Project Summary

The massive quantities of textual communications generated within most organizations constitutes a largely untapped source for insightful, real-time organizational analytics. From understanding external demands placed on organizations to summarizing the pressing intra-organizational players and issues, most salient developments are documented in digitized text. The content and context recorded in an organization's textual record can be leveraged to understand and improve an organization's performance. The basis of this project lies in two recent developments. First, recent research shows that the patterns and structure of communication, formalized as communication networks, are extremely important to effective organizational and individual problem-solving. Second, many organizations, particularly government entities, have developed open textual input platforms in order to improve responsiveness to user (e.g., citizen, customer) needs. This project builds an analytical bridge between intra-organizational communication networks and streams of external input. Specifically, we will develop methods to parse and summarize the the contents of (1) input streams from external sources and (2) intra-organizational communication networks in the same topic-space, and understand the relationships between the external and internal domains.

Methodology: We propose to study the ways in which government officials' communications with those outside of government are related to intra-governmental communications and government outputs. We will use Florida and North Carolina county government email archives acquired via public records requests and online data collection. We will design computational tools related to statistical topic modeling and network analysis that (1) identify topic-specific internal-external communication networks, (2) identify topic-specific internal communication networks and (3) learn the relationships between internal-external communications, intra-organizational communication networks and the contents of public policies. The methods we develop will track the migration of topics to, within and from government. As such, we will characterize the democratic process at a fine-grained, content and context specific level. The data we collect will permit extensive validation and innovative application of the algorithms developed. We will relate the core email data with additional publicly available data on county governments, including regulations/legislation and minutes from county legislatures. The proposed research will be conducted by an interdisciplinary team that brings expertise in tcomputational (Wallach) and social scientific (Desmarais) fields.

Intellectual Merit: This project will offer important contributions to both computational and social sciences. In terms of computational approaches, we will enhance methods for the statistical analysis of text and network data. In particular, we will expand upon extant methods of textual network analysis in developing ways to learn (1) the topics that cut across network domains and (2) functions that characterize domain transfer of topics. On the social science side, the methods we develop and data we collect will advance organizations' ability to connect streams of eternal input to their internal operations. Also, more directly, we will offer an unprecedented fine-grained assessment of government responsiveness at the local level in the US.

Broader Impact: This project will provide essential tools for organizations in providing timely and coherent responses to the demands of external constituencies. This holds potential to, e.g., improve the efficiency with which local governments manage public health needs, address environmental risks and establish revenue and spending policies. The contributions will be cross-disciplinary and will contribute to the broader scientific community. Also, we will provide an enormous data archive of government communication data to be tapped by other researchers.

Organizational Responsiveness to External Demands: A Modeling Approach based on Statistical Text and Network Analysis

1 Introduction

Nearly every organization strives to respond in a timely and accurate manner to the needs and demands of some external constituency. Firms respond to customers, governments respond to citizens and educational institutions respond to students. The rapid advancement in communications technology over the last two decades has forever transformed the nature, volume and sources of input and feedback available to organizations. Also, electronic communications have drastically improved the ability of organizations to document and communicate their internal developments. These complimentary developments have had a transformative impact on governance - moving to what **CITE** call 'we government'. Most elected officials can be directly contacted electronically through simple internet pools. Citizens can advertise and sign petitions on the web and attend internet 'town meetings' with their representatives. Regarding the internal activities of government; citizens can access electronic communications of their officials through public records requests, access meeting minutes on the web and, e.g., watch the floor activities of the US House of Representatives on HouseLive.gov.

In this project we will develop and apply methods for identifying the cycle of input, response and feedback that leaves its fingerprint in the electronic communications record. We will focus on the nexus between government organizations and their constituents, but the methods we develop will be portable to other types of organizations. Government responsiveness to citizen input offers an ideal venue within which to model the relationship between streams of textual records. First, in democratic societies there is a common expectation that the government will respond to public demands. Second, most of the input modes on which we will focused were designed precisely for the objective of providing input to which public officials could respond. Third, and perhaps of greatest practical importance, due to the scope of freedom of information laws in the US, we as researchers can access the public input and internal communications data associated with a multitude of government organizations.

We frame this project by associating different phases in the cycle of governance with four different types of textual streams - public input (e.g., emails from citizens to government officials, informal internal communications (e.g., emails among officials), formal deliberations (e.g., legislative meeting minutes) and policy outputs (e.g., regulations, laws). We seek to understand these textual themes through the lens of statistical topic models **CITES**. We will develop and apply models that permit the identification of the ways in which topics rise and fall within domains and, crucially, are related across domains. The result will be an analytical approach that permits an organization to distill and investigate the dynamics of input, responsiveness and feedback through a common framework of statistical text analysis. The methods we develop will offer answers regarding several pertinent questions about organizational management of outside input, e.g., is organizational attention to a topic proportional to its attention in outside input, how does an organization adapt to the rise of issues that are novel relative to its current foci, is responsiveness timely?

Topic models infer discrete topics from a corpus of documents. A topic is simply a relative frequency distribution of words and each document is probabilistically associated with each topic identified **CITES**. Statistical topic models provide a dually qualitative and quantitative inferential summary of textual corpora. Qualitative in that the textual content of a corpus is maintained and

words themselves form the basis of the quantitative analysis. Dynamic topic models provide an excellent framework within which to understand input to, output from and feedback to organizations that document their activities at various stages in a textual format. Since the seminal work on statistical topic models **CITE**, the basic framework has been extended and adapted to focus on several aspects of textual corpora **CITE**, including author-specific attributes of text **CITE**, dyadic (i.e., author-recipient) aspects of messages **CITE**, dynamics, the underlying communication network, and joint text-metadata models of documents **CITE**. In the current project, we will undertake an ambitious set of extensions that integrate several of these extensions - jointly modeling separate streams of text that influence each other, are informed by rich meta-data, incorporate the underlying communication network, and characterize the over-time aspect of the text streams.

The benefit from connecting these innovations in statistical topic modeling is that we will leverage a medium common to each domain relevant to a cycle of organizational feedback and responsiveness - textual documentation to connect the domains as well as domain-specific metadata types. For example, in the case of governance, we will tie together the identities of citizens and groups providing outside input, the structure of communication networks underlying informal communications within government and voting coalition patterns within legislatures; all through the medium of the co-evolving, domain-specific text streams.

Figure 1 Illustrates the cycle of organizational responsiveness that we intend to model through the guise of co-evolving textual streams. Considering the case of governance, substantial research exists that focuses on parts of this cycle. For example, a large body of research exists that documents recent developments in tools for citizens to provide precise, timely and voluminous input to government officials **CITES**. There is also a large body of research focusing on legislative adaptation to broad ideological trends among constituents **CITES**. And, yet another literature that addresses the processes by which topics rise from informal awareness among officials and outside parties to the legislative agenda **CITES**.. However, due to the historical inaccessibility of timely and common data modes related to each component of the governance cycle, little research has endeavored to connect all of the dots. We will provide such a complete picture, leveraging the common, available, and timely mode of text streams.

The disadvantage of analyzing relationships between domains in a separate, pairwise manner is that it is impossible to reconstruct a complete picture of the cycle of input, response and feedback relevant to an organization. For instance, in the example of governance, analysis of public input and any one government domain could provide a misleading account of government responsiveness. It may be the case that legislative meeting minutes document consideration of issues that are the subject of considerable public input. However, if final legislation is not responsive, it is not the case that democratic representation has run its complete course. This happened in the case of **CITE** where the British House of Lords briefly brought up an issue raised in an online petition signed by...., but never legislated on it. Our approach will permit us to tap multiple streams of

The advantage of our approach will be that we will tap multi-channel input and response processes. This will allow us to identify fast-tracks and bottlenecks in the representation cycle. Identifying the dynamics of the entire system would inform those interested in providing outside input, those seeking to understand the overall responsiveness of an organization and those interested in affecting the organization's responsiveness.

2 Open Outside Input and Governance

At every level of government in the US, substantial resources have been dedicated to developing online platforms for citizen input to government. The e-Rulemaking **CITE** process is the archetype of these efforts. Federal agencies are required to post proposed rules to the website `regulations.gov` and provide a period for open commenting on the proposed regulation. Another mainstay of government operations in the information age is online tools to provide direct messages to public officials **CITE**. And, recently, President Obama established `Change.gov` to provide for direct citizen input to the administration's activities **CITE**. These developments mark a potential for rapid, massive and innovative "citizen-sourcing" of public policy - a form of democratic representation that is much more timely and rich than that realized through periodic elections and other forms of slower, less interactive input **CITE**.

However, these developments raise several questions about the utility of these input and feedback modes. For example, do the actions of public officials and, ultimately, the content of public policy, reflect the inputs provided by citizens? Do public officials have the capacity to organize and summarize outside inputs? What characteristics of outside input predict the timely integration into public policy? All of these questions are critical to determining the value of these e-government or 'we-government' technologies.

We endeavor to answer these questions and more. Using fine-grained textual and contextual data on several stages of the input-response-feedback cycle, we will assess the dynamics of government responsiveness to open outside input on public policy. This will be made possible through the development of machine learning tools that connect multiple textual streams and a massive database of US county government records assembled through public records requests. This project is headed by a multidisciplinary team, which has already realized success in developing innovative machine learning tools to analyze novel databases on government communication networks.

3 Background: Public Records Data

At the national, state and local levels, the US is the global leader in the scope and reliability of laws that guarantee access to information on government **CITE**. The seminal legislation in this area is the federal Freedom of Information Act, which marks all information produced by executive agencies as public, unless the information meets at least one of seven criteria for exemption. All US states have laws that mimic the federal legislation **CITE**, and most state laws subject localities (i.e, cities and counties) to public records archiving and disclosure requirements. North Carolina offers one of the broadest laws. For instance, email communications among local government officials are established by statute to be public record.

Public record disclosure requirements establish a treasure trove for researchers. Many government organizations post significant text streams - from email communications with elected officials to meeting minutes - directly on websites. For categories of information not posted to the web, it is possible to make requests for the data. At all stages of the democratic process, local governments are required to archive textual records and provide them to the public upon request. This constitutes the primary advantage for focusing on government organizations in developing multi-stream models of the input-response-feedback cycle.

4 Project Team

This project will be completed by a multidisciplinary team of computer scientist (Wallach) with background in developing machine learning methods for the statistical analysis of text and networks and a political scientist (Desmarais) with background in developing large scale statistical and network analytic models of political decision-making within US government institutions. The team is part of the UMass Amherst Computational Social Science Initiative and has already successfully completed and published pilot research that is highly relevant to the proposed project. This project will combine large-scale data collection, innovative methods development and social scientific analysis, resulting in the following deliverables.

- We will develop machine learning methods of statistical text analysis capable of characterizing the relationships between topics in separate dynamic corpora. In addition to extending dynamic topic models to multi-corpora relations, we will integrate corpora-specific-contextual data models. These methods will integrate several separate innovations in statistical topic modeling and provide a comprehensive textual and contextual data modeling framework that captures cross-domain migration of topics.

- We will provide a first-of-its kind, exhaustive micro-level assessment of the function of democratic representation at the local level. The research we have on democratic responsiveness in the US focuses mainly on the federal or state governments. Our research will address local politics. Also, existing research on democratic representation addresses highly aggregated measures: e.g., showing that legislators from highly conservative districts vote, on average, more conservatively than those from more liberal districts. Our work will provide a fine-grained characterization of government responsiveness to the rise of public demands for action on specific public policy topics.

5 Description of Pilot Research

As an initial phase of research for this project, the team endeavored to leverage its expertise in the computational and social sciences, as well as the availability of rich data sources through the public record to address a challenge in the study of communication networks. A substantial body of research finds that the structure of a communication network has a substantial influence on the problem-solving abilities of the network as a whole and the performance of the individuals in the network **CITES**. Moreover, recent research indicates that the topic or context of communication is important, meaning the measurement of a communication network should involve topic-of-discussion specificity **CITES**. Thus, in order for an organization to diagnose and design its internal communication network, it must first be able to discern topic-specific communication networks.

Our pilot research addresses this problem. Noting that email constitutes the cornerstone of most organizations' electronic communications, we focused on measuring and analyzing topic-specific communication networks using email archives. Using a public records request, we collected one months worth of in and out-box contents for all managerial-level employees of New Hanover County, North Carolina. We then developed a model that combined statistical topic modeling and the latent space model of social networks (a model that projects a network into a Euclidean space, within which actors who are close are more likely to connect than those who are far apart). Specifically, the

model learns topic-specific latent spaces. This produces a coherent probabilistic generative model of corpus of messages annotated with sender-receiver information that permits intuitive visualization and analysis of the underlying topic-specific communication networks.

6 Proposed Research

We will advance machine learning methods for dynamic natural language processing and, using the methods we develop, provide a novel and insightful analysis of democratic responsiveness at the county level of government in the US. The tools we develop will be applicable to any level of government and other organization types. Our analysis of county governments will serve as a proof of concept and prototype for the use of our methods. Also, we will assemble a comprehensive multifaceted database of textual data on county governments that will prove useful to other researchers in natural language processing and the social sciences.

In the following sections we describe our approach to modeling the multi-domain cycle of input-response-feedback by first describing proposed datatypes and domain-specific model specifications then discussing our approach to tying the different domains together into a multi-component system.

6.1 Modeling Outside Input

To define a model for citizen input it is important to first identify the form of the data we will gather on citizen input. Via public records requests, we will collect all emails sent to county government officials by those outside of government. Though some government organizations have experimented with different electronic input modes, the email message is still the workhorse of direct advocacy **cite**. We actually do not need to submit a separate request for this data, as they come with the in and out-boxes of government officials.

We will model citizen input with a dynamic author-recipient topic model. The author-recipient topic model learns an overall topic model for the entire corpus as well as deviations for author-recipient pair **CITE**. We will consider both complete-email-address identifiers of 'authors' as well as email address domains as author signature - which may better handle input sent from the same organization but different people. The dynamic component will be handled with a logistic-normal model of topic dynamics akin to **CITE**. This approach will permit us to understand (1) the overall content of citizen input, (2) source-specific peculiarities in the content of input, (3) recipient-specific peculiarities of input content and (4) the over-time change in the content of input.

6.2 Modeling Informal intra-governmental Communications

The data for informal intra-governmental communications will also utilize the email data collected through public records requests. Since we will have all email for the officials, we will be able to model the complete communication network over time. In this phase of the model we will make the most explicit use of the team's pilot research. The model will constitute a dynamic topic-specific model of communication networks with the networks projected into Euclidean latent spaces. The topic dynamics will again be modeled using a logistic-normal specification. This characterization will permit us to understand how the content and network context of intra-governmental communications evolve.

6.3 Modeling Formal intra-governmental Communications

Policy development that takes place at the county level occurs within county legislatures. Digitized archives of legislative meeting minutes are typically available on the web, but are certainly accessible via public records requests. If public records requests are fulfilled via paper delivery, we will use OCR to digitize the meeting minutes. This data will offer a window into whether issues that are being communicated to the government from outside actors and/or topics that arise through informal intra-governmental communications make their way to the legislative agenda.

Our modeling approach with regard to legislative meeting minutes is informed by the political science literature on decision-making in legislatures. Most of this research focuses on the process of coalition-building in legislative processes **CITES**. Since the vast majority of legislatures require majority support to establish policy, the primary task of a legislator is to lobby, placate and persuade his or her colleagues regarding proposed legislation. The coalition-building process gives rise to factionalism, which induces high positive association between the priorities of those in the same faction and high negative association between the priorities of those in different factions.

We will implement an integrated version of the correlated topic model **CITE** and author topic model **CITE** and take advantage of the multiple-meeting structure of the data to assess the associations underlying the generation of meeting proceedings. Like the author topic model, there will be a common set of topics discussed by all legislators at a given point in time, but proportional attention to topics will vary across legislators. The seminal correlated topic model embeds a correlation structure among topics. We will instead examine correlation between legislators' attention to topics. Specifically, we will parse each meeting into statements made by each county legislator. Then, we will fit a logistic-normal parameterization of the author-topic model that estimates a covariance matrix over legislators regarding their attention to topics in each meeting. This will capture the varying priorities among legislators and the correlations among them that arise through the legislative bargaining process.

6.4 Modeling Legislative Output

We will gather data on the statutes created by each county in order to examine implementation of topics that arise in the public, informal intra-governmental communications, and on the legislative agenda. These data are available on the web for most counties, and are certainly a matter of public record and would be available upon request. For some counties, it is straightforward to extract how legislators voted on a measure from the meeting minutes (e.g., see **Figure 2**). For other counties, however, this is not as simple and may not be feasible.

Regardless of whether we can extract votes on legislation, we will use a dynamic topic model with a logistic-normal autocorrelation structure **CITE** to model the content of legislation. If the votes are available, we will augment the dynamic topic model with a topic-specific network structure similar to that in **CITE**. The network will, however, take on a different form. The fully-visible Boltzmann Machine constitutes a probability model of jointly observed binary switches that is parameterized to model the tendency for each switch to be 'on' as well as the pairwise association between the states of each dyad of switches. This can be used to model binary votes with voters akin to switches **CITE**. This will allow us to jointly model the attention to topics in legislation as well as the associations among legislators' final votes on policy.

7 Timeline and Division of Labor

Here we give a clear breakdown of how the work will be divided and what will be completed.

8 Broader Impact

Here we discuss broader scientific impact, training and education, as well as potential non-academic impacts.

9 Results From Prior NSF Support

Data Management Plan

A. Project Information

B. General Data Management Plan Information

C. Policies

D. Legal Guidelines and Requirements

E. Access, Sharing and Re-use of Data

F. Data Standards and Capture

G. Security, Storage, Management and Back-Up of Data

H. Preservation, Review and Long-Term Management of Data

Biographical Sketch: Bruce A. Desmarais

(a) Professional Preparation:

- Eastern Connecticut State University, Economics and Public Policy, B.A. (2002)
- University of North Carolina at Chapel Hill, Political Science, M.A. (2008)
- University of North Carolina at Chapel Hill, Political Science, Ph.D. (2010)

(b) Appointments:

- *University of Massachusetts Amherst* Assistant Professor, 2010 - Present

(c) Publications

(i) Publications Directly Related to the Proposed Project

- Cranmer, Skyler J. and Bruce A. Desmarais. 2011. "Inferential Network Analysis with Exponential Random Graph Models." *Political Analysis.* 19(1): 66-86.
- Desmarais, Bruce A. and Skyler J. Cranmer. 2012."Statistical Inference for Valued-Edge Networks: The Generalized Exponential Random Graph Model" *PLoS-ONE.* 7(1):e30136.
- Desmarais, Bruce A. and Skyler J. Cranmer. 2012. "Statistical Mechanics of Networks: Estimation and Uncertainty." *Physica A* 391(4): 1865-1876.
- Desmarais, Bruce A. and Skyler J. Cranmer. 2012. "Micro-Level Interpretation of Exponential Random Graph Models with Application to Estuary Networks" *Policy Studies Journal.* 40(3): 402-434.
- Cranmer, Skyler J., Tobias Heinrich, and Bruce A. Desmarais. Accepted 2012. "Reciprocity and the Structural Determinants of the International Sanctions Network." *Social Networks.*

(iii) Other Significant Publications

- Desmarais, Bruce A. 2012."Lessons in Disguise: Multivariate Predictive Mistakes in Collective Choice Models." *Public Choice.* 151(3–4): 719-737..
- Cranmer, Skyler J., Bruce A. Desmarais, and Elizabeth J. Menninga. 2012. "Complex Dependencies in the Alliance Network" *Conflict Management and Peace Science* 23(3).
- Cranmer, Skyler J., Bruce A. Desmarais, and Justin H. Kirkland. 2012. "Towards a Network Theory of Alliance Formation" *International Interactions.* 38(3): 295-324.
- Harden, Jeffrey J. and Bruce A. Desmarais. 2011. "Linear Models with Outliers: Choosing Between Conditional Mean and Conditional Median Methods" *State Politics and Policy Quarterly.* 11(4): 371-389.
- Desmarais, Bruce A. and Jeffrey J. Harden. 2012. "Comparing Partial Likelihood and Robust Estimation Methods for the Cox Regression Model" *Political Analysis.* 20(1): 113-135.

(d) Synergistic Activities

- Participated in the founding and administration of the Triangle Political Methodology Group (2009–2010) – `http://www.unc.edu/depts/polisci/methods/`.
- Co-organized an interdisciplinary speaker series in Computational Social Science at UMass Amherst (2010–Present) – `http://cssi.umass.edu/seminars.html`.
- Editorial board member, *State Politics & Policy Quarterly*, (2011–Present).
- Member of the fellowship committee for the 2012 Political Networks Conference.

(e) Collaborators and other Affiliations

(i) Collaborators

- Skyler Cranmer, University of North Carolina at Chapel Hill
- Jeffrey J. Harden, University of North Carolina at Chapel Hill
- Hanna Wallach, University of Massachusetts Amherst
- Brian Schaffner, University of Massachusetts Amherst
- Vincent Moscardelli, University of Connecticut
- Tobias Heinrich, Rice University
- Allison Freeman, Center for Community Capital (UNC Chapel Hill)
- Elizabeth Menninga, University of North Carolina Chapel Hill
- Justin Kirkland, University of North Carolina, Chapel Hill
- Rachel Shorey, University of Massachusetts Amherst
- Stuart Benjamin, Duke University
- Peter Krafft, University of Massachusetts Amherst

(ii) Graduate and Post-Doctoral Advisors

- Thomas Carsey, University of North Carolina Chapel Hill
- Skyler Cranmer, University of North Carolina Chapel Hill
- James Stimson, University of North Carolina Chapel Hill
- Kevin McGuire, University of North Carolina Chapel Hill
- Isaac Unah, University of North Carolina Chapel Hill

(ii) Thesis Advisor

- Rachel Shorey, UMass Amherst Computer Science M.S. Student
- Peter Krafft, UMass Amherst Computer Science M.S. Student
- James Aaron, UMass Amherst Political Science Ph.D. Student
- Michael Kowal, UMass Amherst Political Science Ph.D. Student

Total number of graduate students advised: 4

Budget Justification

Senior Personnel

Other Personnel

Fringe Benefits

Travel

Other Direct Costs

Indirect Costs

Facilities, Equipment and Other Resources

Laboratory

Clinical

Computing

Office

Major Equipment

Other Resources