Data to Study Political Polarization"

Abstract: There is mounting concern that social media sites contribute to political polarization by creating "echo chambers" that insulate people from opposing views about current events. We surveyed a large sample of Democrats and Republicans who visit Twitter at least three times each week about a range of social policy issues. One week later, we randomly assigned respondents to a treatment condition in which they were offered financial incentives to follow a Twitter bot for 1 month that exposed them to messages from those with opposing political ideologies (e.g., elected officials, opinion leaders, media organizations, and nonprofit groups). Respondents were resurveyed at the end of the month to measure the effect of this treatment, and at regular intervals throughout the study period to monitor treatment compliance. We find that Republicans who followed a liberal Twitter bot became substantially more conservative posttreatment. Democrats exhibited slight increases in liberal attitudes after following a conservative Twitter bot, although these effects are not statistically significant. Notwithstanding important limitations of our study, these findings have significant implications for the interdisciplinary literature on political polarization and the emerging field of computational social science.

Jessica Cohen
University of North Carolina

"Network Neuroscience Insights into Developmental Disorders: Functional Brain Network Dysfunction in ADHD"

Network analysis as applied to neuroimaging data has led to novel insights into how brain function underlies both typical and atypical behavior. There has been a large push recently to measure dysfunctional brain network organization as a biomarker for disease, as systematic differences between patients and controls are observed transdiagnostically. This talk will focus on knowledge gained about the brain basis of ADHD using network analysis. I will discuss two sets of analyses in which differences in interactions across distinct brain networks, particularly between the default mode network and task-relevant networks, are observed in ADHD. Both static and dynamic network interactions will be considered. Together, this research identifies promising network-based biomarkers for identifying individuals with ADHD and, more generally, provides an example of an application of network neuroscience to developmental disorders.

David Dunson
Duke University

"Structural Brain Connectomics: new mathematical representations & statistical inference methods"

It has become routine to collect data on the locations of white matter fiber bundles in the brain through diffusion tensor imaging. We discuss novel representations of the structural brain connectome based on data of this type motivated by interest in studying relationships between connectomes and human traits.

Emily Falk
University of Pennsylvania

"Social Networks and the Brain"

Brain dynamics shape learning and behavior, and social context shapes brain structure and function. In this talk, I will present evidence from a series of studies in adolescents and young adults linking brain responses to individual differences in social network structure. Specifically, I will focus on how brain systems implicated in processing social rewards, social threats, and more general understanding of others' mental states are associated with different types of social network properties, and in turn how these differences relate to susceptibility to social influences.

Diego Fregolente
SAMSI

"Competition and Spreading of Low and High Quality Information in Online Social Networks"

The advent of online social networks as major communication platforms for the exchange of information and opinions is having a significant impact on our lives by facilitating the sharing of ideas. Through networks such as Twitter and Facebook, users are exposed daily to a large number of transmissible pieces of information that compete to attain success. Such information flows have increasingly consequential implications for politics and policy, making the questions of discrimination and diversity more important in today's online information networks than ever before. However, while one would expect the best ideas to prevail, empirical evidence suggests that high-quality information has no competitive advantage. We investigate this puzzling lack of discriminative power through an agent-based model that incorporates behavioral limitations in managing a heavy flow of information and measures the relationship between the quality of an idea and its likelihood to become prevalent at the system level. We show that both information overload and limited attention contribute to a degradation in the system's discriminative power. A good tradeoff between discriminative power and diversity of information is possible according to the model. However, calibration with empirical data characterizing information load and finite attention in real social media reveals a weak correlation between quality and popularity of information. In these realistic conditions, the model provides an interpretation for the high volume of viral misinformation we observe online.

Laura Germine
McLean Hospital and Harvard Medical School

"Dynamic Cognitive Assessment in Health and Disease"

This talk will focus on modern methods for high frequency cognitive assessment using the web and mobile devices. Considerations related to psychometrics, accessibility, technology, and participant engagement will be discussed, as well as the opportunities and challenges for understanding dynamic cognitive mechanisms, over time.

Krista Gile
University of Massachusetts

"Clustering Network Tree Data From Respondent-driven Sampling"

There is great interest in finding meaningful subgroups of attributed network data. There are many available methods for clustering complete network. Unfortunately, much network data is collected through sampling, and therefore incomplete. Respondent-driven sampling (RDS) is a widely used method for sampling hard-to-reach human populations based on tracing links in the underlying unobserved social network. The resulting data therefore have tree structure representing a sub-sample of the network, along with many nodal attributes. In this paper, we introduce an approach to adjust mixture models for general network clustering for samples collected by RDS. We apply our model to data on opioid users in New York City, and detect communities reflecting group characteristics of interest for intervention activities, including drug use patterns, social connections and other community variables.

This is joint work with Shuaimin Kang, Pedro Mateu-Gelabert, and Honoria Guarino.

Oscar Gonzalez
University of North Carolina

"Statistical Mediation Analysis from the Causal Inference and Structural Equation Modeling Perspectives"

Statistical mediation uncovers the intermediate variables, known as mediators, that explain how an independent variable caused a change in an outcome. Statistical mediation is often used in the areas of prevention and intervention science to investigate the mechanisms of behavior change through which an intervention causally affects a health behavior. New mediation methods from the causal inference literature are a seminal for mediation analysis because they focus on the causal basis of mediation. However, these methods are not widely known and differ from the traditional methods used in the social sciences. In this talk, I discuss some links between the traditional and the causal frameworks for mediation, along with some of my current work in this area.

Michael Hudgens
University of North Carolina

"Causal Inference with Interference"

A fundamental assumption usually made in causal inference is that of no interference between individuals, i.e., the potential outcomes of one individual are assumed to be unaffected by the treatment or exposure of other individuals. However, in many settings, this assumption obviously does not hold. For example, in social/behavioral studies, we may expect spillover or peer effects between individuals who are socially connected. In this talk we will discuss recent approaches to assessing causal effects in the presence of interference.

Samuel Jenness
Emory University

"Statistical Approaches to Modeling Infectious Disease Epidemics across Temporal Contact Networks"

Infectious diseases are transmitted across highly structured networks of social contacts that form and dissolve over time. Investigating network drivers of epidemics and network-based opportunities for disease control has required the development of statistical approaches to modeling dynamic network structures embedded within broader mathematical models of intra- and inter-host epidemiology, demography, and bio-behavioral pathogen spread. In this talk, I present on temporal exponential random graph models (TERGMs) to model dynamic contact networks using easily collected egocentric network data, the integration of these methods within our epidemic modeling software platform â€" EpiModel (www.epimodel.org) â€" and our recent applications of these tools to the transmission dynamics of HIV, bacterial sexual transmitted infections, tuberculosis, and SARS-CoV-2.

Eric Kolaczyk
Boston University

"Quantitative Methods for Understanding Coalescence and Fragmentation in Dynamic Networks of Epileptic Seizures"

While current technology permits inference of dynamic brain networks over long time periods at high temporal resolution, the hallmark pattern of network coalescence and fragmentation during human seizures remains poorly understood. I will briefly summarize two related projects in this area: (i) a method of dynamic community detection that addresses critical aspects unique to the analysis of dynamic functional networks inferred from noisy seizure data; and (ii) a new class of random graph hidden Markov models, and associated statistical inference machinery, that allows for comparison of percolation regimes as putative (albeit oversimplified) mechanisms around seizure onset.

David Lazer
Northeastern University

"Patterns of Sharing Fake News in 2016 and 2020"

This presentation examines the patterns of sharing and exposure to fake news in 2016 (regarding the Presidential election) and 2020 (regarding COVID-19). Key findings include: fake news is fairly common as measured by content on Twitter, but exposure and, especially, sharing of fake news is highly concentrated among a small number of people. Those sharing being exposed to misinformation are vastly disproportionately older conservatives. However, in a parallel study of misperceptions regarding COVID-19, the partisan relationship is much attenuated, and the age relationship inverted. This set of findings point to the existence of an amplification ecosystem on Twitter that reaches a fairly narrow strata of the population.

Fan Li
Duke University

“A Regression Discontinuity Design for Ordinal Running Variables: Evaluating Central Bank Purchases of Corporate Bonds”

Regression discontinuity (RD) is a widely used quasi-experimental design for causal inference. In the standard RD, the assignment to treatment is determined by a continuous pretreatment variable (i.e., running variable) falling above or below a pre-fixed threshold. In the case of the corporate sector purchase programme (CSPP) of the European Central Bank, which involves large-scale purchases of securities issued by corporations in the euro area, such a threshold can be defined in terms of an ordinal running variable. This feature poses challenges to RD estimation due to the lack of a meaningful measure of distance. To evaluate such program, this paper proposes an RD approach for ordinal running variables under the local randomization framework. The proposal first estimates an ordered probit model for the ordinal running variable. The estimated probability of being assigned to treatment is then adopted as a latent continuous running variable and used to identify a covariate-balanced subsample around the threshold. Assuming local unconfoundedness of the treatment in the subsample, an estimate of the effect of the program is obtained by employing a weighted estimator of the average treatment effect. Two weighting estimators---overlap weights and ATT weights---as well as their augmented versions are considered. We apply the method to evaluate the causal effect of the CSPP and found and find a statistically significant and negative effect on corporate bond spreads at issuance.

Yu-Ru Lin
University of Pittsburgh

“Online Misinformation Consumption: the myth of typical consumers and eye appeal”

With increased social isolation due to the measures taken to slow down the spread of COVID-19, reliable information becomes more crucial than ever. Who are the typical misinformation consumers in the pandemic? What content is powerful but often misleading? In this talk, I will present our recent works that study the information being spread online in this ongoing "misinfodemic" using a large sample of communication traces collected from Twitter. I will discuss (1) some of the typical and atypical characteristics of online users who are more likely to engage in COVID-related misinformation consumption, and (2) the kinds of visual and text content likely to associate with widely-distributed unreliable sources. I will also discuss statistical learning approaches to identify these different types of users and content. Our studies help reveal new opportunities to decelerate or stop the misinformation propagation, such as to predict the potentially at-risk misinformation consumers for early targeting and timely intervention, as well as to contribute to enhancing the public's critical literacy and resilience to misinformation.

Cameron McIntosh
Government of Canada

“Let’s Remain Flexible: Nonparametric Modeling in Causal Analysis”

An observed association between X and Y contains some unknown combination of causal and non-causal (spurious) components. A nonparametric identification strategy allows the investigator to isolate a causal effect from an observed association without imposing any statistical assumptions. When we move to estimation of the causal effects, however, we leave the non-parametric world behind to some extent. Nonetheless, we should make the necessary sacrifices -- in terms of functional form and distributional assumptions -- as minimal as possible. Linear additive modeling is still by far the most popular approach in SEM applications, but is not likely an adequate

representation of our complex non-linear world. Fortunately, econometric advances over the last couple of decades make it possible to estimate general non-linear, non-additive structural equations that are more aligned with the spirit of nonparametric causal identification. This talk gives an overview of these methods and how they can be implemented in SEM research.

Sarah Muldoon
University at Buffalo

"Mapping Systems to Graphs: challenges of modeling neuroimaging data"

Network neuroscience is a rapidly growing field that uses tools from network theory to investigate brain networks across multiple scales and modalities. While the use of network analysis in neuroscience has led to many important insights about brain structure and function, properly mapping neuroimaging data to a graph presents multiple difficulties. In this talk, Iâ€™ll discuss some of the challenges of building networks from neuroimaging data and why the structure of these networks can be difficult to properly interpret. Further, because many network metrics were designed for other systems such as social networks, typical network measures may not be appropriate for detecting and analyzing structure in brain networks. Iâ€™ll therefore end by presenting work that aims to modify existing network metrics and/or design new metrics that are specifically tailored to the features of brain networks.

Bengt Muthen
UCLA and Mplus software

"Effect Estimation with Latent Variables"

This presentation gives an overview of standard and novel mediation modeling with counterfactually-defined causal effects available to SEM analysts as well as others using the general latent variable modeling framework of the Mplus software. The focus is on effect analysis with latent variables, including latent mediator constructs measured by multiple indicators to avoid measurement error bias, complier-average causal effect mixture modeling to handle non-compliance in randomized experiments, multilevel mediation modeling with latent centering and latent variable interactions, and analysis of intensive longitudinal data using multilevel time series modeling with random effects used for propensity score matching as well as moderators and mediators in randomized studies.

Co-author: Tihomir Asparouhov, Mplus

Judea Pearl
University of California, Los Angeles

"Deep Understanding through Structural Equation Models (SEM)"

Links to reading list:
1. The First Law of Causal Inference (https://ucla.in/2QXpkYD)

2. "The Seven Tools of Causal Inference" 2016 Article: https://ucla.in/2HI2yyx
Summary: https://vimeo.com/314324108

Jukka-Pekka Onnela

Harvard University

“Smartphone-based Digital Phenotyping”

Behavior has traditionally been a difficult phenotype to characterize because of its temporal nature and context dependence, and it has been captured using self-reports or clinician-administered surveys. Unfortunately, both are subjective, qualitative, and typically cross-sectional. This is a substantial limitation, both for research and clinical practice, because behavioral changes are present in most illnesses, especially in central nervous system disorders. Consequently, more granular ways to capture behavior could lead to new ways to diagnose, treat, and prevent disease. We have previously defined digital phenotyping as the moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices, in particular smartphones. Smartphone-based digital phenotyping can give rise to temporally dense, longitudinal measurement of behavior in naturalistic or free-living settings. Making sense of data collected from consumer grade devices is however very challenging, but in addition to leading to more precise phenotypes, better data could lead to better mathematical models of complex diseases. I will discuss some of the opportunities and challenges that smartphone-based digital phenotyping presents for data science in the social and behavior sciences.

Brea Perry
Indiana University

“Leveraging Personal Social Network Data to Understand Social and Biological Mechanisms of Cognitive Aging”

Significant advances in the prevention and treatment of Alzheimerâ€™s disease and related dementias (ADRD) in older populations have been slow to emerge. In response, researchers have turned to an examination of social-environmental factors to identify candidates for clinical intervention. Recently, social engagement (variously operationalized) has been linked to a shorter period of cognitive decline, greater independence, and higher quality of life in people with AD. However, the social and biological mechanisms of this relationship are poorly understood. In this talk, I will discuss theoretical pathways, methodological insights, and preliminary results from the Social Networks and Alzheimer's Disease Study (SNAD). SNAD examines associations between personal social network characteristics, biomarkers of underlying neurodegeneration, and cognitive function using a sample of older adults at high risk for ADRD.

Paul Resnick
University of Michigan

“Survey Equivalence: A Procedure for Measuring Classifier Accuracy Against Human Labels”

In many classification tasks, the ground truth is either noisy or subjective. Examples include: which content or explanation is better according to some community? is this comment toxic? what is the political leaning of this news article? We refer to such tasks as survey settings because the ground truth is defined through a survey of one or more human raters. In survey settings, conventional measurements of classifier accuracy such as precision and recall confound the quality of the classifier with the level of agreement among human raters.

We describe a procedure that, given a dataset with K raters per item and predictions from a classifier, rescales any accuracy measure into one that has an intuitive interpretation. The key

insight is to score the classifier not against the best proxy for the ground truth, such as a majority vote of the raters, but against a single human rater at a time. That score can be compared to other predictors' scores, in particular predictors created by combining labels from several other human raters. We describe a model of labeling that allows for both noise and subjectivity among raters. Under this model, we describe an optimal combiner (in an information theoretic sense) and provide an efficient computation procedure for it. Running this combiner for rater subsets of various sizes produces a survey power curve, the expected score of predictions made from surveys of up to K-1 raters. The survey equivalence of any classifier is the minimum number of raters needed to produce the same expected score as that found for the classifier.

Ilya Shpitser
Johns Hopkins University

"Identification and Estimation of Causal Parameters via a Modified Factorization of a Graphical Model"

Interventionist causal inference quantifies cause effect relationships as functions of potential outcome random variables. In simple models, causal parameters of interest are identified by variations of the functional known as the g-formula, and estimated using parametric and semi-parametric frameworks for statistical inference. Identification by the g-formula, and subsequent estimation methods in fully observed models have a clean interpretation in terms of the Markov factorization with respect to a directed acyclic graph (DAG).

I show that this interpretation may be extended to arbitrary hidden variable causal models using a more involved nested Markov factorization with respect to a directed mixed graph. This view leads to a simple characterization of non-parametric identification for many parameters in causal inference, including causal effects, direct, indirect, and path-specific effects, responses to counterfactual policies, and many others. In addition, the nested Markov factorization avoids certain paradoxes associated with causal nulls, and leads to well-behaved model likelihoods that avoid making assumptions on unobserved variables, while capturing the structure these variables induce on the observed marginal distribution. This structure takes the form of all equality constraints induced by the hidden variable model, including conditional independences and ``Verma constraints."

In linear structural equation models, nested Markov likelihoods is related to path-diagram likelihoods associated with arid graphs.

This is joint work with Thomas S. Richardson, Robin J. Evans, James M. Robins

Rick Troiano
National Cancer Institute, NIH

"Wrist Accelerometer Data from the National Health and Nutrition Examination Survey"

The National Health and Nutrition Examination Survey (NHANES) 2011-2014 and the 2012 NHANES National Youth Fitness Survey (NNYFS) included an accelerometer-based physical activity monitor for participants ages 3-80+ years. Summary data at the minute, hour, and day levels were recently released and the high resolution (80 Hz) triaxial data are expected to be released soon. The presentation will describe the accelerometer protocol, available data and how to access them. Potential data applications and limitations will be discussed.

Ashton Verdery
Pennsylvania State University

"Inference from Social Network Samples"

Many studies can collect social network data that enables inference beyond the cases sampled. I review some methods and applications of network scale up methods for understanding population health, with an illustration focused on the reach of bereavement from recent mortality crises.

Thespina "Nina" Yamanis
American University

"Social Influence and Diffusion Effects on Men's HIV Testing During a Randomized Controlled Intervention Trial in Tanzania"

Few studies have assessed social network influence on HIV-related behaviors during the course of HIV intervention trials. Even fewer such studies have occurred in sub-Saharan Africa, the region most affected by HIV. In sub-Saharan Africa, men test for HIV at low rates, with consequences including premature mortality and ongoing HIV transmission. Research is needed to understand the drivers of men's HIV testing behavior in this context. The theory of social influence postulates that individuals adopt behaviors they perceive as normative within their networks. In this study, we assess whether social influence was responsible for changes in HIV testing behavior during a social network intervention trial with 40 sociometric networks comprising 1249 young men in Tanzania. Results of the intervention trial demonstrated efficacy for increasing young men's HIV testing (Maman et al, 2020). For this study, we estimated stochastic actor oriented models (SAOMs) in RSiena that test how social influence changes individual's HIV testing behavior. By simultaneously estimating the evolution of network and behavior dynamics, SAOMs estimate for peer influence while controlling for friendship selection. SAOMs differ from traditional regression techniques in their ability to control for endogenous network mechanisms that may induce similarity in friends' HIV testing behaviors. This is the first study, to our knowledge, to use longitudinal sociometric network data to assess network influence on men's HIV risk behaviors in sub-Saharan Africa. Our results demonstrate that social network influence was responsible for the intervention trial's efficacious effect on changes in HIV testing. Network interventions are, thus, highly promising for scaling up HIV testing among young men in sub-Saharan Africa.

Research cited: Maman S, Mulawa M, Balvanz P, McNaughton Reyes L, Kilonzo M, Yamanis T, Singh B, and Kajula L (2020). Results from a cluster-randomized trial to evaluate a microfinance and peer health leadership intervention to prevent HIV and intimate partner violence among social networks of Tanzanian men. *PLOS One*, Mar 20;15(3):e0230371. https://doi.org/10.1371/journal.pone.0230371

With Brian Aronson, Marta Mulawa, Suzanne Maman, Lusajo Kajula, and James Moody