Peter Krafft (PhD Applicant): Statement of Purpose

New data analysis techniques, massive datasets, innovative technological advances, and forward-looking institutions fostering collaborative relationships between social scientists and methodologists are all allowing an emerging generation of computational social scientists to ask fundamentally new questions. Due to the pervasiveness of social network data and the ready availability of text data, two of the disciplines at the forefront of this movement are network analysis and text analysis. However, despite there being considerable previous work in each of these areas individually, their intersection is under-explored. Statisticians, computer scientists, and other methodologists must develop new methods and models in order to make sense of data that include both text and network attributes, such as from email or other communication services. Bayesian statistics, which has been popular for text analysis and to some extent for network analysis, is appealing because it provides a powerful set of tools for quantifying our scientific uncertainty about the social processes underlying these data, and a flexible framework for developing models that scientists can use to approach important social science questions. I am interested in developing Bayesian models of text and network data, and applying these models to social science questions.

This interest originated in part from my experiences with Prof. Michael Lavine, whose steadfast dedication to the principles of Bayesian statistics inspired me in the graduate-level mathematical statistics courses I took with him in my second year at university, as well as in the two research projects I did with him—one on modeling the light that reaches the floor of Harvard Forest and one on diagnosing an estimation issue in a model from spatial statistics.[1] In these classes and projects, I learned about the importance of using common sense in data analysis, about the power of data exploration and visualization, and about how to diagnose unexpected properties of complicated models both through close consideration of their mathematical form and through simulation. Bayesian statistics is important to these ideas because it provides a more sensical interpretation than classical statistics in many situations, and because it allows Monte Carlo methods such as Gibbs sampling and Metropolis Hastings to be used to reason about model parameters.

I recently started using this undergraduate training to develop Bayesian models for political science data. In the U.S., one of the most societally relevant areas of application for computational methods is in political science. With protests from liberals in the Occupy Movement, action from conservatives in the Tea Party, and approval ratings of congress reaching all-time lows,[2] it is clear that the United States government is losing popularity. The American scientific community, as a community that is funded in large part by the government, has a responsibility to help that government recover. Computational social science provides a new opportunity to contribute to this effort. I am doing my MS project with Prof. Hanna Wallach (a computer scientist and an expert in Bayesian models of text) and Prof. Bruce Desmarais (a political scientist and an expert in network analysis). In this project we are using the email inboxes and outboxes from the county department managers of New Hanover County, North Carolina to infer attributes of managers who are particularly effective at getting their own agendas on their county's legislative docket. This project could

[1] For full papers from these and other projects, please see `www.cs.umass.edu/~pkrafft/papers/`.

[2] http://www.gallup.com/poll/149009/congressional-job-approval-ties-historic-low.aspx

help bureaucrats design more effective communication strategies. However, this research poses a major modeling challenge. Email data contain both text attributes (the subject and body of each email) and network attributes (the author and recipients of each email), but there has been very little work in joint text and network modeling. After brainstorming and discussing many possible models, I decided on one to investigate first. By extending a model from the network literature to a framework that is appropriate for email data and combining it with a standard probabilistic model of text, the model I developed can identify who each managers communicates with about particular topics. This model will allow us to discover whether managers with broader communication strategies are more effective than managers with more targeted communication strategies.[3]

During my PhD, I want to continue my work on developing Bayesian models of text and network data for political science applications. Stanford University is ideal for pursuing this goal. In the computer science department, Prof. Daniel Jurafsky applies natural language processing methods to social science questions, and Prof. Jure Leskovec works in social network analysis. In the political science department, Prof. Justin Grimmer does great work in applying Bayesian methods to political science questions. One specific project I find interesting is Prof. Leskovec's MemeTracker. I met with Prof. Leskovec when he recently visited my department. I attended a meeting with him and several other grad students in which we discussed whether using a more sophisticated text analysis method would help the MemeTracker system. Although Prof. Leskovec mentioned that he had tried to use a straight-forward application of a popular probabilistic text analysis method, he never tried to develop a new model suitable for his data. I would be interested in helping him develop such a method. Another related paper I find interesting is "Who Leads Whom: Topical Lead-Lag Analysis across Corpora" presented at last year's NIPS Workshop on Computational Social Science. This paper, which both Prof. Leskovec and Prof. Jurafsky contributed to, involved both text analysis and network analysis components.

I am well-prepared for researching these types of problems as a full-time PhD candidate. I have a thorough knowledge of Bayesian statistics and machine learning from my undergraduate and MS programs, as well as from being the teaching assistant for my department's graduate machine learning course. I also have recently become involved in the computational social science community both through my research and through volunteering as the student assistant for the Second NIPS Workshop on Computational Social Science.

Using my knowledge and skills to collaborate with the faculty and students in both Stanford's Computer Science Department and Stanford's Political Science Department would undoubtedly position me to achieve my ultimate goals of getting a faculty position at a research university, making fundamental contributions to the field of Bayesian statistics, and using the methods I develop for societally relevant applications. Computational social science is important because it has opened a communication channel between methodologists and practitioners. I look forward to contributing my part to the nascent community at the intersection of statistics, computer science, and political science.

Thank you for your time and consideration.

[3] I am currently writing up our preliminary results in the form of a tech report, which we plan to submit to the next International Conference on Machine Learning.