BIG DATA 2017
CONFERENCE PROGRAM

FRIDAY, AUGUST 18

9:00-9:40AM MOHAMMAD AKBARPOUR, STANFORD UNIVERSITY
Title: "Big data is not good data: A theory of social learning in dynamic environments"

Abstract: We study a model of social learning with "overlapping generations", where agents meet others and share data about an underlying state over time. We examine under what conditions the society will produce individuals with precise knowledge about the state of the world. There are two information sharing regimes in our model: Under the full information sharing technology, individuals exchange the information about their point estimates of an underlying state, as well as their sources (or the precision of their signals) and update their beliefs by taking a weighted average. Under the limited information sharing technology, agents only observe the information about the point estimates of those they meet, and update their beliefs by taking a weighted average, where weights can depend on the sequence of meetings, as well as the labels. Our main result shows that, unlike most social learning settings, using such linear learning rules do not guide the society (or even a fraction of its members) to learn the truth, and having access to, and exploiting knowledge of the precision of a source signal are essential for efficient social learning (joint with Amin Saberi & Ali Shameli).

9:40-10:20AM LUCAS JANSON, HARVARD UNIVERSITY
Title: "Model-Free Knockoffs For High-Dimensional Controlled Variable Selection"

Abstract: Many contemporary large-scale applications involve building interpretable models linking a large set of potential covariates to a response in a nonlinear fashion, such as when the response is binary. Although this modeling problem has been extensively studied, it remains unclear how to effectively control the fraction of false discoveries even in high-dimensional logistic regression, not to mention general high-dimensional nonlinear models. To address such a practical problem, we propose a new framework of model-free knockoffs, which reads from a different perspective the knockoff procedure (Barber and Candès, 2015) originally designed for controlling the false discovery rate in linear models. The key innovation of our method is to construct knockoff variables probabilistically instead of geometrically. This enables model-free knockoffs to deal with arbitrary (and unknown) conditional models and any dimensions, including when the dimensionality p exceeds the sample size n, while the original knockoffs procedure is constrained to homoscedastic linear models with n greater than or equal to p. Our approach requires the design matrix be random (independent and identically distributed rows) with a covariate distribution that is known, although we show our procedure to be robust to unknown/estimated distributions. As we require no knowledge/assumptions about the conditional distribution of the response, we effectively shift the burden of knowledge from the response to the covariates, in contrast to the canonical model-based approach which assumes a parametric model for the response but very little about the covariates. To our knowledge, no other procedure solves the controlled variable selection problem in such generality, but in the restricted settings where competitors exist, we demonstrate the superior power of knockoffs through

simulations. Finally, we apply our procedure to data from a case-control study of Crohn's disease in the United Kingdom, making twice as many discoveries as the original analysis of the same data.

10:20-10:50am Break

10:50-11:30AM NOUREDDINE EL KAROUI, UNIVERSITY OF CALIFORNIA, BERKELEY

Title: "Random matrices and high-dimensional statistics: beyond covariance matrices"

Abstract: Random matrices have played a central role in understanding very important statistical methods linked to covariance matrices (such as Principal Components Analysis, Canonical Correlation Analysis etc…) for several decades. In this talk, I'll show that one can adopt a random-matrix-inspired point of view to understand the performance of other widely used tools in statistics, such as M-estimators, and very common methods such as the bootstrap. I will focus on the high-dimensional case, which captures well the situation of "moderately" difficult statistical problems, arguably one of the most relevant in practice. In this setting, I will show that random matrix ideas help upend conventional theoretical thinking (for instance about maximum likelihood methods) and highlight very serious practical problems with resampling methods.

11:30AM-12:10PM TRACY KE, UNIVERSITY OF CHICAGO

Title: "A new SVD approach to optimal topic estimation"

Abstract: In the probabilistic topic models, the quantity of interest—a low-rank matrix consisting of topic vectors—is hidden in the text corpus matrix, masked by noise, and Singular Value Decomposition (SVD) is a potentially useful tool for learning such a low-rank matrix. However, the connection between this low-rank matrix and the singular vectors of the text corpus matrix are usually complicated and hard to spell out, so how to use SVD for learning topic models faces challenges.

We overcome the challenge by revealing a surprising insight: there is a low-dimensional simplex structure which can be viewed as a bridge between the low-rank matrix of interest and the SVD of the text corpus matrix, and which allows us to conveniently reconstruct the former using the latter. Such an insight motivates a new SVD-based approach to learning topic models.

For asymptotic analysis, we show that under a popular topic model (Hofmann, 1999), the convergence rate of the l1-error of our method matches that of the minimax lower bound, up to a multi-logarithmic term. In showing these results, we have derived new element-wise bounds on the singular vectors and several large deviation bounds for weakly dependent multinomial data. Our results on the convergence rate and asymptotical minimaxity are new. We have applied our method to two data sets, Associated Process (AP) and Statistics Literature Abstract (SLA), with encouraging results. In particular, there is a clear simplex structure associated with the SVD of the data matrices, which largely validates our discovery.

12:10-12:25PM DATA SCIENCE LIGHTNING TALKS

12:25-1:30pm Lunch

1:30-2:10PM NIKHIL NAIK, HARVARD UNIVERSITY

Title: "Understanding Urban Change with Computer Vision and Street-level Imagery**"**

Abstract: Which neighborhoods experience physical improvements? In this work, we introduce a computer vision method to measure changes in the physical appearances of neighborhoods from time-series street-level imagery. We connect changes in the physical appearance of five US cities with economic and demographic data and find three factors that predict neighborhood improvement. First, neighborhoods that are densely populated by college-educated adults are more likely to experience physical improvements. Second, neighborhoods with better initial appearances experience, on average, larger positive improvements. Third, neighborhood improvement correlates positively with physical proximity to the central business district and to other physically attractive neighborhoods. Together, our results illustrate the value of using computer vision methods and street-level imagery to understand the physical dynamics of cities. (Joint work with Edward L. Glaeser, Cesar A. Hidalgo, Scott Duke Kominers, and Ramesh Raskar.)

2:10-2:50PM ALBERT-LÁSZLÓ BARABÁSI, NORTHEASTERN UNIVERSITY

Title: "Taming Complexity: From Network Science to Controlling Networks"

Abstract: The ultimate proof of our understanding of biological or technological systems is reflected in our ability to control them. While control theory offers mathematical tools to steer engineered and natural systems towards a desired state, we lack a framework to control complex self-organized systems. Here we explore the controllability of an arbitrary complex network, identifying the set of driver nodes whose time-dependent control can guide the system's entire dynamics. We apply these tools to several real networks, unveiling how the network topology determines its controllability. Virtually all technological and biological networks must be able to control their internal processes. Given that, issues related to control deeply shape the topology and the vulnerability of real systems. Consequently unveiling the control principles of real networks, the goal of our research, forces us to address series of fundamental questions pertaining to our understanding of complex systems.

2:50-3:20pm Break

3:20-4:00PM MARENA LIN, HARVARD UNIVERSITY

Title: "Optimizing climate variables for human impact studies"

Abstract: Estimates of the relationship between climate variability and socio-economic outcomes are often limited by the spatial resolution of the data. As studies aim to generalize the connection between climate and socio-economic outcomes across countries, the best available socio-economic data is at the national level (e.g. food production quantities, the incidence of warfare, averages of crime incidence, gender birth ratios). While these statistics may be trusted from government censuses, the appropriate metric for the corresponding climate or weather for a given year in a country is less obvious. For example, how do we estimate the temperatures in a country relevant to national food production and therefore food security? We demonstrate that high-resolution spatiotemporal satellite data for vegetation can be used to estimate the weather variables that may be most relevant to food security and related socio-economic outcomes. In particular, satellite proxies for vegetation over the African continent reflect the seasonal movement of the Intertropical Convergence Zone, a band of intense convection and rainfall. We also show that agricultural sensitivity to climate variability differs significantly between countries. This work is an example of the ways in which in-situ and satellite-based observations are invaluable to both estimates of

future climate variability and to continued monitoring of the earth-human system. We discuss the current state of these records and potential challenges to their continuity.

4:00-4:40PM ALEX PEYSAKHOVICH, FACEBOOK

Title: TBA

Abstract: TBA

4:40-5:20PM TZE LEUNG LAI, STANFORD UNIVERSITY

Title: "Gradient boosting: Its role in big data analytics, underlying mathematical theory, and recent refinements"

Abstract: We begin with a review of the history of gradient boosting, dating back to the LMS algorithm of Widrow and Hoff in 1960 and culminating in Freund and Schapire's AdaBoost and Friedman's gradient boosting and stochastic gradient boosting algorithms in the period 1999-2002 that heralded the big data era. The role played by gradient boosting in big data analytics, particularly with respect to deep learning, is then discussed. We also present some recent work on the mathematical theory of gradient boosting, which has led to some refinements that greatly improves the convergence properties and prediction performance of the methodology.

SATURDAY, AUGUST 19

9:00-9:40AM NATESH PILLAI, HARVARD UNIVERSITY

Title: "Accelerating MCMC algorithms for Computationally Intensive Models via Local Approximations"

Abstract: We construct a new framework for accelerating Markov chain Monte Carlo in posterior sampling problems where standard methods are limited by the computational cost of the likelihood, or of numerical models embedded therein. Our approach introduces local approximations of these models into the Metropolis–Hastings kernel, borrowing ideas from deterministic approximation theory, optimization, and experimental design. Previous efforts at integrating approximate models into inference typically sacrifice either the sampler's exactness or efficiency; our work seeks to address these limitations by exploiting useful convergence characteristics of local approximations. We prove the ergodicity of our approximate Markov chain, showing that it samples asymptotically from the exact posterior distribution of interest. We describe variations of the algorithm that employ either local polynomial approximations or local Gaussian process regressors. Our theoretical results reinforce the key observation underlying this article: when the likelihood has some local regularity, the number of model evaluations per Markov chain Monte Carlo (MCMC) step can be greatly reduced without biasing the Monte Carlo average. Numerical experiments demonstrate multiple order-of-magnitude reductions in the number of forward model evaluations used in representative ordinary differential equation (ODE) and partial differential equation (PDE) inference problems, with both synthetic and real data.

9:40-10:20AM RAVI JAGADEESAN, HARVARD UNIVERSITY

Title: "Designs for estimating the treatment effect in networks with interference"

Abstract: In this paper we introduce new, easily implementable designs for drawing causal inference from randomized experiments on networks with interference. Inspired by the idea

of matching in observational studies, we introduce the notion of considering a treatment assignment as a quasi-coloring" on a graph. Our idea of a perfect quasi-coloring strives to match every treated unit on a given network with a distinct control unit that has identical number of treated and control neighbors. For a wide range of interference functions encountered in applications, we show both by theory and simulations that the classical Neymanian estimator for the direct effect has desirable properties for our designs. This further extends to settings where homophily is present in addition to interference.

10:20-10:50am Break

10:50-11:30PM ANNIE LIANG, UNIVERSITY OF PENNSYLVANIA

Title: "The Theory is Predictive, but is it Complete? An Application to Human Generation of Randomness"

Abstract: When we test a theory using data, it is common to focus on correctness: do the predictions of the theory match what we see in the data? But we also care about completeness: how much of the predictable variation in the data is captured by the theory? This question is difficult to answer, because in general we do not know how much "predictable variation" there is in the problem. In this paper, we consider approaches motivated by machine learning algorithms as a means of constructing a benchmark for the best attainable level of prediction. We illustrate our methods on the task of predicting human-generated random sequences. Relative to a theoretical machine learning algorithm benchmark, we find that existing behavioral models explain roughly 15 percent of the predictable variation in this problem. This fraction is robust across several variations on the problem. We also consider a version of this approach for analyzing field data from domains in which human perception and generation of randomness has been used as a conceptual framework; these include sequential decision-making and repeated zero-sum games. In these domains, our framework for testing the completeness of theories provides a way of assessing their effectiveness over different contexts; we find that despite some differences, the existing theories are fairly stable across our field domains in their performance relative to the benchmark. Overall, our results indicate that (i) there is a significant amount of structure in this problem that existing models have yet to capture and (ii) there are rich domains in which machine learning may provide a viable approach to testing completeness.

11:30AM-12:10PM ZAK STONE, GOOGLE

Title: TBA

Abstract: TBA

12:10-1:30pm Lunch

1:30-2:10PM JANN SPIESS, HARVARD UNIVERSITY

Title: "(Machine) Learning to Control in Experiments"

Abstract: Machine learning focuses on high-quality prediction rather than on (unbiased) parameter estimation, limiting its direct use in typical program evaluation applications. Still, many estimation tasks have implicit prediction components. In this talk, I discuss accounting for controls in treatment effect estimation as a prediction problem. In a canonical linear regression framework with high-dimensional controls, I argue that OLS is dominated by a natural shrinkage estimator even for unbiased estimation when treatment is random; suggest

a generalization that relaxes some parametric assumptions; and contrast my results with that for another implicit prediction problem, namely the first stage of an instrumental variables regression.

2:10-2:50pm Bradly Stadie, Open AI, University of California, Berkeley

Title: Learning to Learn Quickly: One-Shot Imitation and Meta Learning

Abstract: Many reinforcement learning algorithms are bottlenecked by data collection costs and the brittleness of their solutions when faced with novel scenarios.
We will discuss two techniques for overcoming these shortcomings. In one-shot imitation, we train a module that encodes a single demonstration of a desired behavior into a vector containing the essence of the demo. This vector can subsequently be utilized to recover the demonstrated behavior. In meta-learning, we optimize a policy under the objective of learning to learn new tasks quickly. We show meta-learning methods can be accelerated with the use of auxiliary objectives. Results are presented on grid worlds, robotics tasks, and video game playing tasks.

2:50-3:20pm Break

3:20-4:00pm Hau-Tieng Wu, University of Toronto

Title: "When Medical Challenges Meet Modern Data Science"

Abstract: Adaptive acquisition of correct features from massive datasets is at the core of modern data analysis. One particular interest in medicine is the extraction of hidden dynamics from a single observed time series composed of multiple oscillatory signals, which could be viewed as a single-channel blind source separation problem. The mathematical and statistical problems are made challenging by the structure of the signal which consists of non-sinusoidal oscillations with time varying amplitude/frequency, and by the heteroscedastic nature of the noise. In this talk, I will discuss recent progress in solving this kind of problem by combining the cepstrum-based nonlinear time-frequency analysis and manifold learning technique. A particular solution will be given along with its theoretical properties. I will also discuss the application of this method to two medical problems – (1) the extraction of a fetal ECG signal from a single lead maternal abdominal ECG signal; (2) the simultaneous extraction of the instantaneous heart/respiratory rate from a PPG signal during exercise; (3) (optional depending on time) an application to atrial fibrillation signals. If time permits, the clinical trial results will be discussed.

4:00-4:00pm Sifan Zhou, Xiamen University

Title: "Citing People Like Me: Homophily, Knowledge Spillovers, and Continuing a Career in Science"

Abstract: Forward citation is widely used to measure the scientific merits of articles. This research studies millions of journal article citation records in life sciences from MEDLINE and finds that authors of the same gender, the same ethnicity, sharing common collaborators, working in the same institution, or being geographically close are more likely (and quickly) to cite each other than predicted by their proportion among authors working on the same research topics. This phenomenon reveals how social and geographic distances influence the quantity and speed of knowledge spillovers. Given the importance of forward citations in academic evaluation system, citation homophily potentially put authors from

minority group at a disadvantage. I then show how it influences scientists' chances to survive in the academia and continue publishing. Based on joint work with Richard Freeman.