Statement of Interest in Computational Social Science

RSF Summer Institute in Computational Social Science Application

JOSHUA G. MAUSOLF

THE UNIVERSITY OF CHICAGO

2/17/2017

As a budding computational social scientist, I have taken great strides to develop my computational prowess. In this brief statement, I will cover three primary dimensions of these efforts, namely: (1) academic coursework and teaching, (2) professional experience in computation and quantitative research, and (3) my independent research initiatives.

ACADEMIC INCLINATION TO COMPUTATION

Academically, I have gravitated toward computational and quantitative coursework, taking a raft of courses on topics such as machine learning algorithms and applications, relational databases, natural language processing, web scraping, web-applications, time-series, maximum likelihood, longitudinal analysis, and hierarchical models to name a few. Computational courses used a variety of languages including Python, R, SQL, PostgreSQL, Bash, Git, and Markdown. To a lesser extent, these courses also incorporated JSON, HTML, CSS, JavaScript, and Java. Conversely, most statistical courses utilized R or Stata. Below, I highlight an example project from a course I took, entitled "Machine Learning for Public Policy":

Predicting Dropouts in Montgomery County Public Schools. This project applied a machine learning approach to the problem of high school dropouts in Montgomery County Public Schools (MCPS). I collaborated with fellow students Bridgit Donnelly and Christine Cook. We created a scalable ranking system using machine learning that would allow MCPS to target their interventions to individuals most immediately at risk for dropping out of high school. To detail my involvement in this project, I conducted background research on the school districts and merged this data with the student data using PostgreSQL on a remote AWS server. I wrote the majority of the initial machine-learning pipeline, the feature development, and the classifier initiation and evaluation functions, which can be found on Github. Our team effort resulted in finding an appropriate classifier that improved on current MCPS methods. A copy of the report is available online.

Outside of coursework, I have also sought to bolster my methodological skills by acting as a TA in computational and statistical courses. Courses include both graduate and undergraduate coursework such as "Computing for the Social Sciences," "Machine Learning for Public Policy," and "Statistical Methods for Research," among others. These courses include lectures and lab sessions teaching a variety of topics using Python, R, and Stata.

PROFESSIONAL EXPERIENCE WITH DATA SCIENCE AND QUANTITATIVE RESEARCH

Professionally, I have applied my skills to a variety of data-driven opportunities, including graduate research assistantships and data science fellowships. For example, I picked up several graduate research assistant positions during my time at the University of Chicago. While working for Jenny Trinitapoli, I led a team of junior research assistants in cleaning, recoding, and manipulating longitudinal data on the TLT Project. As a graduate RA for Kathleen Cagney, I create data visualization and analysis for several projects, chiefly (1) examining the effect of localized crime on health, and (2) examining energy consumption and spending in Chicago. Yet, my most computationally intensive projects for faculty include work for James Evans and Xi Song.

Under the guidance of James Evans, I served as a junior data scientist at the Computation Institute's Knowledge Lab at the University of Chicago. I assisted on a project which applied named entity extraction on CV's to build a network of academics connected by edges such as a common university, location, or publication in a given year. My major contributions to the project are twofold. First, I wrote Python code using NetworkX to systematically manipulate and configure the existing hypergraph data. A key issue was that many of the edges reflected data for multiple years, such as "1971-1979." Instead of having nine edges, one for each year, only one edge existed. My code created individual edges for each year between every such node, replacing the initial multiyear edge. My code iterates over this network of 523 nodes and 440,493 edges. Using these results, I helped visualize the network in Gephi, where node size corresponds to an academic's betweenness centrality. The network visuals were rendered using JavaScript for dynamic web interaction. I have put these results on my website, available for years (1941-1982).

Lastly, I am currently working on another computational project for Xi Song to develop an R-application for a new statistical model. The model is a bivariate-locational scale model, which improves on current methods for examining intergenerational mobility. In particular, the model will be allow researches to concurrently model two longitudinal distributions, such as the relationship between parental and child income over the life-course or the relationship between individuals income and wealth distributions over their lifetime.

Beyond these research positions for faculty, I have sought data science and computational fellowships. Last summer, I enjoyed the opportunity to collaborate with a wonderful team of computational scholars at the Eric and Wendy Schmidt Data Science for Social Good summer fellowship. As a 2016 fellow, I worked with other fellows on a project for the Metropolitan Nashville Police Department. We worked in close collaboration with a second DSSG team working with Charlotte-Mecklenburg police. Collectively, our goal was to develop a comprehensive machine learning pipeline to predict adverse police incidents both at the officer level and dispatch level. A major part of this project was designing a common database schema that would allow reproducible ETL for both departments and any future police departments. In this way, we could create a generalizable machine learning pipeline. My major contributions included generating model evaluation code and database schema using Python and PostgresSQL, launching the Python web app, refining ETL for ACS data, and assisting in the feature generation. The public repository for the project is available on Github. Collectively, our results significantly improved over existing methods of identifying officers at risk of adverse incidents, and both departments are working to implement our models.

INDEPENDENT QUANTITATIVE AND COMPUTATIONAL RESEARCH AGENDAS

Beyond my professional research and fellowship activity in computational and quantitative social science, I have pursued computational and quantitative approaches in my own research, both current and future. Here, I focus on my work to understand the effect of social movements on presidential and congressional discussion of inequality, which I have included as a writing sample.

Analyzing Political Rhetoric of President Obama and Congress. Originally, this paper grew out of several computational courses, wherein I developed Python and Bash web-scrapers to collect all presidential speeches and congressional records from 2009-2015. In sum, the text corpora consisted of 6.62 million and 187.70 million words for President Obama and Congress, respectively. I subsequently created a dataset from those corpora using natural language processing with Python. Using this data, I examined fluctuations in rhetoric on inequality in relation to the number of Occupy Wall Street protesters arrested, media coverage of the Occupy movement, and other covariates controlling for public opinion, political climate, and the economy. I used time-series analysis, particularly ARFIMA models to conduct the study. Theoretically, I challenge the existing paradigm that movements are meaningful only in how they effect legislative or policy gains, suggest that rhetorical gains should instead be considered, and argue that the role of the president should be at the forefront of the analysis of social movements. Ultimately, I demonstrate (1) that the arrests of Occupy protesters not only predict media coverage

but also increased discussion of fairness and inequality by President Obama and Congress, (2) that the president's rhetorical shift influences congressional discussion, (3) that this phenomenon persists after the movement faltered, and (4) that Occupy's use of disruptive protest best predicts this rhetorical response, all else equal. The project is currently under peer review, and I have presented and will present this paper in several computationally focused conference sessions, including a presentation at the 2nd Annual International Conference for Computational Social Science and an upcoming presentation in the "Computational Approaches to Dynamic Social Processes" session at the annual meeting for the *Population Association of America*.

Future Research. Building off the Occupy paper, I plan to expand my emphasis on economic inequality. Although the study of inequality can fundamentally be thought of through the lens of (1) a dichotomy between the poor and affluent or (2) an understanding the conditions engendering the penurious, I propose an alternative perspective. Rather than a distributional or left-tail perspective on inequality, I will examine new empirical boundaries to understanding the maintenance, making, and mobility of American elites as part of a dissertation agenda. In particular, I intend to explore the following topics: (1) Discrimination in Elite Labor Market Entry and Transfers, (2) CEO Networks and Executive Compensation, and (3) Intergenerational Income and Wealth Mobility.

I intend to experimentally examine elite labor markets, namely professional service and technology firms, which are seen as contemporary gateways to top incomes and corporate leadership. In particular, I assess labor market discrimination in terms of race and gender relative to a candidate's university prestige, degree field, and technical skills. In part, the lack of diversity among top firms is blamed on a pipeline problem, which suggests that top firms lack diversity not because they are biased, but rather because of a dearth of qualified diversity candidates from top-four schools. Given perfectly qualified candidates, does discrimination still exist, and at lower levels of prestige, where do differences occur? I intend to see to what degree prestigious qualifications (in university degree, major, and technical skills) create advantage in elite labor markets at different career stages. Methodologically, I intend to utilize online correspondence tests. I will use a variety of computational approaches to assist in generating application materials for job applicants, as well as finding and applying to available jobs. Additionally, I am open to exploring alternate approaches using online experiments for this question.

Next, I will analyze the role of executive network centrality using the concept of N-dimensional interlocks to adjudicate several competing theories of executive compensation, particularly the ways in which benchmarking interacts with interlocks and executive social networks, broadly defined. Currently, I need to collect social network data to match with the executive compensation data I have from ExecuComp Compustat, 1992-2015. Analysis will utilize both network methods and time-series approaches. The collection of network data, if not available from institutional datasets, may include implementing web-scraping and named entity extraction. Beyond exploring the relationship between network centrality and compensation, I envision additional projects such as using machine learning and time-series models to predict executive leadership, promotion, and pay.

Lastly, I seek to assess the life-course accumulation of wealth and its intergenerational transmission, particularly the role of financial literacy, investment behavior, and portfolio composition. Additionally, this work will seek to examine the relation of income versus wealth over the life-course, using the bivariate-locational scale model I am developing with Xi Song. A key process will be simultaneously examining income and wealth distributions, particularly the positive divergence of wealth from personal income, as might occur from regular savings and investment versus high consumption. Here, I envision a twofold process, wherein I first use a temporally validated machine learning model to predict positive wealth divergence, and second utilize the bivariate locational scale model to confirm the statistical significance and causality of top-ranking features. While a variety of survey data exists targeting wealth and income, an ideal solution would be adopting Matthew Salganik's approach of data linkage, a strategy likewise suggested by Russell Sage Foundation's recent call for computational proposals. Ideally, we could connect existing survey datasets with investment portfolio behavior from banking and investment institutions. In this way, we could have more granular data on individual financial decisions and develop more robust models than those that rely solely on survey data.