Research Statement, August 9, 2011
Jeff Gill

Since finishing my degree in 1996 I have worked to build a national and international reputation as a recognized and productive scholar in political methodology and statistics. Through a set of books and referee journal articles I have accumulated a record of contributing to the literature on: Bayesian approaches, difficult numerical estimation problems, entropic measures of uncertainty, modern thinking on hypothesis testing, and Markov chain Monte Carlo procedures. I have also contributed to the political methodology community by: hosting the society's web operations for five years, serving as a section officer (member at large, and now president), organizing national meeting panels, serving as an Associate Editor of *Political Analysis*, teaching in the ICPSR and Essex summer programs, and in the Summer of 2006 hosting the Political Methodology Summer Meeting. I expect to continue in these directions and the following text outlines my primary research accomplishments since tenure as well as indicating future directions.

My primary areas of scholarly research are in political methodology and its applications in American and comparative politics, Bayesian stochastic simulation, and applications in public health At the broadest level, I seek to develop or adopt new methodological tools for the empirical analysis of questions in social and health measurement.

More specifically, my core research develops new statistical methods in Bayesian modeling and data analysis to substantive questions in voting behavior/elections, bureaucratic politics, terrorism studies, and healthcare. Since the Bayesian perspective generally necessitates advanced computing to obtain useful empirical results from realistic model assumptions, most of the new statistical work that I am interested in focuses on developing stochastic simulation techniques to solve various estimation problems. Ongoing work includes the development of tools in nonparametric density estimation and Markov chain Monte Carlo (MCMC) simulation. Other related methodological research interests include: functional data analysis, mixture models, Dirichlet process priors, elicited priors, and queueing theory. Applied work has included: Congressional budgeting, tenure of federal executives, modeling turnout, education policy, legislative/bureaucratic interaction, childhood mental health consequences from exposure to war, and analyzing terrorist organizations as networks. The publications, and working papers listed herein and on the enclosed Curriculum Vitae reflect these interests.

To elaborate on the broad outline of research interests described above, I will now give some specific details on current research activities. Building on my interests in Bayesian modeling and associated computing, there are several projects that seek to improve these practices. One key area of interest is incorporating qualitative prior information into Bayesian statistical specifications. Major current projects are described below. Omitted are early stage or yet-to-be-funded projects on: spatial-temporal voting models with Bayesian hierarchical spatial models (with Jamie Monogan, University of Georgia), models of Warfarin dosage (with Brian Gage, Washington University), and discrete MCMC diagnostics (with Dominik Hangartner, UC Berkeley, and Skyler Cranmer, UNC–Chapel Hill).

- **Bayesian Methods: A Social and Behavioral Sciences Approach, Third Edition.** I am currently under contract to produce a third edition of the Bayesian book from 2002 and 2007. Publication will be in 2012. This is a field that moves rapidly, making an update important. There will be a greater emphasis on MCMC tools and the theory behind their use as well as a continued emphasis on regression-style models. A review of the first edition can be found in the *Journal of Politics* 65(3):909-911, and two more in are available at The Political Methodologist. A review of the second edition is in The American Statistician, and Andy Gelman wrote an `amazon.com` review.

- **Using Statistics to Fight Terrorism.** The safety of millions of people depends on the understanding of the workings of covert networks, especially of terrorist networks. To protect people, governments and nongovernmental organizations invest enormous amounts of time and energy to detect covert networks and to thwart terrorist events and other kinds of attacks. The most fruitful contribution of currently published work is the recognition that terrorist organizations are not unitary actors, and that they are highly decentralized networks. However, there are two major deficiencies in academic studies of terrorism: all available data describes only publicly visible events, and missingness in network nodes and edges is generally treated as a regular statistical missing data problem. We (myself and John Freeman, Department of Political Science, University of Minnesota) provide a new means of prior elicitation of deeply contextual, non-numeric information from qualitative experts leading to innovations in social network analysis. Our technology solves the practical problems with visual approaches to elicitation of missing data. The experiment produces new insights into exactly how different visual treatments affect and improve human probability assessments. This suggests that others can use this technology to supply data that is needed to advance work in social network analysis of terrorist and other criminal groups.

- **The Variable Effect of War on Longterm Childhood Mental Health Outcomes.** While children are routinely exposed to armed conflicts ranging from minor skirmishes to full-scale national wars, there is relatively little scholarship on the psychological and emotional consequences they face in either the short or long term. Furthermore, children are disproportionaly affected due to their lessened ability to protect themselves physically as well as their necessarily incomplete emotional development. Regretfully, militarized conflict in the proximity of civilians is not a declining phenomenon. Current scholarship has focused almost exclusively on post-traumatic stress disorder (PTSD), which is a serious and alarming consequence in children. Yet, there is convincing evidence of additional disturbance, including: depression, anger, reduced concentration ability, adverse educational outcomes, emotional detatchment, anxiety, aggression, and long-term grief. Many of these are longer-term outcomes of exposure to violence than the more commonly studied outcome of PTSD. We (myself and Enbal Shacham, Brown School of Social Work Washington University) initiated an interdisciplinary study where the political science, the epidimiology, the psychology, and the statistical analysis are all done to the highest standards of each discipline. Currently there is

no study offering this combination. Additionaly, we also have immediate access through Moi University to health clinics in Eldorot, western Kenya. In late 2007 and into 2008 this area was wracked by inter-tribal (Kikuyu, Luo) violence and retribution following a divisive and corrupt presidential election, most notably the burning alive of over fifty people in the Kenya Assemblies of God Pentecostal Church, including many children. The core of the statistical analysis is the specification of a Bayesian hierarchical model to directly incorporate grouping that results from clustering of effects, geography, and affliction, as well as demographics.

- **Identifying Structure in Social Data Models using Markov Chain Monte Carlo Algorithms.** With George Casella (University of Florida, Statistics), funded by the *National Science Foundation* (Methodology, Measurement, and Statistics program $378,552). There are often structures in social science data such as: unexplained clustering effects, unit heterogeneity, autocorrelation, or missingness, that cast doubt on the notion of unitary effects in a model. We are concerned here with how nonparametric priors can enhance the increasing use of Bayesian models in the social sciences, particularly in the (near ubiquitous) presence of unobserved heterogeneity. Our product partition model (also called classification likelihood) is incorporated into a Dirichlet process mixtures model on the random effect component of a generalized linear model. We are now able to recover posterior probabilities of latent cluster arrangements and therefore determine those with highest posterior probability with a search process through the Stirling number of the second kind possibilities.

- **Dynamic Adaptive Markov Chains for Stochastic Search.** Self-tuning Markov chain algorithms that respond to surface characteristics and adapt are a relatively new approach and one that still requires theoretical defense as well as applied demonstration. We (Jeff Gill and George Casella) have already shown that with one scheme we can analyze previously intractable voting models in high dimensions. The ongoing work here seeks to generalize the approach with a theoretical justification that stability properties can be preserved. We currently finishing a *National Science Foundation* (Mathematical Social and Behavioral Sciences program $350,000) grant for work to develop a family of nonparametric priors (mixtures of Dirichlet processes) and apply this technology to the resulting estimation challenges.

- **Modeling Qualitative Information with Elicited Priors: Confidence in Judicial and Bureaucratic Institutions in Central America.** Researchers who wish to systematically combine qualitative and quantitative information in the same model have found few helpful procedures thus far. However, elicitation and probabilistic interpretation of expert opinion have the potential to bring together previously disparate approaches in political science. To bridge the divide, we (Jeff Gill and Lee Walker, former student, now at the University of South Carolina, Department of Political Science) developed and explained procedures to format qualitative, descriptive, and narrative information for inclusion into a standard empirical model. We extend this work to other elites in Central America and demonstrate that much more can be learned about nascent governments with the Elicited-Bayesian approach than

by using strictly quantitative methods (i.e. survey research) or strictly qualitative methods (i.e. unstructured elite interviews).

- **Queueing Theory Models for Political Science Data.** Queueing theory is widely used in many literatures to describe assembly-line type processes, and services in which the completion time is indeterminant. However, there are almost no applications of queueing theory in political science. This is curious because political actors queue up for desired benefits under a number of circumstances. This project looks at the theoretical and practical basis for applying of queueing theory to the analysis of institutional politics. Empirical applications include the process of bills through legislatures, the scheduling and hearing of court cases, and initiatives within international institutions.

- **Multiple Imputation for Missing Categorical Data in Political Science.** Multiple imputation is the most substantial improvement in the handling of missing data. The ease of software solutions such at the `mice` package in `R` and `Amelia` have made the process of imputing missing data quite easy. Unfortunately, though, the underlying engine still assumes missingness on a continuous metric. Attempts to rectify this and provide helpful software for categorical multiple imputation with regression-style applications has been only partially successful. Currently Skyler Cranmer (a former student of mine, now tenure-track at UNC–Chapel Hill) and I have developed a modernized hot-decking procedure for imputing missing categorical data along the lines of multiple imputation but with a method that preserves the structure of the discreteness as measured. We are currently in the process of justifying the theoretical properties and developing an `R` package.

- **Specifying Spike and Slab Prior Distributions for Bayesian Hypothesis Testing of Nonlinear Models.** Bayesian spike and slab approaches have been developed for stochastic variable selection by designing a hierarchy of priors over the parameter and model spaces. The spike and slab prior is a two-component mixture distribution with one part centered at zero with very high precision (the spike) and the other as a distribution centered at the research hypothesis (the slab). With the selective shrinkage or penalties, this setup incorporates the zero coefficient contingency directly into the modeling process to produce posterior probabilities for hypothesized outcomes. This works especially well with relatively small sample sizes where conventional methods lack statistical power. To date this technology has been applied only to standard linear models, and so we (myself and Xun Pang, former student, now tenure track Princeton University, Department of Politics) extend it to the generalized linear model to facilitate nonlinear outcomes. To overcome the technical challenges in estimating these forms we developed a hybrid Gibbs sampling algorithm and demonstrate that it has better properties than alternative MCMC approaches.

- **A Multilevel Approach to Energy Balance and Cancer Across the Lifecourse.** Awarded, soon funded by the *National Institutes of Health, National Cancer Institute*, to establish at Washington University a Transdisciplinary Research in Energetics and Cancer

(TREC) Center for Co-Investigator/Core Leader for the Bioinformatics Core (Core D). Total award amount: $9,248,284.00. Award Period: March 2011 to December 2016. This project applies spatial and multilevel analytic approaches to assess the independent and joint effects of environments and policies at home and work on physical activity, diet, and obesity among diverse workers. Workers representing multiple ethnic/minority groups will be sampled across four metropolitan areas in Missouri. This study is significant by addressing health priorities in Missouri where large disparities in obesity and related cancers exist among minority populations. It builds on the research team's extensive experience in (1) developing and testing new instruments for assessing environmental and policy influences on obesity, (2) multilevel, ecological interventions and analytical approaches; (3) building large Geographic Information Systems databases and (4) successful management of large-scale analytic studies. This research is innovative by examining environmental and policy influences of obesity across multiple settings. In addition, the results will yield actionable knowledge about what policy and organizational changes would have the greatest impact on healthy lifestyles to prevent obesity and related cancers.