Organizational Responsiveness to Open Outside Input: 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. In addition, electronic communications have drastically improved the ability of organizations to document and communicate their internal proceedings. These complimentary developments have ushered into governance what has been termed 'we government' [1]. Most elected officials can be directly contacted electronically through simple internet tools. Citizens can advertise and sign petitions on the web and attend internet 'town meetings' with their representatives [2] . 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.

The unprecedented scale, regularity and flexibility of external-internal relations available to contemporary organizations presents an enormous opportunity for systematic, real-time, deep analytics. The emerging field of computational social science seeks to address the most pressing social scientific problems through the development of computational and statistical methods that are ideally tailored to the research task at hand. We propose to develop machine learning methods that are precisely tailored to modeling the *complete* input, deliberation, policy formation and implementation cycle documented in the electronic record surrounding an organization. Critically, the models we develop will jointly represent the content of relevant streams of text (i.e., topics) and the socio-organizational structure surrounding text generation (i.e., networks). In this collaborative project, which brings together researchers from computer science and political science, we will develop novel methods and apply them to the analysis of government responsiveness to public input across many US county and city governments.

In this project we will develop and apply novel quantitative methods for identifying the cycle of input, response and feedback that leaves its fingerprint on 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 embedded in different socio-organizational structural contexts (e.g., activist messages sent from independent citizens and regulations produced by a legislative body). First, in democratic societies there is a common expectation that the government will respond to public demands. Second, the input mode on which we will focus– direct email from the public to government officials – is regularly central to government efforts to encourage direct input to the policy creation and implementation processes. 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 four different phases in the cycle of governance – illustrated in Figure 1 – with four different types of textual streams and sociopolitical organizational contexts: 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 multi-scale multi-corpora dynamic Bayesian latent variable models of textual content (i.e., topics [3]) and organizational structural contexts (i.e., networks). We will develop (1) several domain-specific models of dynamic content and structure that are tailored to the unique characteristics of each corpus and (2) develop a flexible framework for constructing multi-corpora dynamic models that tie the individual domains together within a modular, extensible framework. The result will be an analytical approach that enables the systematic investigation of the dynamics of input, responsiveness and feedback surrounding an organization 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., does organizational attention to a topic respond to outside input, how does an organization adapt to the rise of issues that are novel relative to its current foci, is external input being routed to those in the organization who focus on the topics of the input, are those within the organization who talk about the same topics talking to each other?

Statistical topic models automatically infer groups of semantically-related words, known as topics, from word co-occurrence patterns within documents. A single topic is characterized by a discrete distribution over some vocabulary. Thus every word in the vocabulary is associated with every topic, albeit with varying probabilities. Given a corpus of documents, statistical topic models simultaneously infer the composition of the topics that best describe that corpus, as well a document-specific distribution over these set of topics for each document. In this way, every document is probabilistically associated with every topic. [4]. Since the seminal work on statistical topic models [3], the basic framework has been extended and adapted to model additional structures and features relevant to textual corpora; including author-specific distributions over shared topics [5], dyadic (i.e., author-recipient) aspects of messages [6], the underlying communication network [7], and joint text-metadata models of documents [8]. Dynamic topic models [9] 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 . In the current project, we propose to develop new Bayesian latent variable models that draw upon the now robust framework of specifying complex topic models that represent textual content of corpora and contextual structure surrounding the generation of the text. This ambitious agenda will involve 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.

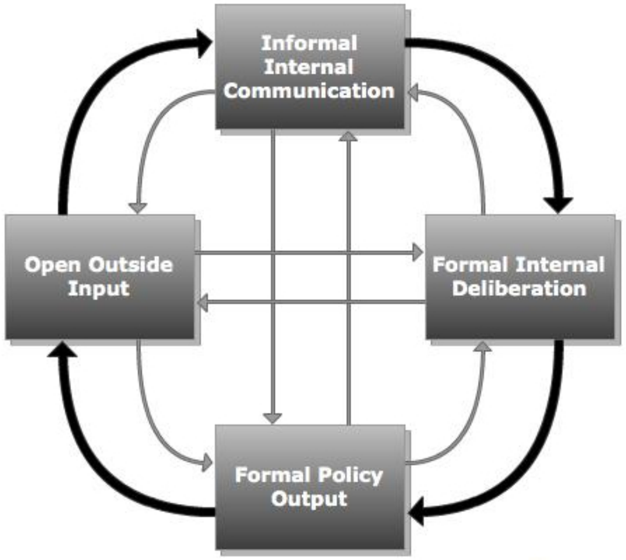


Figure 1: Cycle of input, response and feedback. Black lines denote the normal IRF process and the gray lines represent alternative connections between domains. Direction of arrow indicates the temporal ordering of issue migration.

Figure 1 illustrates the cycle of organizational responsiveness that we intend to model through the guise of co-evolving textual streams. The black arrows indicate the normal progress of an issue from an external stimulus to a formal policy implementation. However, virtually any domain-domain connection is possible, and to understand the whole process of external-internal relations, it is critical to simultaneously incorporate all pathways of topical flow. 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 [10, 11, 12]. There is also a large body of research focusing on legislative adaptation to broad ideological trends among constituents [13, 14, 15]. And yet another literature that addresses the processes by which topics rise from informal awareness among officials (i.e., bureaucratic agency employees) and outside parties to the legislative agenda [16, 17, 18]. However, little – indeed, as far as we are aware, no – research has endeavored to build a complete model of the democratic process. 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, democratic representation has not run its complete course. More practically, if common input topics arise in several corners of government, but internal officials do not connect regarding those topics, an issue may never rise to the legislative agenda. And, initial research indicates that even the most advanced designs for encouraging direct input (i.e., e-petitions that trigger mandatory legislative attention) may end up having little to no policy influence [19].

In this project we will develop and apply a framework for deriving a complete, multivariate, and interdependent model of external input, internal deliberation and organizational response. We will use topics represented in texts related to each domain of governance as the binding substance - building bridges between stages of organizational activity and decision-making. 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 Substantive Domain: 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. E-Rulemaking [20] – the process by which proposed regulations are posted to the internet and publicly deliberated on the web – is the archetype of these efforts. US 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 [21]. Another recent example of open outside input was provided by President (elect) Obama in 2008. He established `Change.gov` during the transition to provide for direct citizen input to the administration's future priorities and activities [22]. 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 [1].

Open questions regarding open input: These developments, which utilize information technology to build a direct bridge between policymakers and their constituents raise several critical questions about their utility. We will develop and apply novel quantitative methods that answer questions such as these:

- 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 socio-organizational 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 Bayesian latent variable models and corresponding inference algorithms that connect multiple textual streams and a massive database of US local government records.

3 Contributions: Advancing Computational Social Science

The PIs in this project are global leaders in the emerging field of computational social science with substantial publication records in their respective disciplines of computer science and political science, as well as a robust collaborative agenda. The research will be lead by a 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 contributions. These innovations will not only appeal to an audience of interdisciplinary scholars in computational social science, but will also constitute important central advancements in computer science and political science.

- **Central computer science contributions:** Offering substantial contributions to the field of machine learning, we will develop novel Bayesian latent variable models capable of characterizing the relationships between topics in separate dynamic corpora and the socio-organizational contexts within which the text was generated. The contributions offered by these innovations will be of two types: (1) domain-specific dynamic Bayesian latent variable models of textual content and socio-organizational context and (2) the development and implementation of a flexible and extensible approach to building models that represent dynamic interdependence across corpora. These methods will provide a comprehensive textual and contextual data modeling framework that captures cross-domain migration of topics.

- **Central social science contributions:** 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.

4 Description of Pilot Research

As an initial phase of research for this project, we endeavored to leverage our 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. this work has been presented at a variety of computational social science workshops—including the Workshop on Information in Networks, New Directions in Analyzing Text as Data, and the Annual Political Networks Conference—and appeared at Neural Information Processing Systems conference (one of the top machine learning conferences) in late 2012. 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 [23]. 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 [24]. 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 [7]. 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 month's 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. Figure 2 illustrates the results from our pilot topic-partitioned communication network analysis. These are four selected topic-network-spaces. The word types listed at the top of the plots are the five most likely for the respective topic. Nodes are placed closer together in the inferred latent space if they are more likely to exchange emails on the respective topic. This modeling approach provides estimates that permit intuitive exploration of the structure of a communication network within a given topic.

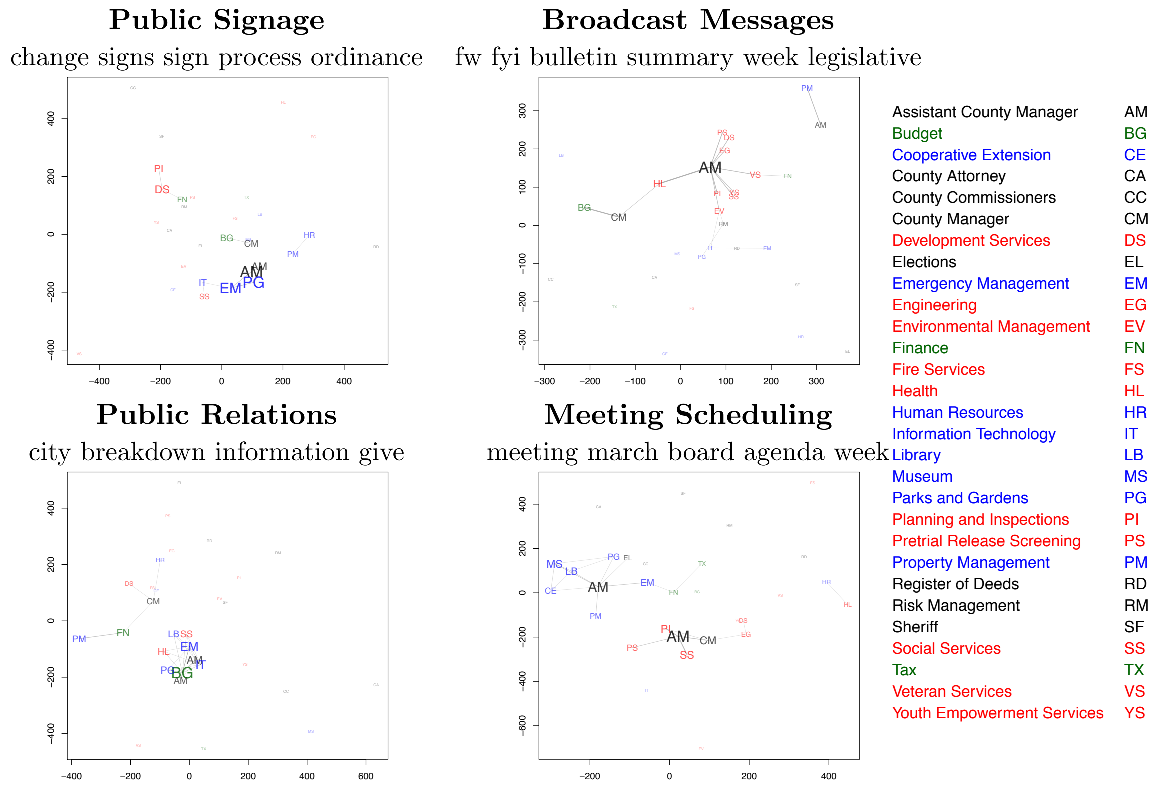


Figure 2: Four topic-specific communication patterns inferred from the NHC email network. Each pattern is labeled with a human-selected name for the corresponding topic, along with that topic's most probable words in order of decreasing probability. The size of each manager's acronym in topic t's pattern (given by $0.45 + 1.25\sqrt{d_a^{(t)} \,/\, \max_a d_a^{(t)}}$, where $d_a^{(t)}$ is the degree of actor a in that subnetwork) indicates how often that manager communicates about that topic. Managers' acronyms are colored according to their respective division in the New Hanover County organizational chart. The acronym "AM" appears twice in all plots because there are two assistant county managers.

5 Proposed Research

We will develop new machine learning methods and associated software for dynamic text analysis. These methods will be specifically intended to provide a novel and insightful analysis of democratic responsiveness at the county level of government in the US. Our methods and tools 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. We will also assemble a comprehensive multifaceted database of textual data on county governments that will prove useful to other researchers in machine learning, natural language processing, and the social sciences.

In the following sections we describe our proposed approach to modeling the multi-domain cycle of input-response-feedback. First, we discuss our ideas for tying different domains together into a multi-component system. Then we describe our proposed data sources and outline some initial domain-specific model specifications. These model specifications serve to illustrate the kinds of models that we propose to develop and deploy over the course of the proposed project.

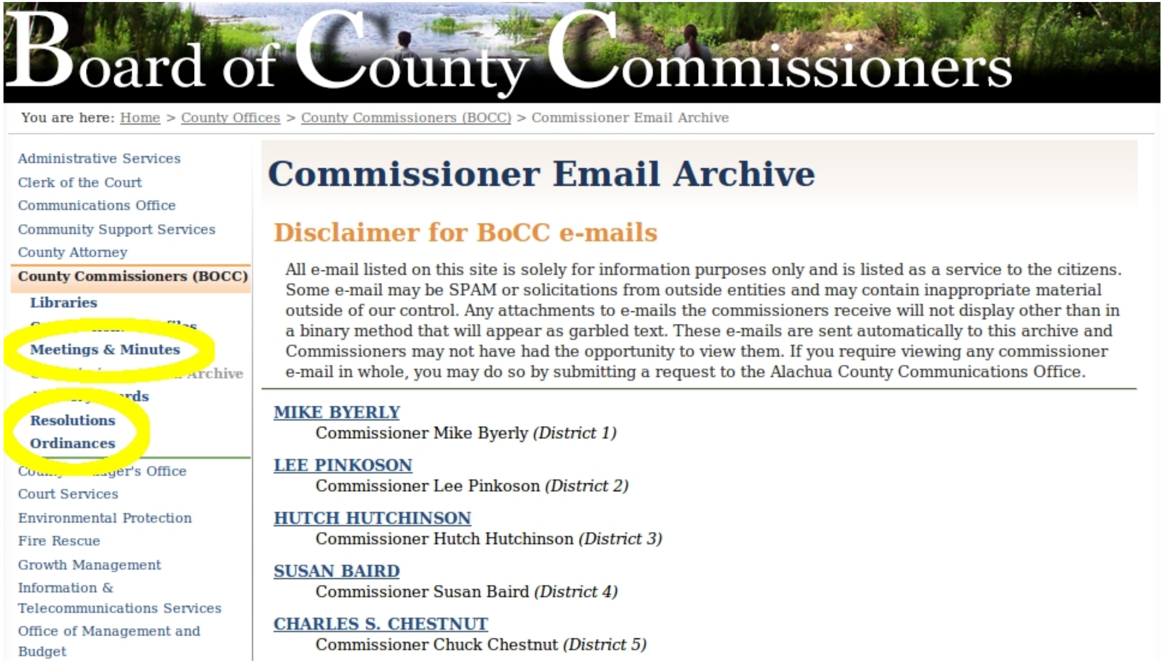
5.1 Public Records Data Collection

Background: 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 [25]. 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. Most US states have laws that mimic the federal legislation [26], and most state laws subject localities (i.e, cities and counties) to electronic public records archiving and disclosure requirements [27].
Benefits: 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 [27]. This constitutes the primary advantage for focusing on government organizations in developing multi-stream models of the input-response-feedback cycle.
Procedure: For most localities, legislative proceedings and legislative output are available on the city and county websites. Archives of external emails sent to government officials and internal emails are typically not posted on websites, though some counties and cities do indeed post them all. The single data source that we plan to use in this project, which requires government officials' compliance with public records requests is government email. We have had considerable success in gathering email corpora via public records requests for our pilot research (described below). For the pilot data collection, we have successfully gathered and processed one-month archives of government officials' in and out-box contents from North Carolina counties including New Hanover, McDowell, Transylvania, Columbus and Moore. With dedicated resources, we are confident we would collect data from most of the 100 counties in North Carolina as well as county and city governments in other states.
Robustness: However unlikely, it is certainly within the realm of possibility that the demeanor of local governments toward large scale public records requests will change course. If this happens, we will still be able to gather more than enough data to carry out the project. This is because for many localities, including several counties in Florida and cities in North Carolina, California and Oregon, all of the communication records that are required for our research are available on government websites. These sources include emails to and between government officials, minutes of legislative meetings, and records of adopted policies. Figure 3 illustrates the web interfaces to this data for two such localities, Alachua County, Florida and the City of Corvallis Oregon. As can be seen from these intuitive and open web interfaces, ample data required to model the interrelationships among domain-specific corpora are readily available.

5.2 Integrating Domain-Specific Models

Below we describe several types of domain-specific models that constitute innovative analytic representations of textual data connected to individual phases in the governance process. Before delving into domain-specific modeling frameworks, we describe a general approach to tying the domains together into an joint probabilistic model that incorporates cross-domain and over-time interdependence. Each of our proposed domain-specific model families draws upon ideas from statistical topic modeling in order to model domain-specific textual content. We seek to understand the re-

Alachua County, Florida
url: `http://www.alachuacounty.us/Depts/BOCC/Pages/CommissionerEmailArchive.aspx`



Board of County Commissioners

You are here: Home > County Offices > County Commissioners (BOCC) > Commissioner Email Archive

- Administrative Services
- Clerk of the Court
- Communications Office
- Community Support Services
- County Attorney
- County Commissioners (BOCC)
  - Libraries
  - [illegible] files
  - Meetings & Minutes
  - [illegible] Archive
  - [illegible] rds
  - Resolutions
  - Ordinances
- [illegible] ger's Office
- Court Services
- Environmental Protection
- Fire Rescue
- Growth Management
- Information & Telecommunications Services
- Office of Management and Budget

Commissioner Email Archive

Disclaimer for BoCC e-mails

All e-mail listed on this site is solely for information purposes only and is listed as a service to the citizens. Some e-mail may be SPAM or solicitations from outside entities and may contain inappropriate material outside of our control. Any attachments to e-mails the commissioners receive will not display other than in a binary method that will appear as garbled text. These e-mails are sent automatically to this archive and Commissioners may not have had the opportunity to view them. If you require viewing any commissioner e-mail in whole, you may do so by submitting a request to the Alachua County Communications Office.

MIKE BYERLY
Commissioner Mike Byerly *(District 1)*

LEE PINKOSON
Commissioner Lee Pinkoson *(District 2)*

HUTCH HUTCHINSON
Commissioner Hutch Hutchinson *(District 3)*

SUSAN BAIRD
Commissioner Susan Baird *(District 4)*

CHARLES S. CHESTNUT
Commissioner Chuck Chestnut *(District 5)*

Figure 3: Web Interface to Alachua County, FL email archive. Links to legislative deliberations (i.e., meeting minutes) and legislative outputs (i.e., ordinances and resolutions) are circled in yellow. Most local governments provide online access to meeting minutes and legsilation; several others provide direct open access to government emails. See, e.g.,
Corvallis, Oregon – `http://www.corvallisoregon.gov/index.aspx?page=65`
Menlo Park, California – `http://ccin.menlopark.org:81/`
Sarasota, Florida – `https://www.scgov.net/email/Pages/default.aspx`

lationships among the topic distributions, which will provide insight into the ways in which topics progress through the different phases of governance. In addition to textual content, each class of domain-specific models also represents the organizational structural context unique to that domain. This approach enables us to assess the relationships between textual content and structural context.

We work through the description of our proposed integrated modeling framework using the graphical model shown in Figure 4. Let K be the number of topics. Each topic is characterized by a multinomial distribution over words with natural parameter vector β. The multinomial distribution over topics (i.e., topic proportions) for each document, denoted by θ, is given by the mean transformation of the natural parameter vector η. Inter-temporal and inter-domain dependence are captured by drawing η from a natural parameter drawn from a Gaussian vector autoregression model [28] (i.e., $\eta \sim \mathcal{N}\left(\alpha_t, a^2 I\right)$ and $\alpha_t \sim \mathcal{N}(\mu_t, \sigma^2 I)$). Given M domains, the mean natural parameter corresponding to topic k in domain m at time t is

$$\mu_{(t,m,k)} = \lambda_{(0,m,k)} + \lambda_{(1,1,k)}\alpha_{(t-1,1,k)} + \lambda_{(2,m,k)}\alpha_{(t-1,2,k)} + \ldots + \lambda_{(M,m,k)}\alpha_{(t-1,M,k)} + \epsilon_{(t,m,k)},$$

where $\lambda_{(0,m,k)}$ is an over-time intercept for the natural parameter corresponding to topic k in domain m, $\lambda_{(v,m,k)}$ is a real-valued parameter that gives the regression relationship between the natural parameter of topic k in domain v at time $t-1$ and the natural parameter of topic k in domain m at time t and $\epsilon_{(t,m,k)} \sim \mathcal{N}(0, \sigma^2_{m,k})$. This parameterization renders the natural parameter corresponding to a given topic (i.e., the prominence of a given topic) in a selected domain dependent upon the natural parameter of that same topic across all other domains in the immediate past. Through λ, this vector autoregressive model represents the dynamics of topic progression across domains. This derivation constitutes the parameterization of over-time and cross-domain dependencies through the use of logistic-normal distributions [9] with vector autoregressive means to define the natural parameters of the multinomial distribution over topics. We will investigate both variational and sampling-based inference algorithms for this framework, drawing on previous work of PI Wallach and others [9, 29].

Another important feature of our framework is that each domain-specific model family incorporates two types of document metadata (i.e., associated organizational context). The first type (x) represents fixed data on which we condition the topic distributions. This is often referred to as 'upstream' conditioning of topic models on metadata [30, 8]. The other metadata type (y) is integrated into the model through a downstream probabilistic representation that conditions the generative model for metadata on the topic of the document via parameters τ. In the interest of generality, we leave the way in which x conditions the topic distribution and the way in which the topic assignment affects τ without complete specification, since the structure and parameterization will vary considerably over the M domains, as described below.

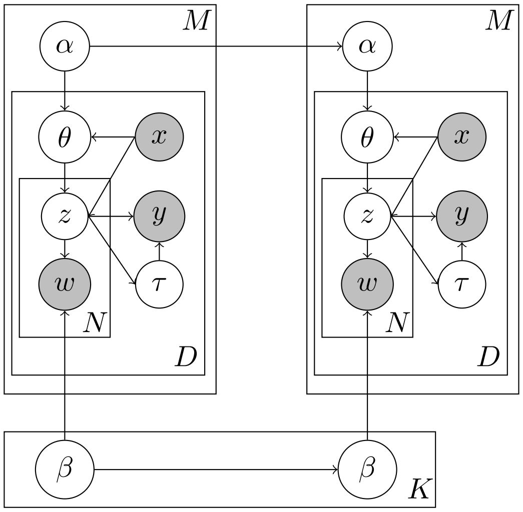


Figure 4: Graphical representation of a dynamic multi-corpus model of text (w) and context (y and x) over two time periods. Each topic's natural parameters (β) evolve over time. The logistic normal determinants of topic proportions (α) co-evolve across the M domains.

The successful implementation and application of this modeling framework will offer government organizations a complete picture of the flow of topical foci throughout different phases in the policymaking process. The biggest challenge to realizing these ends will be the development of scalable and efficient approaches to inference. As we describe in greater detail below, several aspects of the modeling frameworks we will apply pose distinct technical obstacles to implementation on large, real world datasets. Recent developments in stochastic variational [31, 32, 33] and pseudolikelihood approximation [34, 35] will be combined, compared and exhaustively vetted on the variety of datasets we will collect.

5.3 Modeling Outside Input

Data description: 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.

Model description: We will model citizen input by developing a new dynamic cluster–recipient topic model. This model will draw upon ideas from dynamic topic modeling [9], as well as previous work on modeling email [6], including our pilot research [7]. McCallum et al.'s author–recipient topic model learns shared topics for the entire corpus as well as different levels of attention to each topic for each author–recipient pair [6]. For our data, learning specific author–recipient distributions is not appropriate as many of the authors of outside input messages will send only a handful of emails. It will therefore be advantageous to aggregate over several like-minded authors. We will group outside contacts into clusters such that each author belongs to one of the clusters with some probability (i.e., a mixture model) and each cluster is attentive to each topic with varying degrees. We intend to use nonparametric Bayesian clustering methods, which obviate the need to select the number of clusters a priori. PI Wallach has significant expertise in this area [36, 37]. Within the domain of governance, this clustering approach maps very well onto the concept of partisanship. The dynamic properties of the model will be handled using a logistic–normal model of topic dynamics [9]. 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.

Computer science contribution: The primary computer science contribution offered by work on this model will be the development of a new Bayesian latent variable framework for modeling directed communications (described above) and associated inference algorithms. This multifaceted focus provides a coherent approach for dealing with author sparsity and identifying like-minded authors.

Political Science contribution: The primary political science contribution offered by the application of this model will be an analysis of issue (i.e., topic) factions. This will offer a partition of authors that is very consistent with the concepts of partisan faction. Moreover, this analysis will set the framework for differentiating the influence of different factions of authors on the legislative agenda.

5.4 Modeling Informal intra-governmental Communications

Data description: 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.

Model description: In this phase, we will make the most explicit use of the team's pilot research. We will develop a new dynamic topic-specific model of communication networks, in which the networks are projected in to Euclidean spaces to facilitate visualization and exploration. 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.
Computer science contribution: The primary computer science contribution offered by this

component is the development of a new dynamic model (and associated inference algorithm) that builds upon the model developed by the PIs in their pilot research [7] to a dynamic context. We will extend this model to incorporate over-time (Markov) drift in the vertices' positions in the latent space [38], which will enable us to forecast evolution in communication network structure. Recent work has shown that inference via variational approximation drastically improves the scalability of the latent space model for social networks [39]. We will develop and implement a version of variational approximation for use with our latent space topic model of communication networks.

Political Science contribution: The primary political science contribution offered by the application of this model will be an analysis of how government communication network structure and the content of communications co-evolve. This will provide an opportunity to understand when, in the lifecycle of topic attention, network structure changes. Given the literature on the optimal configuration of communication networks [24, 23], these results will permit us to assess how long it takes a government communication network structure around a given topic to configure into an efficient network.

5.5 Modeling Formal intra-governmental Communications

Data description: Policy development that takes place at the county level occurs within county legislatures. Digitized archives of legislative meeting minutes are typically (indeed in every locality that we have checked) available on the web, but are certainly accessible via public records requests. 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.

Model description: 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 [40]. 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 develop new statistical topic models that build upon the correlated topic model [41] and the author–topic model [5] in order to 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.

Computer science contribution: The primary computer science contribution offered by work on this model will be the development of a statistical topic model (and associated inference algorithm) that incorporates interdependence among authors. This will constitute an indispensable component of our input-response-feedback modeling, but would also be independently applicable to countless real-world contexts involving contemporaneous text generated by the same authors over many instances in time (e.g., any sort of meeting for which the minutes are recorded).

Political Science contribution: The primary political science contribution offered by the application of this model will be an analysis of how interdependence manifests in legislative deliberation. We will be able to answer such questions as:

- Do legislators of the same political party focus on the same issues in legislative deliberations?
- Do legislators of the same gender and/or ethnicity focus on the same issues in legislative deliberations?

5.6 Modeling Legislative Output

Data description: 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. For other counties, however, this is not as simple and may not be feasible.

Model description: Regardless of whether we can extract votes on legislation, we will use a dynamic topic model with a logistic–normal autocorrelation structure to model the content of legislation. If the votes are available, we will augment the dynamic topic model with a topic-specific network structure. 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 [42]. This can be used to model binary votes with voters akin to switches, as illustrated by work co-authored by PI Desmarais that analyzes voting on the US Supreme Court [35]. This will allow us to jointly model the attention to topics in legislation as well as the associations among legislators' final votes on policy.

Computer science contribution: The primary computer science contribution offered by work on this topic is the development of a statistical model that simultaneously models the content of a document annotated with a vector of choices and the associated inference algorithm. The model for this domain poses another distinct inference challenge. The problem of inference in Boltzmann Machines is substantially complicated by the intractability of the normalizing constant, with scalable inference being offered by pseudolikelihood approximation [43]. PI Wallach has expertise in sampling based approaches to likelihood approximation [44]. We will offer an important contribution by developing scalable inference for this model. This model will be applicable to any context in which there is voting on a policy documented in text.

Political Science contribution: This model will enable us to assess the likely degree of agreement among legislators surrounding legislation – using topic-specific Boltzmann machines – at any point in time. The primary political science contribution offered by the application of this model will be an analysis of how topics and agreement surrounding topics co-evolve. Of high specific interest will be whether attention to legislation is associated with the degree of agreement among legislators on a given topic.

5.7 Overall Project Outputs

Each of the modeling phases described above constitutes a novel contribution to the literature on natural language processing. The four domain-specific models have not appeared in the literature in their proposed form. Each would therefore constitute a publishable innovation that might appear in such venues as NIPS, AISTATS, and ICML. Decisions about whether to include multiple innovations in the same discrete papers will depend upon the results of empirical experiments, overlap in inferential strategies and several other factors that cannot be perfectly anticipated at this point.

Another, particularly notable, machine learning contribution of this project will be the development of our vector autoregressive framework for building multi-corpora dynamic models of text and context. The framework will be motivated with the problem of modeling organizational responsiveness to outside input and illustrated through the combination of our domain-specific models to model the input-response-feedback cycle in local governments. This will constitute standalone research that is publishable in high profile machine learning conferences.

The major social scientific contributions will come in the form of several analyses of how the sociopolitical contexts and textual content co-evolve. The domain-specific models will individually provide for important innovations in our understanding of government and local government processes. For instance, the analysis of email networks will permit us to assess whether government communication networks are structured to effectively solve problems. As a second example, the legislative meeting minute models will be useful in assessing whether political discussion is as factional at the local level as it is in the US Congress [45]. The biggest contribution to political science offered by this project will be the analysis of the cross-domain relationships inferred using the vector autoregressive logistic-normal model. We will be able to speak in precise detail to a fundamental question of growing important to political scientists and practitioners alike – how do governments respond to open outside input?

6 Broader Impacts

The proposed project will offer broad impacts that advance the core societal mission of the Information Integration and Informatics (III) program, directly enhance educational offerings at the University of Massachusetts and elsewhere and provide valuable interdisciplinary training opportunities related to an emerging area of research – computational social science.

6.1 Societal Mission

The III program supports the development of computational tools and analytical approaches that enable the massive, diverse and complex streams of data to be efficiently utilized to produce scientific, technical and societal advances. Our proposed research addresses this mission, precisely. We will develop methods that enable government organizations to assess their responsiveness to electronic open outside input and diagnose internal obstacles to information flow using corpora of data that are already archived for other purposes (e.g., electronic messages, formal meeting records, distinct policy outputs). We are committed to directing the project so as to optimize the practical applicability of the methods we develop. To this end, we have recently established contact with the Analytics Unit of the Office of Policy and Strategic planning in the Mayor's office of New York City – a unit that has been widely recognized for the use of sophisticated analytics to guide government organization and activity [46]. We have included funds in the budget to support consultation with the NYC government throughout the project in order to assure we are developing tools that offer maximal benefit to government organizations.

6.2 Education and Training

Both PIs are strongly committed to bringing their research into the classroom. PI Wallach regularly teaches a graduate seminar titled, 'Computational Social Science' in which state-of-the art innovations in computational social science are studied at-length. The results of the proposed project would certainly be integrated into the curriculum. PI Desmarais teaches graduate courses in network analysis at the University of Massachusetts and the University of Michigan (summer program). These courses integrate up-to-date methodological innovations in his research. The course material will be updated to reflect the contributions of the proposed project. PI Desmarais has also contributed to the NSF-supported Online Portal for Social Science Education in Methodology (OPOSSEM) - an open source online archive of methodological instructional materials. Instructional materials related to the current project would be posted to OPOSSEM.

The proposed project will support a graduate student in computer science and one in political science at the University of Massachusetts, as well as several undergraduate research assistants. The proposed research represents a genuinely interdisciplinary research program in the emerging field of computational social science. Long-term focused training for graduate students will provide valuable research experience and socialization on interdisciplinary projects - a growing focus of the scientific community. Undergraduate RAs will also gain valuable complimentary experiences. Computer science students will have the opportunity to work on technical problems that address highly important sociopolitical problems, and political science students will be directly exposed to the ways in which technical analysis and innovation can enhance knowledge of the public policy process. Encouraging sincere interdisciplinary engagement is a central mission at UMass Amherst, signified by the establishment and growing prominence of the UMass Computational Social Science Initiative (cssi.umass.edu). Support of the current project will provide an important opportunity for graduate students to engage in a multi-year cross-disciplinary collaborative project.

6.3 Broadening Participation of Underrepresented Groups

Both PIs Wallach and Desmarais have records of making concerted efforts to broaden the participation of underrepresented groups. PI Wallach co-founded the Women in Machine Learning Workshop – an ongoing annual meeting (and associated executive board) of scholars that bolsters the small network of female researchers in machine learning. In 2011, preceding the NSF-sponsored 'Visions in Methodology' conference for women in political methodology at The Ohio State University, PI Desmarais offered a one day workshop on statistical inference with network data (featuring the methods he has developed and applied in his research). Given the results of the proposed projects, both PIs will continue to pursue training opportunities that focus on underrepresented groups.

7 Timeline and Division of Labor

We propose to complete this research within a period of three years. The four major tasks will be (1) development and implementation of new statistical models, (2) data collection and cleaning, (3) application of methods and assessment on empirical data, (4) social scientific analysis of results. We divide the project period into quarters and in Table 1, we provide a breakdown of the primary activities of the project, along with an assessment of when the tasks will be completed.

Table 1: Schedule of Project Activities

Lead PI	**Activity**	**Period**
Wallach	domain-specific model derivation and implementation	Q1– Q8
Desmarais	data collection (e.g., public records requests and web-scraping) and organization	Q1– Q3
Wallach	experimentation and assesment with county government data	Q4 – Q8
Desmarais	social scientific analysis of domain-specific results	Q5 – Q8
Wallach	development and application of integrated vector autoregressive logistic normal model	Q6 – Q12
Desmarais	social scientific analysis of integrated vector autoregressive logistic normal model	Q9 – Q12

8 Results From Prior NSF Support

PI Wallach is also a PI on NSF Award #0965436, entitled "Collaborative Research: New Methods to Enhance Our Understanding of the Diversity of Science" (5/15/2010–4/30/2013; $286,940; co-PIs Fiona Murray, MIT, and Andrew McCallum, UMass Amherst). The grant has contributed to the following selected papers:

- D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. "Optimizing Semantic Coherence in Topic Models." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011.
- E. Talley, D. Newman, D. Mimno, B. Herr II, H. Wallach, G. Burns, M. Leenders and A. McCallum. "A Database of National Institutes of Health (NIH) Research Using Machine Learned Categories and Graphically Clustered Grant Awards." In Nature Methods, 2011.
- Passos, A., Wallach, H., McCallum, A. "Correlations and Anticorrelations in LDA Inference." NIPS Workshop on "Challenges in Learning Hierarchical Models: Transfer Learning and Optimization," 2011.
- A. Bakalov, A. McCallum, H. Wallach, and D. Mimno. "Topic Models for Taxonomies." In Proceedings of the Joint Conference on Digital Libraries, 2012.

PI Wallach is also a PI on NSF Award #1036868 entitled "Collaborative Research: Workshop for Women in Machine Learning" (9/01/2010–08/31/2013; $6,330; co-PI Jennifer Wortman Vaughan, UCLA).