Advance Access publication December 3, 2009

Political Analysis (2010) 18:1–35
doi:10.1093/pan/mpp034

A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases

Justin Grimmer
Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: jgrimmer@fas.harvard.edu (corresponding author)

Political scientists lack methods to efficiently measure the priorities political actors emphasize in statements. To address this limitation, I introduce a statistical model that attends to the structure of political rhetoric when measuring expressed priorities: statements are naturally organized by author. The expressed agenda model exploits this structure to simultaneously estimate the topics in the texts, as well as the attention political actors allocate to the estimated topics. I apply the method to a collection of over 24,000 press releases from senators from 2007, which I demonstrate is an ideal medium to measure how senators explain their work in Washington to constituents. A set of examples validates the estimated priorities and demonstrates their usefulness for testing theories of how members of Congress communicate with constituents. The statistical model and its extensions will be made available in a forthcoming free software package for the R computing language.

1 Introduction

I introduce a statistical model to measure the priorities political actors articulate in texts, which I apply to measure how legislators explain their work to constituents. The *expressed agenda model* incorporates information about the authors of texts and other covariates to create a method explicitly designed to measure how legislators articulate priorities to constituents. A Bayesian approach, coupled with the use of a deterministic method for estimating complex posteriors, makes estimation and inference straightforward, as well as inference about quantities of interest derived from the priorities. I apply the model to an original collection of over 24,000 Senate press releases, collected from each Senate office in 2007, demonstrating that the press release data and statistical model facilitate comprehensive tests of theories about how legislators communicate with their constituents.

Members of Congress invest substantial resources to communicate with constituents, issuing thousands of statements, press releases, and speeches during each legislative term.

Author's note: I thank the Center for American Political Studies and the Institute for Quantitative Social Science for financial support. I have benefited from conversations with Ken Benoit, Matt Blackwell, Daniel Carpenter, Jacqueline Chattopadhyay, Andrew Coe, Brian Feinstein, Rob Franzese, Claudine Gay, Jeff Gill, David Hadley, Frank Howland, Emily Hickey, D. Sunshine Hillygus, Daniel Hopkins, Michael Kellerman, Gary King, Burt Monroe, Clayton Nall, Stephen Purpura, Kevin Quinn, Brandon Stewart, seminar participants at Harvard University, participants at the 2008 Summer Political Methodology meeting, and 2009 Southern Political Science Association meeting.

In spite of the recognized importance of this communication to understanding political representation and legislative behavior (Mayhew 1974; Fenno 1978), political scientists know surprisingly little about the content of these statements and how legislators translate their activities in Washington into statements to constituents. This is due, in large part, to the difficultly in collecting and analyzing the multitude of statements from members of Congress. Most studies of congressional communication employ methods that are too expensive and time consuming to apply to each member of Congress or even large samples of members (Fenno 1978; Lipinski 2004). As a result, much of our knowledge about how legislators explain their work to constituents is derived from observations made about a few members of Congress and a small subset of statements made to constituents.

As an alternative to manual coding, political scientists have recently turned to unsupervised learning methods to analyze attention in large text corpora: methods that simultaneously estimate the categories in a collection of texts and sorts documents into the estimated categories (Quinn et al. forthcoming).[1]

These methods, however, cannot be directly applied to measure the priorities articulated by representatives. When analyzing the attention senators (or other political actors) dedicate to issues there is a hierarchical structure: political statements, at the bottom of the hierarchy, are organized according to their author, at the top of the hierarchy. Previously developed unsupervised learning methods ignore this hierarchy and instead focus upon assigning documents to topics (Banerjee et al. 2005) or introduce structure designed to answer a different (but still important) question about how attention varies over time in the whole legislature (Quinn et al. forthcoming).

I accommodate the hierarchical structure when measuring author attention with the expressed agenda model. The method simultaneously discovers the topics in the data, assigns documents to their likely topic, and measures the attention a set of authors dedicate to the estimated topics. Like other unsupervised learning methods, the expressed agenda model does not require any pre-read documents, estimating the topics in the press releases. The use of Bayesian inference and a recently developed approach to estimation of complex posteriors, variational inference, makes fully Bayesian inference straightforward, whereas previous statistical models for political text assume that all estimates are known with certainty (Quinn et al. forthcoming). Using the new data set and statistical model, I demonstrate that the expressed agenda model is able to identify substantively important topics. A series of validations also demonstrates that the expressed agenda model provides substantively interesting estimates of senators expressed priorities and facilitates tests of important hypotheses across all members of a legislature—a previously infeasible task.

2 Expressed Agendas in Legislatures and Politics

By measuring how legislators explain their work in Washington to constituents, the model and data set provide powerful tools for understanding how political representation operates in America.

Legislators employ the resources of their office to portray how they are *responding* to the priorities and concerns of their constituents (and to distract attention from areas where

[1]There is a burgeoning literature that analyzes communication with *supervised* methods (Hopkins and King forthcoming; Hillard, Purpura, and Wilkerson 2008). Supervised methods provide hand-coded documents and predefined categories to a method in order to teach—supervise—the method how to place documents into categories. This approach is not employed here because there is still substantial uncertainty about the topics legislators could raise in conversation with constituents.

they appear less responsive) (Mayhew 1974; Fenno 1978; Kingdon 1989; Arnold 1992). Understanding the contents of these portrayals are critical to understanding *home style* and are inherently important to explaining how legislators maintain their connection with constituents (Fenno 1978; Kingdon 1989; Arnold 1992; Sulkin 2005), assessing deliberative standards of democracy (Gutmann and Thompson 1996; Mansbridge 2003), identifying the causes of the incumbency advantage (Gelman and King 1990), determining how legislators claim credit for resources secured for their state (Mayhew 1974), and understanding how legislators interact with the media (Arnold 2004). The importance of understanding legislators' statements to constituents is most clearly articulated by Fenno when he remarks ''empirical theories of representation will always be incomplete without theories that explain explaining'' (Fenno 1978, 162).

In this paper, I address a critical component of home style (Fenno 1978): *the issues legislators' emphasize in communication with constituents*. This quantity is of theoretical interest for studies of Congressional communication, ranging from qualitative studies of explanation (Mayhew 1974; Fenno 1978) to quantitative studies of senators' communicated issue priorities (Schiller 2000; Sulkin 2005). Although the expressed priorities of legislators (and other political actors) appear in numerous studies, it lacks a common name across applications. Therefore, I call the attention a senator allocates to issues in public statements her *expressed agenda*. It is an *agenda* because it measures the priorities of each senator, as articulated in press releases. Importantly, it is an *expressed* agenda because the attention dedicated to issues in communication are the issues that senators express as their priorities, not necessarily those issues that receive the most attention from senators while in Washington or from their staff. (But as I demonstrate, a senator's expressed agenda is closely tied to other indicators of legislative activity.)

2.1 Substantive Structure and Unsupervised Learning Methods

While intended to measure the priorities legislators and other political actors express in public statements, the expressed agenda model is also a new model for document *clustering*. A large literature in statistics and computer science advocate using clustering methods as an effective method for grouping together documents of similar content (Ng, Jordan, and Weiss 2002; Blei et al. 2003; Quinn et al. forthcoming; Manning et al. 2008). Although each method has performed well on a subset of problems, it is well known that the optimal application of any clustering method requires that the method be tuned to a particular substantive problem. The need for problem-specific methods arises because classification is only possible by making assumptions (a result known as the ''ugly duckling theorem'' [Watanabe 1969]) and because all clustering methods have the same average performance across all possible problems (a no-free lunch theorem [Wolpert and Macready 1997]). Therefore, arguments about whether to adopt a statistical or algorithmic approach to clustering are misplaced: the debate should focus on whether a specific class of clustering methods are well tuned to discover substantively interesting clusters from a particular collection of documents.

One approach to developing problem-specific clustering methods is to construct a hierarchical model, where the hierarchy includes additional information about each of the documents (Blei et al. 2003; Blei and Lafferty 2006; Teh et al. 2006; Mimno and McCallum 2008).[2] A particularly novel example of this approach is advanced in Quinn et al.

[2]Note, that this is distinct from hierarchical clustering, which creates a dendrogram describing a series of partitions in the data that obey some organizing rule (Hastie, Tibshirani, and Friedman 2001).

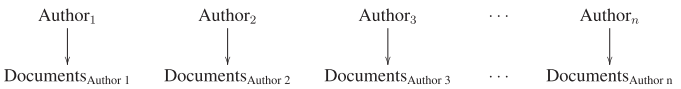

Fig. 1 Hierarchical structure in political texts exploited in expressed agenda model: Documents organized by J.G.

(forthcoming), which analyzes Senate floor-speeches and includes information about the day a speech was made on the Senate floor. This method is therefore tuned to measure how attention varies over time *in the entire legislature.* But this hierarchical structure is ill-suited for the quantity of interest in this paper: *how attention to issues varies across legislators*, because it ignores the authors of particular documents.

The expressed agenda model explicitly includes information about the author of the documents and therefore is designed to address the priorities legislators articulate to constituents. Figure 1 shows the general structure that underlies the model: many statements from each author, with several authors in the collection. While applied to study legislative home style in press releases here, this same structure is employed anytime the *quantities of interest are the priorities a set of actors allocate to issues* and therefore the expressed agenda model has wide range of applications in political science. This includes examinations of issue ownership in political campaigns, where the interest is in comparing the issues Democrat and Republican candidates emphasize during campaigns (Petrocik 1996; Simon 2002; Sigelman and Buell 2004). Likewise, scholars of the news media ask which stories are afforded attention in different newspapers (Armstrong et al. 2006). Scholars of the presidency are often interested in the priorities presidents communicate in their public statements (Lee 2008), and deliberative democrats are interested in exploring the explanations are offered for new policies (Gutmann and Thompson 1996). Table 1 highlights these and other potential applications of the expressed agenda model.

3 Press Releases and Measuring Expressed Agendas

Press releases are an ideal medium for measuring how legislators present themselves to constituents. There are essentially no *formal* constraints imposed upon a press release's content, and they comprise a critical component of how senators explain activities in Washington to constituents (Yiannakis 1982). In contrast, use of media coverage to measure politicians' issue agenda conflates the issues politicians discuss and the media outlet's choice to cover an issue (Sulkin 2005). Surveys of Senate staffers require the strong (and

Table 1 Potential applications of the expressed agenda model

Question	*Example study*
How do campaigns affect attention in congress?	Sulkin (2005)
What do senators discuss in floor statements?	Hill and Hurley (2002)
Do competing candidates emphasize the same issues?	Petrocik (1996)
What do presidents address in daily speeches?	Lee (2008)
What reasons are offered to justify policy?	Gutmann and Thompson (1996)
Do competing political elites discuss the same issues?	Gabel and Scheve (2007)
What issues receive attention from newspapers?	McCombs (2004)

often violated) assumption that the staffer is able to offer an unbiased recollection of a politician's stated priorities (Cook 1988).

Press releases are also ideal because they are regularly used by each Senate office. In 2007, the average Senate office released four and a half press releases per week, and the Senate, as a whole, issued an average of 66 press releases per day. The frequency of press releases stands in contrast to newsletters legislators send to constituents, which are only sent occasionally during a legislative session and have extremely limited space (Lipinski 2004).

3.1 Press Releases and Newspaper Coverage

Press releases are also important because their content reaches citizens through local newspapers. Newspapers—particularly local papers—often have only a small budget dedicated toward covering what representatives do while in Congress (Vinson 2002). To fill this gap, newspaper editors rely upon wire service stories and press releases from Congressional offices (Cook 1989; Arnold 2004; Schaffner 2006).

3.2 "Ventriloquism" and Press Releases

Press secretaries know that they will have high levels of success in generating news coverage with press releases (Cook 1989; Schaffner 2006). In fact, some press releases are run *almost verbatim* in papers: Table 2 collects three press releases (left-hand column) that were subsequently repeated in newspapers (right-hand column) (much like a ventriloquist's dummy). The italic text in Table 2 identifies the duplicated content. Printing of press releases with little modification appears to be common in small-town newspapers. For example, Richard Lugar (R-IN) issued a press release on July 17, 2007, describing why Reynolds, IN was selected to receive federal funding for alternative fuels research. His explanation was repeated, *almost exactly*, on July 18, 2007, in The Times—a local newspaper in heavily Democratic northwest Indiana. The repetition of press releases also occurs in major metropolitan newspapers. On May 30, 2008, Senator Dick Durbin's (D-IL) office issued a press release about funding secured for hybrid buses in Chicago. The Chicago Tribune, which has the fifth largest circulation among American newspapers, used Durbin's release with only slight modification on May 31, 2008 (second row of Table 2). The third example shows that a joint press release from Susan Collins (R-ME) and Olympia Snowe (R-ME) announcing funds secured for laid-off factory workers was reprinted—essentially unchanged—in the Bangor Daily News.

3.3 Measuring the Coverage Rate of Senate Press Releases in Newspapers

Press releases can influence how a newspaper covers a member of Congress without being plagiarized directly by providing a source for statements on a representative's position or drawing attention toward funds secured for a district. To more systematically analyze how often press releases translate into *coverage* in local newspapers, I measured the percentage of press releases from 10 Senate offices (identified in column 2, Table 2) that were quoted, paraphrased, or plagiarized in six local newspapers (identified in column 1). The total number of press releases from a senate office that were covered in a given newspaper is contained in column 3, whereas column 4 identifies the percentage of a senator's total press releases that a newspaper covered.

To determine if a press release from a Senate office was used in a newspaper, I first collected all newspaper stories from 2007 and January 2008 that contained the relevant senator's name.[3] I then used publicly available *cheating detection* software to uncover press releases and newspaper stories that had extremely similar content (Bloomfield 2008). The software provides an efficient method for searching over the 1,069,430 potential newspaper-press release pairs that must be checked to generate Table 3. I then took the pairs of newspaper stories and press releases that the software identified with similar content and manually validated that the newspaper article contained a quote from the press release.

Overall, Table 3 indicates that press releases are regularly used by local newspapers. For example, both Orrin Hatch (R-UT) and Bob Bennett (R-UT) had over a quarter of their press releases covered in the Deseret Morning News. This should not be surprising: the Deseret Morning News has a history of financial problems that constrain its ability to cover politics, and the editor-in-chief, Joseph Cannon, is the former chair of the Utah Republican party. Other newspapers use press releases at a similar rate: Byron Dorgan (D-ND) had 54 of his press releases used in the Bismarck Tribune, and 74 press releases from Susan Collins (R-ME) were used in the Bangor Daily News. Even the San Francisco Chronicle used press releases from both Diane Feinstein (D-CA) and Barbara Boxer (D-CA). The high percentage of press releases used from each Senate office is particularly striking given that Senate offices write press releases for an entire state. Therefore, a sizable portion of press releases from a Senate office is irrelevant to a local paper.

This analysis, although limited to a subset of senators and newspapers, shows that press releases are a common source of information for newspapers. Press releases are regularly quoted in local newspapers—such as the Bismarck Tribune—and are used in major metropolitan papers—like the San Francisco Chronicle. This confirms that press releases are an important medium that legislators use to communicate with constituents, as the messages in press releases are likely to reach constituents.

4 Preparing the Texts for Analysis

Using the thousands of press releases from all Senate offices, the expressed agenda model measures the priorities senators communicate to their constituents through press releases. To perform this analysis, a set of preprocessing steps are performed on the press releases, all of which are well established in the literature on the statistical analysis of text (Manning et al. 2008). The first step discards the order of words in the press release, leaving an unordered set of words remaining (Hopkins and King forthcoming; Quinn et al. forthcoming). Although one might expect the order of words to be crucial to understanding the sentiment expressed in a text, identifying the topic of a press release should be invariant to permutations of word order. Certain topics, such as the Iraq war, should result in specific words appearing with

[3]Stories were collected from the Lexis-Nexis database of newspaper stories. The newspapers selected are not a random sample of papers. I intentionally selected newspapers that I conjectured would display a great deal of variation in their use of press releases. The Deseret Morning News (Salt Lake City, Utah), The Bismarck Tribune (Bismarck, North Dakota), and the Bangor Daily News (Bangor, Maine) are all local newspapers that were likely to be highly reliant on the information senators provided. The Salt Lake Tribune (Salt Lake City, Utah), The Pioneer Press (St Paul, Minnesota), and the San Francisco Chronicle (San Francisco, California) are all large newspapers with greater capacity for covering political news. I also selected a mix of Republican, Democrat, and split delegations; senators from big and small states; and senators who issued a great deal of press releases and senators who issue only a few. Given the nature of the sample selection, inference to a broader population is inappropriate from these data, but is the subject of future research.

Table 2 ''Ventriloquism'': press releases in local newspapers

Richard Lugar (R-IN), 7/17/2007	*The Times (IN), 7/18/07*
The Town of *Reynolds was selected in 2005 to demonstrate* to the nation and the world *that a community's energy needs can be fully met through locally produced renewable sources, including electricity, natural gas replacement, and vehicular fuel* (Lugar 2007).	*Reynolds*, located about 20 miles north of Lafayette, *was chosen in 2005 to demonstrate that a community's energy needs can be fully met through locally produced renewable sources, including electricity, natural gas replacement, and vehicular fuel* (AP 2007).
Dick Durbin (D-IL), 5/30/2008	*Chicago Tribune (IL), 5/31/2008*
U.S. Senator Dick Durbin (D-IL) announced today that the *U.S. Department of Transportation (DOT) has awarded a $9.6 million grant to the city of Chicago* that *will allow the Chicago Transit Authority to purchase approximately 13 additional articulated diesel hybrid buses. Hybrid buses* are quieter, cleaner, *burn less gas*, and run more smoothly *than conventional diesel.* (Durbin 2008).	*U.S. Senator Dick Durbin says the city of Chicago will receive $9.6 million from the federal government* to buy hybrid buses. Durbin said Friday that the grant from the *U.S. Department of Transportation will allow the Chicago Transportation Authority to buy about 13 more articulated diesel hybrid buses.* In March, the CTA announced plans to lease 150 hybrid buses at the cost of $13.4 million a year. *Hybrid buses burn less gas than conventional diesel buses* (AP 2008).
Susan Collins (R-ME), 11/1/2007	*Bangor Daily News (ME), 11/2/2007*
U.S. Senators Olympia J. Snowe and Susan Collins today announced that the U.S. Department of Labor has approved their request for $894,918 in National Emergency Grant funding for Domtar and Fraser Mill workers. Last month Senators Snowe and Collins sent Secretary Chao a letter urging the Department of Labor to quickly review and approve the NEG funding request for the 300 workers who lost their jobs at Domtar Industries in Baileyville and Fraser Papers of Madawaska. "This is great news for 300 workers in Northern and Eastern Maine who lost their jobs through no fault of their own" said Senators Snowe and Collins (Collins 2007).	*U.S. Senators Olympia J. Snowe and Susan Collins Thursday announced that the U.S. Department of Labor has approved their request for $894,918 in National Emergency Grant funding for Domtar and Fraser Mill workers. Last month, the senators sent Secretary* Elaine *Chao a letter urging the Department of Labor to quickly review and approve the NEG funding request for the 300 workers who lost their jobs at Domtar Industries in Baileyville and Fraser Papers of Madawaska. "This is great news for 300 workers in Northern and Eastern Maine who lost their jobs through no fault of their own," said* the senators (Staff 2007).

high frequency (troop, war, iraqi) irrespective of whether the senator supports or opposes the war.

Next, all the words are placed into lower case and all punctuations are removed. Then, I applied the Porter stemming algorithm to each word (Porter 1980). The stemming algorithm takes as an input a word and returns the word's basic building block, or *stem*. For example, the stemming algorithm takes the words family, families and returns famili.

After stemming the words in each document, I counted the number of occurrences of each word in the *corpus*, the total set of press releases. All words that do not occur in at least 0.5% of press releases were removed (Quinn et al. forthcoming). Finally, I removed all *stop* words (e.g., around, whereas, why, whether), along with any word that appears in over 90%

Table 3 Measuring the coverage rate in Senate press releases

Newspaper	*Senator*	*Number quoted*	*Percent of press releases*
Deseret Morning	Bennett (R-UT)	35	32.4
Deseret Morning	Hatch (R-UT)	67	27.2
Salt Lake Tribune	Bennett (R-UT)	21	19.4
Bangor Daily	Collins (R-ME)	74	18.2
Salt Lake Tribune	Hatch (R-UT)	43	17.4
Bismarck Tribune	Dorgan (D-ND)	54	16.8
Bismarck Tribune	Conrad (D-ND)	33	16.3
Pioneer Press	Klobuchar (D-MN)	29	13.1
Pioneer Press	Coleman (R-MN)	32	12.2
Bangor Daily	Snowe (R-ME)	44	11.9
San Francisco Chronicle	Boxer (D-CA)	11	7.2
San Francisco Chronicle	Feinstein (D-CA)	24	6.3

Note. This table presents the coverage rate of press releases in local newspapers and shows that constituents are likely to read the contents of their representative's press releases in local newspapers. The first column contains the name of the newspaper and the second column identifies which senator's press releases were used. The third column presents the number of press releases that had content appear in a story in the local newspaper. To compute this number, I used freely available cheating detection software to uncover sentences that were the same or highly similar (Bloomfield 2008). The fourth column presents the percentage of press releases from a Senate office that was covered in the newspaper.

of any individual senator's press releases. This ensures that each senator's press releases are not grouped together based upon language unique to each senator, yet unrelated to the topic of the document.

After preprocessing the press releases, 1988 unique stems remain, along with 3,715,293 stem observations in the 24,236 press releases. Each document is represented as a $w \times 1$ vector, where w are the number of stems that remain after the preprocessing (in this example, $w = 1988$).

5 A Statistical Model for Expressed Agendas

When measuring the attention political actors allocate toward topics in texts, the data are naturally organized hierarchically, with press releases grouped according to the Senate office that authored the statement. At the top of the hierarchy we suppose that there are a set of senators, indexed by $i = 1, \ldots, n = 100$. Each senator decides how much attention to dedicate to each topic k ($k = 1, \ldots, K$) present in her press releases. The vector describing the attention a senator dedicates to each topic is her *expressed agenda* and probabilistically determines how often each of the K topics appear in her press releases.

At the bottom of the hierarchy are each senator's press releases. Represent press release j ($j = 1, \ldots, D_i$) from senator i with the $w \times 1$ vector y_{ij}. Typical element of y_{ij}, y_{ijz}, measures the number of times the zth stem occurs in the jth document from the ith senator. To connect the senator's priorities with the content of her press releases, suppose that each press release has only one topic. Although a common assumption in statistical topic models, this assumption is particularly appropriate for Senate press releases. Press releases are written in a style similar to short news stories, designed to draw attention to one particular aspect of a senator's activities in Washington. Thus, most press releases address one particular topic. The topic of each press release is a random draw, with the probability of a specific topic occurring determined by the attention senator i dedicates to the issue.

Conditional upon this sampled topic, a press release's content is drawn from a distribution that is specific to each topic. Formally, the expressed agenda model is a hierarchical mixture model where the mixture weights (senators' expressed agendas) are allowed to vary across senators, but the components of the mixture (topics) are fixed across authors to ensure that the priorities of senators are comparable (see Section 6 below). To complete our preliminary notation, suppose that there are a total of $D = \sum_{i=1}^{100} D_i$ press releases and collect all the press releases into the $D \times w$ matrix $\boldsymbol{Y}$.

5.1 Senator-Level Parameters: Senators' Expressed Agendas

The expressed agenda for each senator determines the probability that topics appear in documents. Call the attention senator i allocates to issue k, π_{ik}. Equivalently, π_{ik} represents the expected probability that a press release is generated by the kth topic. Each senator's expressed agenda, $\boldsymbol{\pi}_i$, is then defined as the $K \times 1$ vector describing the attention she dedicates to each topic, $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK})$. In order for $\boldsymbol{\pi}_i$ to be interpreted as the probability of each topic appearing in a press release, its elements must sum to one, $\sum_{k=1}^{K} \pi_{ik} = 1$, and every entry must be greater than zero, $\pi_{ik} \geqslant 0$ for each $k = 1, \ldots, K$. Substantively, this assumption implies that senators are resource constrained when allocating attention to issues and cannot distract from an issue any more than not issuing a press release on the issue.

5.2 Document-Level Parameters: Topics and Words

Conditional on a senator's expressed agenda, $\boldsymbol{\pi}_i$, we draw the topic of each press release. Represent press release y_{ij}'s topic with the $K \times 1$ indicator vector $\boldsymbol{\tau}_{ij}$: if press release y_{ij} was generated by the kth topic, then $\tau_{ijk} = 1$ and the other $K - 1$ elements of $\boldsymbol{\tau}_{ij}$ are equal to 0.[4]

The topic of each press release $\boldsymbol{\tau}_{ij}$ is a draw from a multinomial distribution,

$$\boldsymbol{\tau}_{ij} \big| \boldsymbol{\pi}_i \sim \text{Multinomial}\big(1, \boldsymbol{\pi}_i\big). \tag{5.1}$$

Equation (5.1) connects the topics of press releases to a senator's expressed agenda. The expected proportion of senator i's press releases allocated to the kth topic is π_{ik}.

Conditional on the sampled topic, $\boldsymbol{\tau}_{ij}$, we draw the content (words) of each press release. One possibility would be to model the contents of each press release y_{ij} directly as a draw from a normal distribution (Fraley and Raftery 2002). But using normal distributions to cluster documents will tend to group press releases together based upon the number of words used in the document or the length of y_{ij} (Banerjee et al. 2005). If the length of a document does not contain information about the topic of a document, then using the normal distribution is inappropriate.

To eliminate the influence of word count when clustering press releases, I normalize each press release to have unit length. The unit length representation of y_{ij} is given by $\boldsymbol{y}_{ij}^*$, with $\boldsymbol{y}_{ij}^* = \frac{\boldsymbol{y}_{ij}}{||\boldsymbol{y}_{ij}||}$ where $||\cdot||$ is defined as the Euclidean norm, $\big|\big|\boldsymbol{y}_{ij}\big|\big| = \left(\boldsymbol{y}_{ij}^{'}\boldsymbol{y}_{ij}\right)^{1/2}$. $\boldsymbol{y}_{ij}^*$, now measures the relative rate words, are used in each press release rather than the total number of times each stem is used in a document.

After normalizing each press release, we suppose that $\boldsymbol{y}_{ij}^*$ is a draw from a distribution that is defined only the set of unit-length vectors (or a unit hypersphere): the von Mises-Fisher

[4]Collect the indicator vectors for all of senator i's press releases into the $D_i \times K$ matrix $\boldsymbol{\tau}_i$.

(vMF) distribution (Banerjee et al. 2005). The vMF distribution is characterized by a $w \times 1$ vector that governs the distribution's center, $\boldsymbol{\mu}$, and a scalar that determines the distribution's dispersion, κ. $\boldsymbol{\mu}$ points to the location on the unit hypersphere, where the vMF distribution reaches its mode. κ is an inverse dispersion parameter: as $\kappa \rightarrow 0$ the vMF density approaches the uniform distribution on a sphere, as $\kappa \rightarrow \infty$ the vMF converges upon a spike at the center, $\boldsymbol{\mu}$.

Suppose that there are K vMF distributions and represent the center and dispersion parameter for the kth vMF distribution as $\boldsymbol{\mu}_k$, κ_k. $\boldsymbol{\mu}_k$ can be thought of as the prototype document for the kth topic. A press release's topic, $\boldsymbol{\tau}_{ij}$, determines the vMF distribution used to generate the content of a press release. Formally, if $\tau_{ijk} = 1$, then

$$\boldsymbol{y}^*_{ij} \big| \left(\tau_{ijk} = 1\right), \boldsymbol{\mu}_k, \kappa \sim \text{vonMises}-\text{Fisher}_w\left(\boldsymbol{\mu}_k, \kappa_k\right). \tag{5.2}$$

The vMF distribution has sampling density $f\left(\boldsymbol{y}^*_{ij} \big| \boldsymbol{\mu}_k, \kappa_k\right) = c(\kappa_k)_w \exp\left(\kappa_k \boldsymbol{\mu}_k^{'} \boldsymbol{y}^*_{ij}\right)$, $c(\kappa_k)_w$ is a normalizing constant given by $c_w(\kappa_k) = \frac{\kappa_k^{w/2-1}}{(2\pi)^{w/2} I_{w/2-1}\left(\kappa_k\right)}$ and $I_{w/2-1}$ is a modified Bessel function of the first kind.[5] It will be convenient to collect the center of each topic's vMF distribution into the $w \times K$ matrix, $\boldsymbol{\mu}$, and the inverse dispersion parameter for each topic into the $K \times 1$ vector $\boldsymbol{\kappa}$.

5.3 Priors for the Expressed Agenda Model

I place a prior on each senator's expressed agenda, $\boldsymbol{\pi}_i$, to partially pool information across senators to allow for more efficient inferences (Gelman and Hill 2007). Suppose that each $\boldsymbol{\pi}_i$ is a draw from a Dirichlet distribution,

$$\boldsymbol{\pi}_i | \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \tag{5.3}$$

where $\boldsymbol{\alpha}$ is a $K \times 1$ vector of shape parameters which govern the Dirichlet distribution.[6] Rather than assume specific values of $\boldsymbol{\alpha}$, we estimate the parameters to determine the amount of pooling from the data. We suppose that each α_k is a draw from a Gamma distribution and assume parametric values of the Gamma distribution to limit the amount of pooling across senators.[7]

5.3.1 Including covariates in the prior

In Appendix B.6, I modify the prior on senators' priorities to allow for the inclusion of covariates, using a Dirichlet-multinomial regression, a modified version of the prior

[5] Alternatively, the model could be developed using a multinomial distribution for the words in documents, and this option is available in the statistical package.

[6] Any distribution on the simplex will suffice for the general setup of the model. The Dirichlet distribution was selected because it makes inference straightforward and it has limited influence on the results. The Dirichlet distribution assumes that there is a negative covariance between the attention dedicated to each topic (Aitchison 1986). This assumption is dangerous only if certain components of the composition have a large, positive covariance. As an alternative, a logistic normal distribution could be used to pool the expressed agenda across senators. A version of the model with a logistic normal distribution is available in the software package. Tests with the logistic normal prior indicate that the more general model does not yield different estimates of expressed priorities than the model with the Dirichlet prior.

[7] Suppose that the Gamma distribution has sampling density, $g(\alpha_k | \lambda, \delta) = \alpha_k^{\delta-1} \frac{\exp\left(-\frac{\alpha_k}{\lambda}\right)}{\lambda^{\delta}(\delta)}$. We set $\lambda = 1$ and $\delta = 1$, which ensures that the value of each α_k remains relatively small, limiting the pooling.

introduced in Mimno and McCallum (2008). These covariates allow for the inclusion of additional information that allows for smoothing across groups of senators who share similar characteristics—such as a senator's political party or a dummy variable for a specific senator who served in different years. For expository purposes, this paper proceeds with a model that does not include additional covariate information, but this model is available in the software package.[8]

I fix all K κ's to a single value in the results below, κ is set to 100 (Zhong and Ghosh 2003).[9] Conditional on κ we assume a conjugate prior for the center of vMF distributions $\boldsymbol{\mu}_k|\kappa \sim \text{vMF}_w(\boldsymbol{\eta}, \kappa)$. The center of the prior vMF distribution, $\boldsymbol{\eta}$, has typical element $\frac{1}{\sqrt{w}}$.[10]

5.4 Posterior Distribution for the Expressed Agenda Model

Figure 2 provides a graphical display of the complete expressed agenda model: the directed acyclic graph consistent with the model and priors that comprise the expressed agenda model. The arrows in the graph depict the parameters that each random variable's density is dependent upon. For example, the directed edge $\boldsymbol{\alpha} \rightarrow \boldsymbol{\pi}^i$ denotes that the sampling density of senator i's expressed agenda, $\boldsymbol{\pi}^i$, depends upon $\boldsymbol{\alpha}$. Notice the hierarchical structure

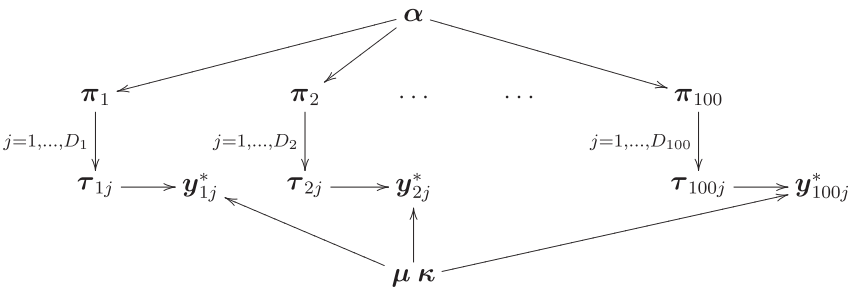

Fig. 2 Bayesian graph of expressed agenda model. This figure presents the expressed agenda model. We assume that each senator's expressed agenda $\boldsymbol{\pi}_i$ is a draw from a Senate-wide Dirichlet distribution Dirichlet($\boldsymbol{\alpha}$). Conditional on $\boldsymbol{\pi}_i$, each press release's topic is a draw from a Multinomial(1, $\boldsymbol{\pi}_i$). Conditional upon this draw, we assume that each document is then drawn from a vMF distribution, with center $\boldsymbol{\mu}_j$ and inverse dispersion parameter κ_j. Note, that all senators select from the same set of topics to ensure that their priorities are comparable. Further, notice the hierarchical structure inherent in the model, with press releases organized according to their author.

[8]The exclusion of covariates in the model does not introduce omitted variable bias as we might expect when using regression to make causal inferences. Rather, the inclusion of covariates improves the information that is borrowed across senators during smoothing. We might expect additional covariates to improve the performance of the model, and therefore, presenting the model with no covariates represents a disadvantage against the expressed agenda model in the evaluations performed below.

[9]Fixing κ across clusters is similar to the approach in Zhong and Ghosh (2003) for estimating mixtures of vMF distributions. The model has been estimated with κ ranging from 50 to 500, and the substantive results remain unchanged.

[10]This is the least informative conjugate prior on $\boldsymbol{\mu}$ that treats all coordinates of the vMF distribution identically because the vMF distribution measures the relative rate at which words occur. To see this, suppose we choose an arbitrarily small quantity $\epsilon > 0$ for each component. The length of $\boldsymbol{\epsilon} = (\epsilon, \ldots, \epsilon)$ is $||\boldsymbol{\epsilon}|| = \sqrt{\omega} \times \epsilon$. Then the normalized vector $\boldsymbol{\epsilon}^* = (\frac{1}{\sqrt{\omega}}, \ldots, \frac{1}{\sqrt{\omega}})$.

present in the model: press releases organized by their author. Appendix B provides the full posterior distribution.

5.5 Inference for the Expressed Agenda Model

Due to the large number of components necessary to capture the variety of topics in press releases, computationally intensive approaches to inference—such as MCMC—are prohibitively slow. Sampling-based methods also face difficulty because permutations of the cluster labels result in the same height of the posterior density, greatly complicating simulation-based inference. One proposed solution is to constrain the parameter space, but this hinders the convergence of the Markov chain (McLachlan and Peel 2000). As an alternative, current best practice recommends running the chains without constraints and then post-processing, identifying the same clusters using a clustering algorithm on the output. This is a useful approach with a small number of mixture components, but MCMC methods have difficulty exploring the posterior as the number of mixture components increases.

Alternatively, one could employ the expectation maximization (EM) algorithm to generate maximum a posteriori (MAP) estimates (McLachlan and Krishnan 1997). This is a reasonable method for inference when MCMC is infeasible, but generating uncertainty estimates from the results of an EM algorithm can be computationally challenging in large mixture models, due to the large number of parameters (McLachlan and Peel 2000). As a result, estimates from an EM model are often analyzed under the assumption that there is no uncertainty in the estimates (Quinn et al. forthcoming). This is an unattractive assumption for two reasons. First, some uncertainty is always present when measuring senator's expressed agendas. Second, the amount of information present about a senator's priorities varies considerably by Senate office. As a result, uncertainty about a senator's expressed agenda should vary by senator as well.

To avoid the difficulties associated with sampling methods and to estimate the entire posterior distribution on each senator's expressed agenda, I use a *variational approximation* to derive an analytical—rather than computational—approximation to the posterior distribution for each senator's expressed agenda (Jordan et al. 1999). Like EM algorithms, variational methods avoid the identification problem because optimization occurs according to a deterministic algorithm based upon starting values and the posterior distribution. Rather than generating MAP parameter estimates, variational methods analytically estimate the entire posterior distribution on each senator's expressed agenda. To perform this estimation, we first restrict the model to a simpler family of distributions. Then, we use the *calculus of variations* to select the member of this distributional family that is closest to the true posterior distribution, where proximity between the distributions is measured using the *Kullback-Leibler* (KL) *divergence* (Bishop 2006). In Appendix B, I derive the update equations used to estimate the posterior distributions.

5.6 Details of Estimation

The results presented in this paper use the variational algorithm derived in Appendix B and assume there are 43 topics present in the data. I varied the number of assumed topic from only five topics, up to 85 different topics. Assuming too few topics resulted in distinct issue being lumped together, whereas too many topics results in several clusters referring to the same issues. During my tests, 43 issues represented a decent middle ground. I corroborated

this number of clusters using a nonparametric model for text clustering, based upon the Dirichlet process prior. This model identified 40–45 clusters in the data set under a wide range of hyperpriors (Blei and Lafferty 2006). The variational algorithm described in Appendix 12 was randomly restarted 100 times, and the analysis was performed on the "best" run.[11]

6 Comparison to Ad-Hoc Approaches to Measuring Author Priorities

Existing methods for unsupervised learning and text clustering are designed to assign documents to topics or to measure the attention in an entire collection of documents—ignoring the information about authors. As a result, these methods are either unable to measure author-specific attention or would require ad-hoc modifications that fail to have the many benefits of the expressed agenda model.

6.1 Clustering Each Senator's Press Releases Separately

To use existing clustering algorithms to measure senators' expressed priorities, one could apply a clustering algorithm to each senator's press releases separately and equate senator attention with the proportion of press releases assigned to each topic.

This method would fail, however, because the estimated topics would be different across senators and the set of topics must be fixed across senators to allow for priorities to be comparable. If a senator issues a press release about a topic (say the Iraq war) only occasionally, an unsupervised learning method will lump together press releases about a topic with other press releases about similar, though distinct, topics (defense spending and veteran affairs, perhaps). The clustering solution for a senator who allocates a great deal of attention to the issue, however, will identify the Iraq war as a distinct topic. As a result, a press release with identical content issued from two different senators could give the impression that the two senators are focusing upon *different* issues.

To demonstrate this problem, consider the press releases of two senators with similar explanatory styles: Robert Menendez (D-NJ) and Frank Lautenberg (D-NJ). I used a mixture of vMF distributions to separately cluster Lautenberg's and Menendez's press releases (Banerjee et al. 2005). To show that two press releases with identical content can be allocated to different topics, I used a *joint press* release—a press release from two senators with identical text—issued by Lautenberg and Menendez on July 31, 2007, that touted the senators' efforts to improve reporting standards about toxic waste disposal (Lautenberg 2007; Menendez 2007). The clustering result from Lautenberg's press releases placed the joint press release in a bureaucratic regulation cluster, with identifying stems push, require, law, bureau.[12] The clustering solution from Menendez assigned the *same document* to a cluster about economic growth (with stems economi, future, econom, studi, growth), because Menendez dedicates considerably less attention to bureaucratic regulation than Lautenberg. This shows that *the same press release can create the appearance that two senators are focusing on different issues* if applied to each senator's press releases separately. The expressed agenda model avoids this problem by fixing the topics across senators.

[11] In Appendix B, I show that optimization occurs by increasing a lower bound on the marginal log-posterior of the data. We analyze the run that had the highest lower bound.

[12] The stems were identified using the mutual information between words and documents. See Section 7.2.

6.2 The Inadequacy of Ad-Hoc Modifications of Existing Methods

Ad-hoc modifications of existing clustering models could provide estimates of authors' priorities. For example, one could run an off-the-shelf clustering method on the entire collection of press releases from each Senate office, then tally the proportion of a senator's press releases that fall into each of the topics. This would create a measure of author-specific attention where the topics are fixed across senators.

This ad-hoc approach, however, is inadequate for several reasons. Most importantly, ad-hoc modifications are unable to provide uncertainty estimates about author-specific attention and subsequently, uncertainty about auxiliary quantities of interest. In contrast, the expressed agenda model estimates the posterior distribution on each author's priorities and is easily extended to posterior distributions on other quantities of interests derived from priorities.

An ad-hoc modification of existing methods also fails to exploit the additional information available to the analyst: the author of each press release. The expressed agenda model uses this additional information to aide in the discovery of topics, assign documents to topics, and measure the priorities authors express relative to the topics. A generative statistical model makes clear the assumptions of the statistical model and how the model could be extended to include across senator and over-time dependence. The statistical model also facilitates the borrowing of information across senators, allowing for efficient inference (Gelman and Hill 2007).

7 Labeling and Validating Topics

An advantage of the expressed agenda model is that the analyst does not need to prespecify the topics in the data. Rather the topics are estimated from the texts. In order to sensibly interpret the expressed agenda of each author, we must reliably label each of the topics and also validate that we are estimating reasonable topics from the data. I use three approaches to perform this evaluation: reading a subset of randomly chosen documents to provide a label, automatically generating distinctive stems to label clusters using the *mutual information* between stems and a topic, and exploiting over time variation in salience to check the reasonableness of cluster labels.

7.1 Labeling Clusters through Manual Document Checking

As a first step to assess the validity of the topics and to generate labels for topics, I randomly selected 10 documents from each topic with a high posterior probability of belonging to that topic (Quinn et al. forthcoming). I then read each of the 10 documents to generate the label found in the first column of Table 4.

On the whole, the clusters seemed to group together documents that referred to the same topic. For example, one group of texts discussed judicial nominations. Press releases in this category include releases from senate delegations to ''Announce Recommendations for Eastern District Federal Judgeships'' (Webb 2007), from members of the Judiciary committee who publicize that the ''Senate Approves Kyl-Feinstein Provision Adding Judgeship to Ninth Circuit'' (Kyl 2007), or declare that ''The United States Senate unanimously confirmed Norman Randy Smith today to serve on the Ninth Circuit Court of Appeals'' (Craig 2007). Another randomly chosen set of press releases dealt with energy policy. Among the press releases selected from this category is an announcement from a group of senators who ''introduced legislation that will increase American drivers' access to ethanol at fuel

Table 4 The topics estimated by the expressed agenda model

Description	*Stems*	*Identifier*
FEMA	disast,fema,storm,damag,declar,emerg,flood,recoveri,rebuild,recov	disast
Food safety	food,fda,agricultur,contamin,recal,inspect,product,nutrit,drug,consum	food
Worker rights	worker,employe,wage,employ,labor,workplac,job,minimum,fair,workforc	worker
AG/Justice	gener,justic,gonzal,judiciari,confirm,resign,investig,million,polit,nomine	gener
Agriculture	farmer,agricultur,crop,produc,rancher,usda,livestock,nutrit,conserv,food	farmer
SCHIP	children,insur,uninsur,schip,kid,enrol,chip,reauthor,parent,incom	children
Public land	land,forest,manag,fish,wildlif,public,recreat,area,natur,speci	land
Pres. Veto/SOTU	presid,bush,veto,depart,iraq,announc,speech,union,democrat,facil	presid
Loan crisis	mortgag,loan,lender,borrow,homeown,lend,bank,crisi,rate,market	mortgag
Border security	border,homeland,immigr,patrol,secur,cross,agent,mexico,illeg,dh	border
Illegal immgr.	immigr,border,illeg,reform,legal,debat,enforc,broken,alien,citizenship	immigr
Honorary	honor,provid,rememb,friend,program,celebr,depart,prayer,tribut,legaci	honor
Global warming	climat,warm,emiss,greenhous,global,carbon,chang,pollut,reduct,environment	climat
Science	scienc,math,competit,compet,technolog,innov,engin,research,global,edg	scienc
Higher edu.	colleg,higher,graduat,loan,univers,maximum,aid,school,grant,afford	colleg
Iraq war	iraq,troop,iraqi,war,withdraw,polit,militari,strategi,petraeu,baghdad	iraq
Veterans' affairs	veteran,affair,medic,mental,wound,war,deserv,traumat,militari,afghanistan	veteran
Tax policy	tax,relief,taxpay,deduct,incom,perman,revenu,credit,code,minimum	tax
Prescrip. drugs	drug,prescript,fda,medicin,medicar,food,patient,market,medic,consum	drug
Energy policy	energi,fuel,oil,renew,sourc,gallon,ethanol,depend,biofuel,effici	energi
Nat/Coast Guard	guard,deploy,mission,militari,duti,soldier,defens,coast,command,iraq	guard
NCLB/School	school,teacher,district,classroom,academ,child,elementari,grade,children,teach	school
Air Force	air,forc,aircraft,base,wing,mission,airlin,militari,plane,defens	air
Women's issues	women,sexual,violenc,woman,assault,victim,domest,awar,prevent,abus	women
Consumer sfty.	consum,product,recal,commiss,manufactur,store,danger,regul,ban,lead	consum
Judicial nom.	judg,nomin,confirm,nomine,circuit,judici,district,judiciari,appeal,legal	judg
Stem cells/research	research,diseas,cure,institut,univers,scientif,scienc,scientist,cell,stem	research
Intl. trade	trade,china,agreement,market,export,manufactur,unfair,worker,product,intern	trade

Continued

Table 4 (continued)

Description	*Stems*	*Identifier*
Gov.Reg/Ethics Ref.	govern,rule,reform,transpar,democraci,account,program,report,elect,foreign	govern
Wounded soldiers	militari,defens,soldier,wound,armi,warrior,arm,veteran,men,walter	militari
Approp: Def. Proj.	million,appropri,defens,project,militari,fiscal,navi,research,air,armi	million
Approp: Water Proj.	water,corp,wrda,engin,project,flood,drink,armi,navig,restor	water
Approp: Econ. Dev.	econom,develop,grant,announc,growth,invest,job,economi,rural,award	econom
Approp: Home State	000,project,500,univers,appropri,hospit,youth,colleg,labor,human	000
Approp: Firefight	firefight,homeland,grant,award,volunt,respond,afg,equip,afgp,depart	firefight
Approp: Airport	airport,aviat,faa,transport,dot,tourist,announc,travel,aircraft,air	airport
Approp: Public Works	project,appropri,fiscal,omnibu,approv,transport,hous,announc,signatur,develop	project
Approp: DHS	secur,homeland,terrorist,dh,threat,attack,terror,11,risk,respond	secur
Approp: School Grants	program,school,youth,particip,grant,children,provid,success,reauthor,teach	program
Approp: Health Care	patient,medicar,hospit,medic,qualiti,medicaid,access,doctor,insur,healthcar	patient
Approp: Crime	crime,enforc,justic,law,crimin,polic,violent,prosecut,gang,local	crime
Approp: HUD	hous,urban,hud,afford,homeless,incom,low,technolog,rehabilit,develop	hous
Approp: Transp.	transport,rail,transit,commut,congest,traffic,corridor,infrastructur,railroad,passeng	transport

pumps'' (Harkin 2007) or Saxby Chambliss (R-GA) stating that he ''addressed members of the Governor's Ethanol Coalition'' (Chambliss 2007), a summary of an investigation into oil companies' attempt to ''prohibit or strongly discourage the sale of alternative fuels'' (Grassley 2007), and legislation introduced to ''dramatically expand renewable fuel sources'' (Bingaman 2007). These press releases all deal with energy—and in particular biofuel as an alternative fuel source.

7.2 An Automatic Cluster Labeling Method

A second approach to applying labels to topics uses the output from the model to identify words that distinguish the documents in a particular topic. The goal is to identify words that are common among documents that discuss the same topic and rare in documents that were generated by another topic. To identify the set of words that satisfy these properties, I select 10 words with the highest mutual information with a topic to label the clusters, which provides a principled method for cluster labeling appropriate for any unsupervised learning technique.

The mutual information between a topic and word measures the amount of information a word provides about whether a topic generated a document randomly chosen from the corpus. Suppose that after estimating the topics using the expressed agenda model, we want to compute the probability that a randomly chosen document $\boldsymbol{y}_{ij} \in \boldsymbol{Y}$ was generated by topic k. Define the event that the document was generated by topic k as $\zeta = I(\tau_k = 1)$, and $\Pr(\zeta = 1)$ is the probability that topic k generated the randomly chosen document. We can summarize our uncertainty about this classification by calculating the *entropy* that k generated a document, $H(k)$ (MacKay 2003),

$$H(k) = -\sum_{t=0}^{1} \Pr(\zeta = t)\log_2 \Pr(\zeta = t), \tag{7.1}$$

where $\log_2$ is used because uncertainty is usually measured in bits. Entropy encodes uncertainty about whether a topic generated a document. It reaches a minimum if all the mass of the probability distribution is centered upon one value (all documents assigned to the same cluster) and reaches a maximum if the probability mass is evenly spread over the possible events (the documents are spread evenly across topics, MacKay 2003).

Conditioning upon additional information, such as a word w, can reduce the uncertainty about whether a topic generated a document. To represent the uncertainty after conditioning upon the additional information, first define the event that a word, w, appears in a document $\boldsymbol{y}_{ij}$ as $\omega = I(w \in \boldsymbol{y})$ and the probability that a word w appears in a randomly chosen document is given by $\Pr(\omega = 1)$. We can now define the entropy for a topic, conditional on word w, $H(k|w)$, as

$$H(k|w) = -\sum_{t=0}^{1}\sum_{s=0}^{1} \Pr(\zeta = t, \omega = s)\log_2 \Pr(\zeta = t|\omega = s). \tag{7.2}$$

As one would expect $\mathrm{H}(k) \geqslant H(k|w)$ for all k and w, with equality only if w provides no information about the clustering, or if the distribution of words in the cluster and outside of the cluster is identical (MacKay 2003).

To generate labels for each topic, we select stems that provide a great deal of information about whether a randomly chosen document belongs to a topic. Intuitively, we want to measure how much a stem reduces the uncertainty in $H(k)$, which we can compute as the

difference between equations (7.1) and (7.2). Define this difference as the mutual information for topic k with stem w, and denote this quantity with $I(k|w) = H(k) - H(k|w)$ (MacKay 2003). If a word w provides no information about whether a topic generated a document, then $H(k) = H(k|w)$ and $I(k|w) = 0$. But, if word w removes all uncertainty about whether a document was generated by topic k, then $H(k|w) = 0$ and $I(k|w)$ obtain its maximum possible value, $H(k)$. Further, as the information a word provides about the probability a document was generated by topic k increases, $I(k|w)$ will increase as well (until reaching its maximum). Thus, the stems with the highest mutual information with each topic provide effective labels for a topic. In Appendix C, I provide the formula used to evaluate the mutual information.

In column 2 of Table 4, I have placed the stems with the 10 largest mutual information with each of the 43 categories. The words identified using the mutual information indicate that the expressed agenda model has uncovered well-defined topics. For example, stems with a high mutual information with the FEMA topic include disast, FEMA, storm, damag, declar, emerg, flood, recoveri, rebuild, recov. The Veteran Affairs topic has a high mutual information with stems veteran, affair, medic, mental, wound, war, deserve, traumat, militari, afghanistan.

In addition to these formal validation methods, a heuristic look at Table 4 suggests the model was able to identify important issues in press releases from senators. The model estimated categories of press releases discussing the Walter Reed scandal and the subsequent Wounded Warrior legislation, the Iraq war, illegal immigration, global warming, the mortgage crisis, and a topic for press releases written to honor constituents and historic events. This suggests that the expressed agenda model was able to recover substantively interesting topics from the data. The final column of Table 4 provides the unique identifier that will be used for each topic throughout the paper.

7.3 Using Senate Debates and External Events to Validate Topics

Following a validation outlined in Quinn et al. (forthcoming), we can use the daily number of press releases generated by each topic as another validity check on the estimated topics. Consider the debate around the Comprehensive Immigration Reform Act of 2007 (S. 1348). President Bush's proposed immigration reforms were met with fierce resistance in the Senate and failed on two separate occasions. Both cloture votes in the Senate were high profile events, garnering a large amount of media and public attention. If the expressed agenda model captures meaningful communications from senators, we should expect to see a spike in the number of press releases about immigration around the cloture votes.

The top plot in Fig. 3 shows the number of press releases placed in the immigration category over 2007.[13] The two days with largest number of press releases about immigration correspond with the two cloture votes in the Senate. The model also detects the debate about the Development, Relief and Education for Alien Minors Act (DREAM) act that would have allowed the children of illegal immigrants to be eligible for college scholarships and enlist in the military. The other two plots in Fig. 3 further illustrate that the model is accurately capturing the content of press releases. Daily press releases about the Attorney General spike during the no-confidence vote for Alberto Gonzales and his subsequent resignation. Honorary press releases—press releases that discuss holidays and honor the

[13]The topic of press release i from senator j was assumed to be the largest element of $\boldsymbol{\tau}_{ij}$.

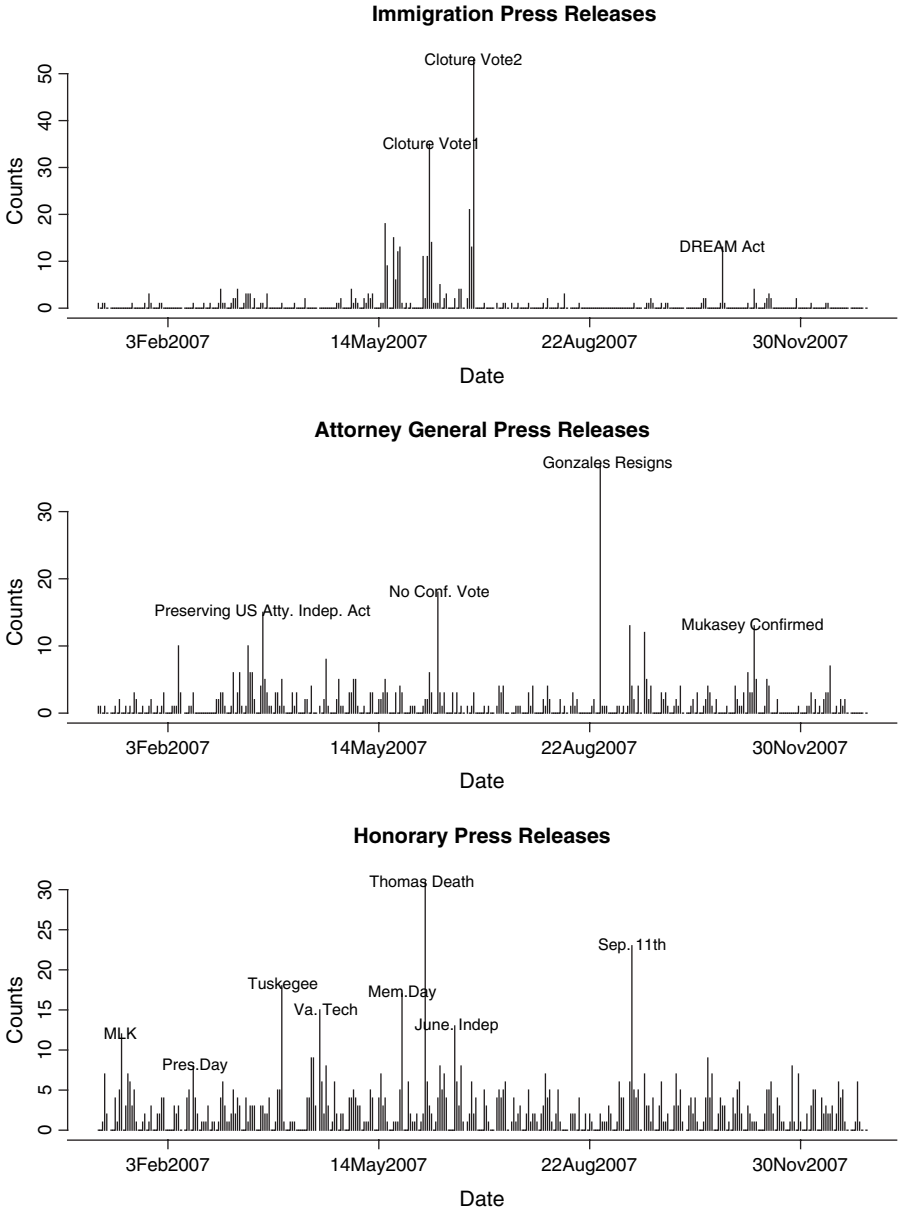

Fig. 3 Senate debates and external events explain spikes in the daily press releases from each topic.

recently deceased—also have spikes corresponding to national holidays, unforseen tragedies, and the death of Senator Craig Thomas (R-WY).

8 Assessing Validity of Estimated Priorities

To validate the estimated expressed agendas from Senate press releases, I use a set of well-established facts about legislative behavior that also have intuitive appeal. If the expressed

agenda model agrees with these patterns first observed in smaller scale qualitative studies, we can have more confidence in applying the results of the model to test more contentious theories of legislative home style. These examples also demonstrate how easily the expressed agenda model can be used to assess how political actors explain work to constituents by incorporating information from every member of a legislature.

8.1 Validation 1: Committee Leaders Focus on their Committee's Issues

Members of Congress have strong incentives to emphasize their positions of power within the legislature. Fenno explains, ''House members explain their use of power in Congress because they believe it will help them win renomination and reelection'' (1978, 139). Elected officials also portray themselves as powerful to be perceived as creating effective policy (Fenno 1978), and legislators are likely to have strong personal interest in the issues that come before committees they lead (Fenno 1973). An implication of Fenno's (1978) argument is that we should observe leaders of Senate committees—chairmen and ranking members—allocate more attention to issues that fall under the jurisdiction of their committee than other senators. This straightforward explanation provides an ideal test of the validity of the estimated expressed agenda model.

Employing the results from the expressed agenda model, Fig. 4 carries out the comparison between prominent committee leaders and the rest of the Senate.[14] In Fig. 4, committee leaders' average attention dedicated to an issue under their committee's jurisdiction is compared with the average attention among the other 98 senators for 47 committee-topic pairs.[15] The left-hand vertical axis denotes the topics that were used for the comparison, and the right-hand vertical axis contains an abbreviated committee or appropriations subcommittee name. The solid dot represents the expected difference between committee leaders and the rest of the Senate, the thick lines are 80% and 95% highest posterior density (HPD) intervals, respectively. If committee leaders discuss issues related to their committee more often, then the estimates should be to the right of the vertical dotted line at zero.

Figure 4 shows that committee leaders allocate more attention to issues under their committee's jurisdiction than the average senator. In all but seven instances committee leaders allocate more attention to the issues under their committee's jurisdiction than other senators, and in some instances, leaders of Senate committees allocate substantially more attention to issues under their jurisdiction than other senators. For example, Joseph Lieberman (ID-CT) and Susan Collins (R-ME), chair and ranking member of the Homeland Security and Governmental Affairs committee, each allocate almost 10 percentage points more attention to Homeland Security issues than other senators, on average. The largest difference between committee leaders and the rest of the Senate corresponds to the Veterans' Affairs committee whose chairman, Daniel Akaka (D-HI), discusses Veterans' issues in 36% of his press releases—20 percentage points more than the closest senator. This example demonstrates that the Expressed Agenda Model is able to retrieve Fenno's (1978) observation that legislators will attempt to highlight their position of power in communications.

[14]In addition to committee leaders on standing committees, I also included subcommittee chairs on the Appropriations committee, due to the prominence of committee membership and the large and diverse nature of the legislation considered by this committee.

[15]Veterans' affairs were calculated with Richard Burr (R-NC) as the ranking member and not Larry Craig (R-ID); the results remain unchanged if Craig is used in place of Burr.

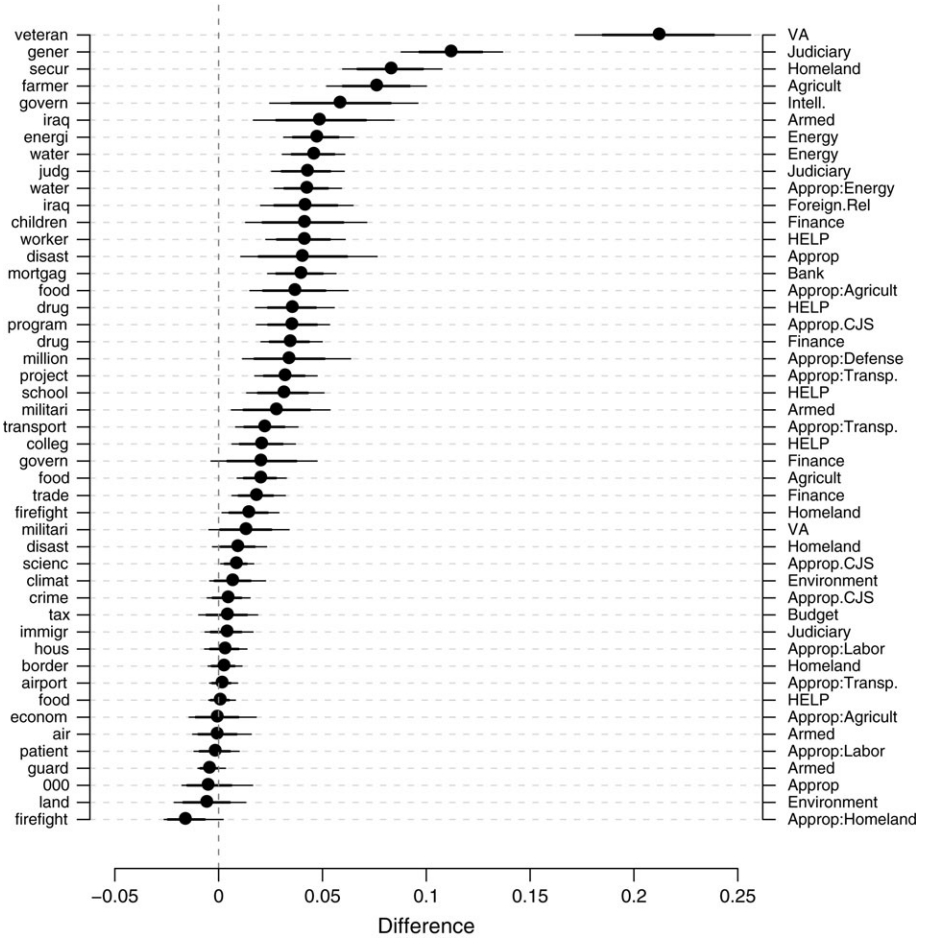

Fig. 4 Chairman and ranking members of committees allocate more attention to issues under their committees' jurisdiction than other senators. This figure compares the attention that Senate committee leaders—chairs or ranking members—dedicate to topics under their committee jurisdictions to the attention allocated by the rest of the Senate. The solid dots represent the expected difference, the thick lines are 80% credible intervals, and the thin lines are 95% intervals. Along the left-hand vertical axis, the topics are listed, and on the right-hand side, the corresponding committee names are listed. In all but seven cases, the dot is to the right of the zero line, indicating that leaders of committees discuss issues that highlight their power in the institution more often than other senators.

8.2 Validation 2: Expressed Agendas Cluster Geographically

Studies of legislative behavior have found that the priorities legislators pursue in Washington and emphasize to constituents vary by location. Arnold argues that this variation occurs because of geographic specific costs and concentrated benefits to many of the policies enacted by Congress (1992, 26). This provides legislators an incentive to the emphasize issues persistently important to their constituents. For example, Fenno describes a senator from a Western state seeking a seat on a committee with jurisdiction over issues important to the ''public-land'' states (1973, 139–40). If the expressed agenda model and

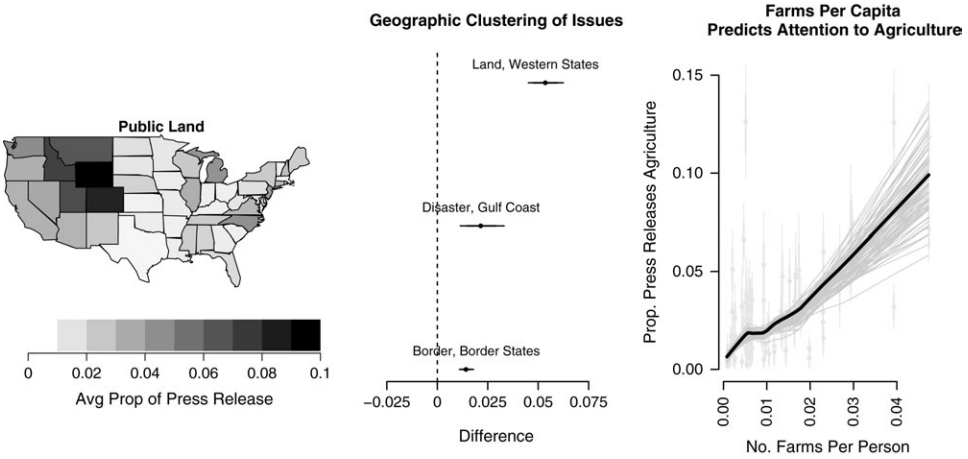

Fig. 5 Attention to issues follows expected geographic patterns. this figure demonstrates that senators' expressed agendas are grouped geographically. The left-hand plot shows that senators from western states allocate substantial attention to public-land issues. Darker shades indicate that the average expected attention from the state's delegation to public-land issues is larger. The center plot carries out a comparison of three different regional issues: public-land and western states (top estimate), hurricanes and gulf coast states (middle estimate), and border-security and states that share an international border (bottom estimate). The point in each plot represents the expected difference between the attention to senators in a geographic area allocate to an issue and the attention senators from other areas of the country dedicate to the same issue. The thick and thin lines are 80% and 95% HPD intervals for this difference. Each point is to the right of the zero, indicating that the issues receive more attention in the geographic areas we would expect. The right-hand plot shows that senators from states with a large number of farms per person also tend to allocate more attention to agriculture issues. The horizontal axis represents the number of farms per resident of the state (one measure of agriculture's importance to a state), and the vertical axis indicates the proportion of press releases allocated to agricultural issues. The gray lines are lowess curves indicates the relationship between the number of farms per capita and the attention to agriculture, whereas the black line is the average relationship.

press release data are recovering valid estimates of legislative behavior, then we should observe this geographic clustering along some issues in the estimated expressed agendas.

The left-hand plot in Fig. 5 shows that this clustering is found in expressed agendas. This plot demonstrates that senators from Western states allocate substantial attention toward public-land topics—indicating a concern with this issue similar to the Western senator in Fenno (1973). The color of each state represents the average expected attention the state's delegation allocated to public land issues. The darker the state, the more attention to the issue and we see that the Western states are nearly black. A manual check shows that western delegations allocate substantial attention to public land. Wyoming's Senate delegation (John Barasso [R-WY] and Mike Enzi [R-WY]) dedicate an average of 18% of their releases to discussions of public land issues and Colorado's delegation (Ken Salazar [D-CO] and Wayne Allard [R-CO]) allocate 14.3% of their releases to land.

The center plot of Fig. 5 carries out the comparison between the attention western and non-western delegations allocate to public-land directly, along with two other geographic comparisons. This plot exhibits the geographic clustering we would intuitively expect from qualitative studies. The top-point represents the expected difference between the attention

to public-land issues for Western senators and the attention to public-land issues among other senators, whereas the thick and thin lines are 80% and 95% HPD intervals for the difference.[16] This shows that there is a very high-posterior probability that senators from Western states allocate more attention to public-land issues than senators from other parts of the country, corroborating an expected geographic comparison. The next two points indicate two other kinds of geographic clustering: senators from the Gulf coast states allocate more attention to disaster (hurricane)-related issues than other senators, and senators from states that share a border with Canada and Mexico issue a larger proportion of press releases about Border security (separate from immigration).

States that do not share borders may bear similar costs or receive similar benefits from policies. As a result, senators from these states with similar interests should attend to similar issues. For example, numerous states have a high-density of farms, but these states are not necessarily grouped in one location. Nonetheless, senators from the high-density agriculture states may be expected to address farm-related issues more than other senators. The right-hand plot shows that this is the case: senators from agricultural states allocate more attention to farming than other senators. The horizontal axis represents the number of farms per resident of the state (one measure of agriculture's importance to a state), and the vertical axis indicates the proportion of press releases allocated to agricultural issues.[17] The light gray lines are lowess curves indicating the relationship between the number of farms per capita and the attention to agriculture. Each gray line represents this relationship for one draw from each senators expressed agenda, whereas the solid black line indicates the average relationship between farms per capita and the proportion of press releases allocated to agriculture.[18] The gray lines slope upwards quickly, demonstrating that senators from states with a high concentration of farms also tend to invest attention in highlighting agricultural issues.

Taken together, the three plots in Fig. 5 demonstrate that the expressed agenda model is able to retrieve geographic and interest-based clustering in expressed agendas: an intuitive property of explanations well established in the qualitative literature on Congressional communication.

8.3 Validation 3: Attention to Appropriations Predicts Opposition to Earmark Reform

Senators who rely upon appropriations secured for their state in press releases have strong incentive to support institutions that allow them to continue to secure particularistic goods (Mayhew 1974). Senators who regularly tout appropriations secured for a state are likely to view these appropriations as essential to their electoral security (Fenno 1978; Cain, Ferejohn, and Fiorina 1987; King 1991). Senators may also feel pressure to ensure that their actions in Washington are consistent with the priorities emphasized to constituents, lest the legislator be portrayed as a hypocrite in future elections (Fenno 1978). In this section, I use a unique vote in the US Senate to show that the results of the expressed agenda model predict aspects of legislative behavior beyond ideal points.

On March 13, 2008, the Senate voted on the Demint-McCain amendment: a proposal introduced by Jim Demint (R-SC) and John McCain (R-AZ) to place a 1-year moratorium on earmarks in senate appropriations bills. Given the incentives to support institutions

[16]Western senators were identified using the region classification from the census bureau.
[17]The numbers of farms per state were obtained from the U.S. Department of Agriculture.
[18]The gray points in the background represent each senator's expected attention to farming, whereas the thick and thin lines are 50% and 90% HPD intervals for this quantity.

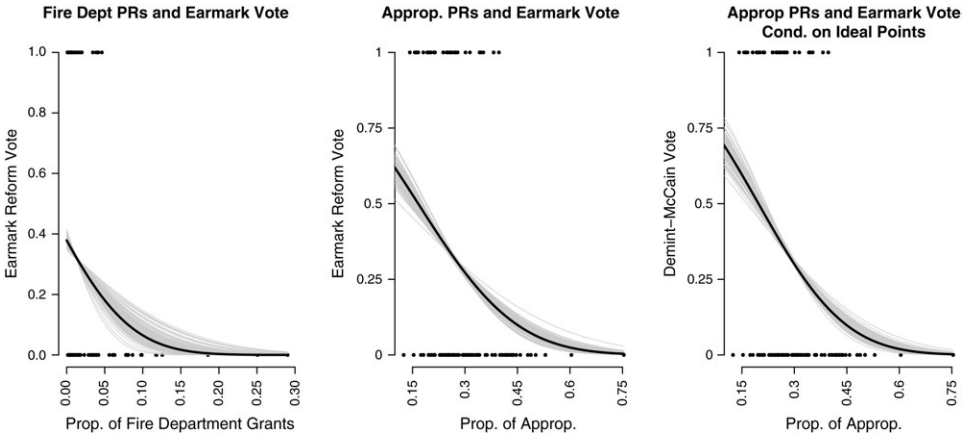

Fig. 6 Senators who dedicate more attention to appropriations were more likely to oppose Demint-McCain. This figure shows that senators who dedicate more attention to appropriations in their press releases are more likely to oppose the Demint-McCain amendment. The vertical axis plots the vote on the amendment, and along the horizontal axis is the average proportion of press releases dedicate to discussing appropriations secured for fire departments. To generate the light gray lines, I took 100 draws from each senator's posterior expressed agenda and then regressed the earmark vote on the draw from the posterior. The gray lines represent the expected probability of supporting the Demint-McCain amendment, and the solid black line is the expected value of the relationship, averaged over the draws from the posterior distribution on the expressed agenda. The left-hand figure shows that senators who discuss fire department grants more often were more likely to oppose the Demint-McCain amendment, and the center plot shows that this relationship was even stronger for an aggregate appropriations category. The right-hand plot shows that the relationship remains even after conditioning upon estimated ideal points of senators, suggesting that consistency explains components of voting behavior beyond ideal point estimates.

essential to maintaining their incumbency advantage and to remain constistent, senators who allocate a large proportion of their press releases toward discussions of appropriations secured for the home state should be more likely to oppose the Demint-McCain amendment.

Figure 6 displays the relationship between senators' vote on the Demint-McCain amendment and two components of the expressed agenda: the proportion of press releases allocated to discussing fire department grants (left-hand plot) and a composite measure of appropriations (center- and right-hand plots). In the left-hand plot in Fig. 6, each senator's vote on the Demint-McCain amendment is predicted using the proportion of press releases dedicated to discussing grants secured for local fire departments—one measure of how often a senator discusses appropriations with constituents.[19] The vertical axis plots the vote on the amendment, and the horizontal axis represents the expected proportion of press releases discussing fire department grants. The gray lines account for the uncertainty inherent in measuring the legislators' priorities by taking 100 draws from each senator's posterior expressed agenda and then regressing the earmark vote on the draws using a probit regression. The black lines represent the average relationship over 1000 draws.

[19]The Demint-McCain amendment was defeated 29-71. I did not include Roger Wicker (R-MS) and Trent Lott (R-MS) due to the change in senate seat after the 2007 session.

Figure 6 shows that senators' votes on the Demint-McCain amendment tended to be consistent with the priorities articulated to constituents. In the left-hand plot, as the proportion of press releases dedicated to fire department grants increases, senators were less likely to support the moratorium on earmarks. The center plot exhibits the relationship between the Demint-McCain vote and an aggregated appropriations category (constituted of the bottom 13 topics from Table 4).[20]

This shows an even stronger relationship: senators who allocate more attention to appropriations were much less likely to vote for the Demint-McCain amendment. The right-hand plot in Fig. 6 shows that the results of the expressed agenda model provides predictive power beyond low-dimensional summaries of previous roll-call votes: the relationship between a senator's vote on Demint-McCain and the proportion of press releases discussing appropriations is still strong and negative, even after conditioning upon a senator's ideal point.[21] Taken together, these three plots show that the results of the expressed agenda model relate as expected to votes on the Senate floor. This provides another validation that the expressed agenda model estimates quantities of theoretical interest.

9 Applying the Expressed Agenda Model

In this section, I show that the estimated expressed agendas are ideal to address theoretically important questions about legislators' home styles. The use of Bayesian inference allows for direct inference about quantities of interest derived from the estimated expressed agendas. Further, by efficiently using all the press releases from each Senate office, the expressed agenda model allows comprehensive tests of hypotheses, in contrast to the limited tests that had been previously carried out in the literature.

9.1 Partners Not Rivals in the Senate

The structure of representation in the Senate is distinctive from other legislative bodies, with each state allocated two senators. Schiller argues that the dual representation in the Senate forces senators representing the same state to articulate distinctive priorities due to the persistent competition for media and public attention (2000, 65). Schiller (2000) provides evidence and a persuasive argument for this novel hypothesis, but is hindered by the existing methods and data, only comparing the statements of a handful of legislators from newspapers. While newspaper stories are an excellent measure of the kind of information available to citizens, newspaper stories conflate senators' priorities with the depictions offered by news writers. The expressed agenda model and the Senate press releases allow a direct and comprehensive test of whether senators from the same state articulate a distinctive set of priorities in press releases.

Schiller's (2000) argument asserts that senators from the same state respond to each other's priorities by advocating a different set of issues, which implies that we should observe senators from the same state articulating a distinctive set of issues. But, Schiller's

[20] Two methods were used to label these topics. First, I used topics with appropriations-related labels and increased attention around the passage of an appropriation bill. Second, I used a topic hierarchy to identify groups of issues that clearly referred to appropriations.

[21] Ideal points were estimated using a one-dimensional item-response theory model, as implemented in MCMCpack (Martin and Quinn 2008). The regressions in the right-hand plot incorporate uncertainty from both the priorities and the ideal points. In each plot, the senators' votes were regressed on draws from the posterior distribution on the priorities and the ideal points for each senator. The simulated lines were generated by varying the attention allocated to appropriations.

(2000) argument could be incorrect and we could still observe some differences in the stated priorities of same-state senators, due to idiosyncratic differences between the two senators in a delegation: such as distinct personal interests, divergent backgrounds, support among different constituencies located in the same state, and different partisanship. All these factors are unrelated to the strategic considerations outlined in Schiller (2000). Therefore, the critical test is not whether two senators articulate a different set of priorities: all senators will have some differences in their stated priorities. Rather, a test of Schiller's (2000) hypothesis depends upon whether senators from the same state have priorities that are *more distinctive* than a comparison group of senators who have no incentive to intentionally articulate different priorities. To perform this test, I compare the differences in priorities among senators who represent the same state to the differences in priorities among senators who represent different states. Under Schiller's (2000) hypothesis, senators who represent different states do not have incentive to carve out distinctive expressed agendas, and therefore, senators who represent different states provide a reasonable group to compare the differences that should be expected due to idiosyncratic variation.

To measure the distance between two expressed agendas, I use the distance metric on the simplex defined in Billheimer et al. (2001), which generalizes intuition about properties of distance in Euclidean space to the simplex. In Appendix 14 I define this metric. Define the distance between two expressed agendas $\boldsymbol{\pi}^j, \boldsymbol{\pi}^i$, $\text{Distance}_{i,j} = g\left(\boldsymbol{\pi}^j, \boldsymbol{\pi}^i\right)$ where $g(\cdot, \cdot)$ is the distance metric developed in Billheimer et al. (2001). To test Schiller's (2000) hypothesis, I compare the average distance between expressed agendas of senators who represent different states to the average distance between expressed agendas of senators who represent the same state.[22] If Schiller's (2000) hypothesis is correct, average distance between expressed agendas of senators who represent different states to the average distance between expressed agendas of senators who represent the same state should be negative: implying that senators from the same state tend to have expressed agendas further apart than senators from different states.

The left-most line in Fig. 7 presents this quantity estimated from Senate press releases. This shows that senators who represent the same state have expressed agendas that are *more similar* than senators who represent different states. The solid dot in Fig. 7 represents the expected value of average distance between expressed agendas of senators who represent different states to the average distance between expressed agendas of senators who represent the same state and is above the horizontal dotted line, indicating that senators from the same state have more similar priorities, on average, than senators who represent different states. The thick lines and thin lines are 50% and 90% HPD intervals for the difference and both fail to intersect the zero line, indicating that there is a high posterior probability that senators from the same state tend to emphasize similar issues. This holds regardless of the partisanship of the state delegation: the next three lines show that the expressed agendas of split, Republican, and Democratic delegations

[22]To derive the comparison between the expressed agendas of senators from the same state and different states, collect the $\binom{100}{2}1002 = 4950$ pairs of senators into the set $\mathcal{P}$. For example, one pair of senators in this set is $\text{Grassley (R}-\text{IA)}, \text{Murray (D}-\text{WA)}$. Define the set $\mathcal{S}$ as the set of 50 pairs of senators who represent the same state, such as (Bayh (D-IN), Lugar (R-IN))$\in \mathcal{S}$. And define $\mathcal{S}' = \mathcal{P} \backslash \mathcal{S}$ as the 4900 pairs of senators who represent different states, for example, (McCain (R-AZ), Obama (D-IL))$\in \mathcal{S}$. Formally,

$$\text{Diff}(\mathcal{S}', \mathcal{S}) = \sum_{(i,j)\in\mathcal{S}'} \frac{g(\pi^i, \pi^j)}{4900} - \sum_{(k,m)\in\mathcal{S}'} \frac{g\left(\pi^m, \pi^k\right)}{50} \tag{9.1}$$

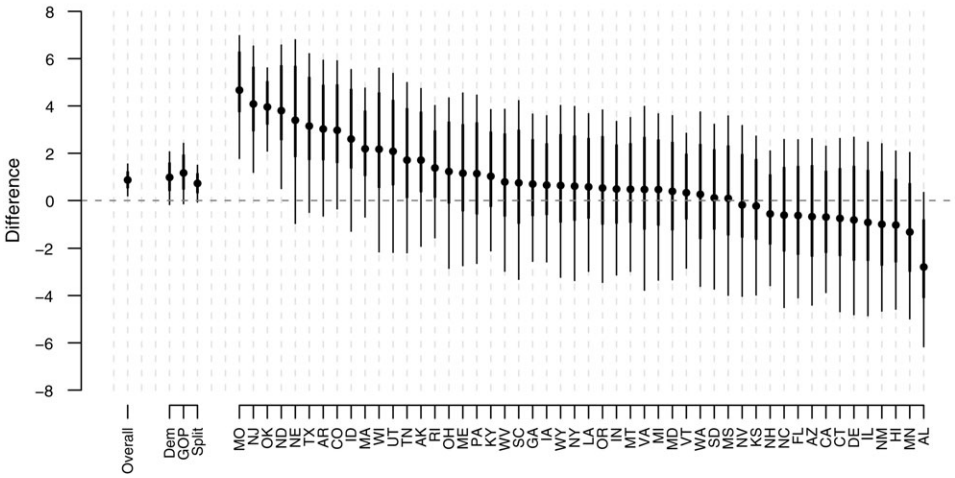

Fig. 7 Senators who represent the same state have more similar expressed agendas than senators from other states. This figure compares the average distance among senators who represent different states to the average distance of senators who represent the same states. The solid dots represent the expected difference, and the thick and thin lines are 50% and 90% HPD intervals, respectively. If senate delegations have more similar expressed agendas than senators who represent different states, then the estimates should be above the horizontal dashed zero-line. The first line compares the average distance between expressed agendas from senators from different states with the average distance between expressed agendas of senators from the same state, showing that senators from the same state communicate a more similar set of priorities than senators from other states. This same pattern holds regardless if the delegation is split, Republican, or Democrat. Further, most states' delegations have more similar expressed agendas than senators who represent different states.

are closer, on average, than the average distance between priorities for senators who represent different states. The final set of lines compare the distance in each senate delegation with the average distance between the expressed agendas of senators who represent different states, and the lines are color-coded according to the partisanship of the delegation. This shows that the majority of state's delegations tend to be more similar than the average distance between the priorities of senators who represent different states, although there is substantial variation in this quantity across states.

This shows that contrary to the prediction's from Schiller's (2000) theory, senators from the same state emphasize a *more similar set of priorities than senators who represent different states*, in press releases from 2007. This similarity could occur because senators from the same state may rely upon similar groups as part of their ''reelection'' constituency (Fenno 1978), subsequently leading senators to identify a similar set of priorities to please this constituency. Alternatively, senators might be able to multiply the effectiveness of their own communication by coupling their efforts with the other senator from their state—forming a partnership to help ensure reelection for both senators.

10 Conclusions and Future Work

This paper has introduced a new method for analyzing the expressed priorities of political actors, as articulated in political texts: the expressed agenda model. This method is capable

of handling thousands of texts from hundreds of political actors to estimate the topics in a data set, assign documents to topics, and measure the proportion of press releases each political actor dedicates to the topics. Using a Bayesian model and a recently developed estimation procedure allows for efficient inference about each senator's priorities. I apply this method to an original collection of press releases from Senate offices and show that the expressed agenda model is capable of retrieving a theoretically relevant set of topics and that press releases are an ideal medium for measuring how senators portray themselves to constituents. Through a series of applications I validate the estimated priorities and topics and show that the model facilitate tests of theoretically important questions about congressional communication.

The statistical model developed in this paper is applicable to a variety of political situations beyond the study of home style and therefore has broad implications for the way political scientists study political communication. The expressed agenda model is ideal whenever scholars are interested in comparing the priorities that authors articulate in text, an important problem in large literatures studying campaign strategy (Petrocik 1996), media-content (Armstrong et al. 2006), and presidential communication (Lee 2008). The expressed agenda model can also be applied to study other forms of Congressional communication, like the attention allocated to issues in Senate floor speeches (Quinn et al. forthcoming) or the issues raised during Senate committee hearings. The forthcoming software package (implemented in the R computing language) makes applying the expressed agenda model straightforward.

The press release data used to analyze senator's expressed agendas provide a comprehensive collection of statements senators make to constituents, which facilitates testing a number of theories. For example, the press releases, coupled with stories from local newspapers, suggest a new approach to studying the connection between politicians and the media. Previous studies of this interaction have relied upon time-series regressions to measure how the priorities politicians articulate covaries with the issues discussed in the media (Bartels 1996). This method provides only suggestive evidence of how politicians and the press interact. In contrast, the press release coverage rate provides a direct measure of how politicians ensure that their message is repeated (and amplified) by the press. Expanding upon this analysis is an important topic for future research. Measuring how often newspapers cover elite statements would provide an answer to a number of theoretically important questions, including how reliant local newspapers are on information from Senate offices, identifying the role of partisanship in determining how often a newspaper prints a legislator's message, and determining how a newspaper's reliance upon information from Congressional offices influences the incumbency advantage.

Appendix A. Collecting the Press Releases

The data set used in this paper contains all the Senate press releases from the 2007 calendar year or the first session of the 110th Congress. Due to the large number of press releases, they require a large expenditure of resources to analyze manually (Yiannakis 1982; Lipinski 2004). I overcome this problem by using automatic data collection methods: I wrote a set of ''screen scraping'' scripts in the Python computing language. Each script collects all the press releases from a senator's Web site, removes any extraneous content unrelated to the text of the press release, and then stores the text. The result of this automated collection process is a data set of 24,236 press releases for 2007.

Appendix B. Deriving the Update Equations for Variational Inference

Given the model and priors outlined above the posterior is given by

$$\begin{aligned}
&\alpha_k \big| \delta, \lambda \sim \text{Gamma}\big(\lambda, \delta\big) \quad \text{for all } k = 1, \ldots, K \\
&\boldsymbol{\pi}_i \big| \boldsymbol{\alpha} \sim \text{Dirichlet}\big(\boldsymbol{\alpha}\big) \quad \text{for all } i = 1, \ldots, n \\
&\boldsymbol{\tau}_{ij} \big| \boldsymbol{\pi}_i \sim \text{Multinom}\big(\boldsymbol{\pi}_i\big) \quad \text{for all } j = 1, \ldots, D_i; i = 1, \ldots, n \\
&\boldsymbol{\mu}_k \big| \boldsymbol{\eta}_k, \kappa \sim \text{vMF}_w\big(\boldsymbol{\eta}, \kappa\big) \quad \text{for all } k = 1, \ldots, \\
&\boldsymbol{y}^*_{ij} \big| \boldsymbol{\mu}, \kappa, \tau_{d_i,j} = 1 \sim \text{vMF}_w\big(\boldsymbol{\mu}_j, \kappa\big) \quad \text{for all } j = 1, \ldots, D_i; i = 1, \ldots, n
\end{aligned}$$

with parametric form

$$\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\tau} | \boldsymbol{Y}) \propto & \prod_{k=1}^{K} \exp(-\alpha_k) \exp(\kappa \boldsymbol{\eta}' \boldsymbol{\mu}_k) \times \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \left(\alpha_k\right)} \\
& \times \prod_{i=1}^{100} \left[\prod_{k=1}^{K} (\pi_{ik})^{\alpha_k - 1} \prod_{j=1}^{D_i} \prod_{k=1}^{K} \left[\pi_{ik} \exp(\kappa \boldsymbol{\mu}' \boldsymbol{y}^*_{ij}) \right]^{\tau_{ijk} = 1} \right]
\end{aligned} \tag{B.1}$$

In the supplemental notes, I provide the model and derive the estimation algorithm for the expressed agenda model with a multinomial distribution used to model document content.

B.1. Approximating Distribution

I adopt a standard *mean-field* approach to estimation of equation (B.1).[23] Specifically, I approximate the full posterior with a family of distributions that contain additional independence assumptions *but no specific parametric forms are assumed* and then select the member of this distributional family that minimizes the Kullback-Leibler divergence between the true posterior and the approximating distribution. Call the approximating distribution $q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})$ and assume that this distribution factors into $q(\boldsymbol{\pi})q(\boldsymbol{\tau})q(\boldsymbol{\mu})q(\boldsymbol{\alpha})$. We will estimate the full posterior for topics $q(\boldsymbol{\tau})$ and senators' priorities $q(\boldsymbol{\pi})$ and then obtain *Maximum a Posteriori* estimates for $\boldsymbol{\mu}_k$ and $\boldsymbol{\alpha}$.[24] This implies that we can write the approximating distribution as $q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha}) = \prod_{i=1}^{N} q(\boldsymbol{\pi})_i \prod_{i=1}^{N} \prod_{j=1}^{D_i} q(\boldsymbol{\tau}_{ij}) \prod_{k=1}^{K} \delta_{\boldsymbol{\mu}^*_k} \delta_{\boldsymbol{\alpha}^*}$ where $\delta_{(\cdot)}$ is the Dirac delta function, $\boldsymbol{\mu}^*_k$ represents the MAP estimates for the kth category and $\boldsymbol{\alpha}^*$ represents the MAP estimates for $\boldsymbol{\alpha}$.[25]

B.2. Minimizing KL Divergence

The standard approach to minimizing the KL divergence between the true posterior and the approximating distribution in variational approximations is to solve an equivalent problem:

[23] The derivation throughout this appendix is a fairly standard in the application of variational inference to mixture models and therefore should have similarities to the derivations in Bishop (2006) and Blei et al. (2003). Note, that I provide these derivations because variational inference in political science is nonstandard.

[24] I estimate the full posterior (with distributions) for the model with multinomial distributions—the integral with vMF distributions are difficult to compute. Furthermore, not much is gained by maintaining a full posterior on the components of the mixture because of the large number of stems used in the analysis.

[25] Recall that the Dirac delta function is a probability distribution that places all of the mass on a single number, given by the term in the subscript.

maximizing a lower bound on the *evidence* or the marginal probability of the data. To derive the lower bound, first write the log evidence as,

$$\log p(\boldsymbol{Y}) = \log \sum_{\tau} \int \int \int p(\boldsymbol{Y}, \pi, \alpha, \mu)\, \mathrm{d}\pi \mathrm{d}\mu \mathrm{d}\alpha.$$

Insert the approximating distribution $q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})$ by multiplying by 1,

$$\log p(\boldsymbol{Y}) = \log \sum_{\tau} \int \int \int \frac{q(\pi, \tau, \mu, \alpha)}{q(\pi, \tau, \mu, \alpha)} p(\boldsymbol{Y}, \pi, \alpha, \mu)\, \mathrm{d}\pi \mathrm{d}\mu \mathrm{d}\alpha.$$

Applying Jensen's inequality yields the lower bound

$$\log p(\boldsymbol{Y}) \geqslant \sum_{\tau} \int \int \int q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha}) \log \left\{ \frac{p(\boldsymbol{Y}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\mu})}{q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})} \right\} \mathrm{d}\boldsymbol{\pi} \mathrm{d}\boldsymbol{\mu} \mathrm{d}\boldsymbol{\alpha}. \tag{B.2}$$

We will define the right-hand side of Inequality (B.2) as $\mathcal{L}(q)$. A straightforward proof (in supplemental notes) shows that maximizing $\mathcal{L}(q)$ with respect to q is equivalent to minimizing the KL divergence between the approximating and true posterior.[26] This is the lower bound used to evaluate convergence of the model as well.

B.3. Distributional Forms

To maximize $\mathcal{L}(q)$ with respect to q, we need to obtain the parametric form of the approximating distribution and select the correct member of that family (maximize the parameters for a given distribution). Either from direct derivation or by applying results on the use of mean-field approximations to exponential families (Jordan et al. 1999), we can obtain the functional forms.[27] This derivation shows that $q(\boldsymbol{\pi})_i$ is a Dirichlet distribution and represents the $K \times 1$ vector of shape parameters that characterize this distribution $\boldsymbol{\theta}_i$. The same derivation shows that $q(\boldsymbol{\tau})_{ij}$ is a multinomial distribution and call $\boldsymbol{r}_{ij}$ the $K \times 1$ vector of parameters for jth document from senator i.

B.4. Iterative Algorithm for Estimation

Each iteration of the estimation algorithm proceeds in several steps. Define the values of the parameters from the previous iteration as $\boldsymbol{\mu}^{\text{old}}, \boldsymbol{\alpha}^{\text{old}}, \boldsymbol{\theta}^{\text{old}}, \mathbf{r}^{\text{old}}$. In each step we update the parameters to maximize the lower bound $\mathcal{L}(q)$ with respect to each independent component of the approximating distribution. The following describes each step in more detail.

[26] Note that $\mathcal{L}(q)$ is a functional: an operator that maps from a space of functions to the real line (Bishop 2006). In the case of exponential family models, the lower bound is convex in the approximating distribution—facilitating iterative (EM-like) algorithms for estimation.

[27] This derivation is standard in variational inferences, see Bishop (2006).

B.4.1. Update step for $\boldsymbol{r}_{ij}$

Typical element of senator $\boldsymbol{r}_{ij}$ for senator i, r_{ijg}^{new} is equal to

$$r_{ijg}^{\text{new}} = \frac{\exp\left[\Psi\left(\theta_{ig}^{\text{old}}\right) - \Psi\left(\sum_{k=1}^{K}\theta_{ik}^{\text{old}}\right)\right]\exp\left[\kappa\boldsymbol{\mu}_{g}^{\text{old}}\boldsymbol{y}_{ij}^{*}\right]}{\sum_{k=1}^{K}\left(\exp\left[\Psi(\theta_{ik}^{\text{old}}) - \Psi(\sum_{j=1}^{K}\theta_{ij}^{\text{old}})\right]\exp\left[\kappa\boldsymbol{\mu}_{k}^{\text{old}}\boldsymbol{y}_{ij}^{*}\right]\right)}. \tag{B.3}$$

where $\Psi(\cdot)$ is the digamma function.

B.4.2. Update step for $\boldsymbol{\theta}^{i}$

Typical element θ_{ig} of $\boldsymbol{\theta}_i$ has update step (Blei et al. 2003),

$$\theta_{ig}^{\text{new}} = \alpha_{g}^{\text{old}} + \sum_{j=1}^{D_i} r_{ijg}^{\text{new}}. \tag{B.4}$$

B.4.3. Update step for $\boldsymbol{\mu}_k$

The update step for $\boldsymbol{\mu}_k^{\text{new}}$ is given by Banerjee et al. (2005),

$$\boldsymbol{\mu}_{k}^{\text{new}} = \frac{\boldsymbol{\eta} + \sum_{i=1}^{N}\sum_{j=1}^{D_i} r_{ijk}^{\text{new}}\boldsymbol{y}_{ij}^{*}}{\left\|\boldsymbol{\eta} + \sum_{i=1}^{N}\sum_{j=1}^{D_i} r_{ijk}^{\text{new}}\boldsymbol{y}_{ij}^{*}\right\|}. \tag{B.5}$$

B.4.4. Update step for $\boldsymbol{\alpha}$

Unfortunately, a closed form for the shape parameters $\boldsymbol{\alpha}$ does not exist, so we use a Newton-Raphson algorithm, developed in Blei et al. (2003) to perform the optimization.

B.5. Using the Model

This estimation algorithm is deterministic and therefore easy to implement in a standard package. This version of the expressed agenda model, along with various extensions, is available in the free R software package expAgenda, which is forthcoming.

B.6. Generalizing the Expressed Agenda Model: Including Covariates

Suppose that we observe an $M \times 1$ set of covariates for each author, $\boldsymbol{X}_i$ (including an intercept term as well). The following extends the Dirichlet-multinomial regression suggested in Mimno and McCallum (2008) to allow for the inclusion of covariates to facilitate more efficient smoothing. Specifically, we modify the model to include a regression at the top of the hierarchy,

$$\begin{aligned}
&\boldsymbol{\beta}_k \sim \text{Normal}\left(0, \sigma^2 I\right) \quad \text{for all } k = 1, \ldots, K \\
&\alpha_{ik} = \exp\left(\boldsymbol{X}_i'\boldsymbol{\beta}_k\right) \\
&\boldsymbol{\pi}_i \left| \boldsymbol{\alpha}_i \sim \text{Dirichlet}\left(\boldsymbol{\alpha}_i\right) \right. \quad \text{for all } i = 1, \ldots, N. \\
&\boldsymbol{\tau}_{ij} \left| \boldsymbol{\pi}_i \sim \text{Multinom}\left(1, \boldsymbol{\pi}_i\right) \right. \quad \text{for all } j = 1, \ldots, D_i; \;\; i = 1, \ldots, N \\
&\boldsymbol{\mu}_k \left| \boldsymbol{\eta}_k, \kappa \sim \text{vMF}_w\left(\boldsymbol{\eta}, \kappa\right) \right. \quad \text{for all } k = 1, \ldots, K \\
&\boldsymbol{y}_{ij}^{*} \left| \boldsymbol{\mu}, \kappa, \tau_{d_i,j} = 1 \sim \text{vMF}_w\left(\boldsymbol{\mu}_j, \kappa\right) \right. \quad \text{for all } j = 1, \ldots, D_i; \;\; i = 1, \ldots, n
\end{aligned}$$

where σ^2 represents the prior variance on the regression coefficients. The inclusion of covariates allows the model to identify subsets of senators who express similar priorities and therefore include additional information in the model that can aid in classification.

B.6.1. Modifying the variational approximation

In this section, I show how the algorithm in Appendix B can be extended to include the regression at the top of the hierarchy.[28] The first modification is an update step for the regression coefficients for each topic $\boldsymbol{\beta}_k$. Collect the coefficient vectors into the $M \times K$ matrix $\boldsymbol{\beta}$. We focus upon MAP estimates for the coefficients, and a closed form update for the regression coefficients is unavailable. Therefore, we apply a BFGS algorithm to maximize the following,

$$f(\boldsymbol{\beta}) = -\sum_{k=1}^{K}\frac{1}{2\sigma^2}(\boldsymbol{\beta}_k'\boldsymbol{\beta}_k) + \sum_{i=1}^{N}\left[\log\Gamma\left(\sum_{k=1}^{K}\alpha_{ik}\right) - \sum_{k=1}^{K}\Gamma(\alpha_{ik})\right] + \sum_{i=1}^{N}\sum_{k=1}^{K}\left[(\exp(\boldsymbol{X}_{ikt}\boldsymbol{\beta}_k)-1)(\Psi(\gamma_{ik})-\Psi\sum_{k=1}^{K}\gamma_{ik})\right].$$

The only other modification to the update step for $q(\boldsymbol{\pi}_i)$ to include the additional information in the prior α_{ik},

$$\gamma_{ik} = \alpha_{ik} + \sum_{j=1}^{D_i} r_{ijk}.$$

The algorithm otherwise remains unchanged.

Appendix C. Deriving an Expression for Mutual Information

To derive an expression for mutual information, we apply the definitions of $H(k)$ and $H(k|w)$ to obtain

$$H(k) - H(k|w) = \sum_{t=0}^{1}\sum_{s=0}^{1}\Pr(\zeta = t, \omega = s)\log_2\frac{\Pr(\zeta = t, \omega = s)}{\Pr(\zeta = t)\Pr(\omega = s)}. \tag{C.1}$$

To evaluate equation (C.1), we compute the necessary probabilities. Define the number of documents in which word w_j appears as $n_j = \sum_{i=1}^{D}\omega_j^i$ and the number of documents where w_j does not appear as $n_{-j} = D - n_j$. Define the effective number of documents assigned to cluster k and the effective number of documents not in cluster k as $n_k = \sum_{i=1}^{D} r_{i,k}$ and $n_{-k} = D - n_k$. To finish the relevant counts, we need to attend to the four possible joint counts of words and topics,

$$n_{j,k} = \sum_{i=1}^{D} r_{i,k}\omega_j^i;\ n_{j,-k} = \sum_{i=1}^{D}(1-r_{i,k})\omega_j^i;\ n_{-j,k} = \sum_{i=1}^{D} r_{i,k}(1-\omega_j^i);\ n_{-j,-k} = \sum_{i=1}^{D}(1-r_{i,k})(1-\omega_j^i).$$

[28] Mimno and McCallum (2008) suggest stochastic EM to estimate a mixture model with the Dirichlet-multinomial regression prior. To my knowledge, this is the first suggestion of a variational-maximization approach.

The probabilities are then defined as,

$$\Pr(\zeta = 1, \omega_j = 1) = \frac{n_{j,k}}{D};\ \Pr(\zeta = 1, \omega_j = 0) = \frac{n_{j,-k}}{D};\ \Pr(\zeta = 0, \omega_j = 1)$$
$$= \frac{n_{-j,k}}{D} \Pr(\zeta = 0, \omega_j = 0) = \frac{n_{-j,-k}}{D};\ \Pr(\zeta = 1) = \frac{n_k}{D} \Pr(\zeta = 0) = \frac{n_{-k}}{D} \Pr(\omega_j = 1)$$
$$= \frac{n_j}{D} \Pr(\omega_j = 0) = \frac{n_{-j}}{D}.$$

This implies the following formula for $I(k|w_j)$ (Manning et al. 2008),

$$I(k|w_j) = \frac{n_{j,k}}{D} \log_2 \frac{n_{j,k} D}{n_j n_k} + \frac{n_{j,-k}}{D} \log_2 \frac{n_{j,-k} D}{n_j n_{-k}} + \frac{n_{-j,k}}{D} \log_2 \frac{n_{-j,k} D}{n_{-j} n_k} + \frac{n_{-j,-k}}{D} \log_2 \frac{n_{-j,-k} D}{n_{-j} n_{-k}}.$$

Appendix D. Defining Distance on the Simplex

In Section 9, I rely upon the distance metric on a simplex developed in Billheimer et al. (2001). In this appendix, I define this metric. Define the *composition* operator, $\mathcal{C}(\boldsymbol{p}) = \left(\frac{p_1}{\sum_{i=1}^{j} p_i}, \ldots, \frac{p_j}{\sum_{i=1}^{j} p_i} \right)$ and define $\triangle^{k-1}$ as the $k-1$ dimensional simplex. Define the additive logistic map $\phi : \triangle^{k-1} \rightarrow \Re^{k-1}$ $\phi(\boldsymbol{c}) = \left(\log\left(\frac{c_1}{c_k}\right), \ldots, \log\left(\frac{c_{k-1}}{c_k}\right) \right)$, where $\boldsymbol{c} \in \triangle^{k-1}$ (Aitchison 1986). Suppose that $\mathcal{N}_{k-1}^{-1} = \boldsymbol{I}_{k-1} - \frac{1}{k} 1 1'$ and that $\boldsymbol{I}_{k-1}$ is a $k-1 \times k-1$ identity matrix and $\boldsymbol{1}$ is a vector of 1's. For two points in a simplex, $\boldsymbol{\pi}^j, \boldsymbol{\pi}^i \in \triangle^{k-1}$, define $g : \triangle^{k-1} \times \triangle^{k-1} \rightarrow \Re_+$, $g(\boldsymbol{\pi}^j, \boldsymbol{\pi}^i) = \phi\left(\mathcal{C}\left(\frac{\pi_1^j}{\pi_1^i}, \ldots, \frac{\pi_k^j}{\pi_k^i} \right) \right) \mathcal{N}_{k-1}^{-1} \phi\left(\mathcal{C}\left(\frac{\pi_1^j}{\pi_1^i}, \ldots, \frac{\pi_k^j}{\pi_k^i} \right) \right)$.

References

Aitchison, John. 1986. *The statistical analysis of compositional data*. New York: Chapman and Hall.

Armstrong, Elizabeth, Daniel Carpenter, and Marie Hojnacki. 2006. "Whose deaths matter? Mortality, advocacy, and attention to disease in the mass media." *Journal of Health Politics, Policy and Law* 31(4):729–72.

Arnold, R. Douglas. 1992. *The logic of congressional action*. New Haven, CT: Yale University Press.

———. 2004. *Congress, the press, and political accountability.* Princeton, NJ: Princeton Press.

Associated Press. 2007. "'Biotown' receives federal grant." *Times of Nortwest Indiana* (accessed May 15, 2008).

———. 2008. "Chicago to receive 9.6 million for hybrid buses". *Chicago Tribune* (accessed June 10, 2008).

Banerjee, Arindam, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. "Clustering on the unit hypersphere using von Mises-Fisher distributions." *Journal of Machine Learning Research* 6:1345–82.

Bartels, Larry. 1996. "Politicians and the press: Who leads, who follows?" *Presentation at the Annual Meeting of APSA,* San Francisco, CA.

Billheimer, D., Peter Guttorp, and William F. Fagan. 2001. "Statistical interpretation of species composition." *Journal of the American Statistical Association* 96(456):1205–15.

Bingaman, Sen. Jeff. 2007. "Bingaman and Domenici introduce legislation to dramatically expand renewable fuel sources." http://bingaman.senate.gov/ (accessed January 1, 2008).

Bishop, Christopher. 2006. *Pattern recognition and machine learning*. New York: Springer.

Blei, David, and John Lafferty. 2006. "Dynamic topic models." *Proceedings of the 23rd International Conference on Machine Learning,* Pittsburgh, PA, June 25–29, 2006. 113–20.

Blei, David, Andrew Y. Ng, and Michael Jordan. 2003. "Latent Dirichlet allocation." *Journal of Machine Learning and Research* 3:993–1022.

Bloomfield, Louis. 2008. "WCopyFind." Software. http://plagiarism.phys.virginia.edu/Wsoftware.html (accessed June 1, 2008).

Cain, Bruce, John Ferejohn, and Morris Fiorina. 1987. *The personal vote: Constituency service and electoral independence*. Cambridge, MA: Harvard University Press.
Chambliss, Sen. Saxby 2007. ''Chambliss Touts focus on BioFuels in Next Farm Bill.'' (accessed January 1, 2008).
Collins, Sen. Susan 2007. ''Senator Collins announces $894,918 for Domtar, Fraser mill workers.'' http://collins.senate.gov/public/ (accessed January 1, 2008).
Cook, Timothy. 1988. ''Press secretaries and media strategies in the House of Representatives: Deciding whom to pursue.'' *American Journal of Political Science* 32(4):1047–69.
———. 1989. *Making laws and making news: Media strategies in the US House of Representatives*. Washington, DC: Brookings.
Craig, Sen. Larry 2007. ''Senate confirms Randy Smith.'' http://craig.senate.gov/ (accessed January 1, 2008).
Durbin, Sen. Richard 2008. ''Durbin announces a 9.6 million DOT grant for CTA hybrid buses.'' http://durbin.senate.gov/ (accessed June 10, 2008).
Fenno, Richard. 1973. *Congressmen in committees*. Boston: Little Brown and Company.
———. 1978. *Home style: House members in their districts*. Boston: Addison Wesley.
Fraley, Chris, and Adrian Raftery. 2002. ''Model-based clustering, discriminant analysis, and density estimation.'' *Journal of the American Statistical Association* 97(458):611.
Gabel, Mathhew, and Kenneth Scheve. 2007. ''Estimating the effect of elite communications on public opinion.'' *American Journal of Political Science* 51(4):1013–28.
Gelman, Andrew, and Gary King. 1990. ''Estimating incumbency advantage without bias.'' *American Journal of Political Science* 34(4):1142–64.
Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
Grassley, Sen. Chuck. 2007. ''Grassley questions big oil's commitment to lessening US dependence on foreign oil.'' http://grassley.senate.gov/ (accessed January 1, 2008).
Gutmann, Amy, and Dennis Thompson. 1996. *Democracy and disagreement*. Cambridge, MA: Harvard University Press.
Harkin, Sen. Tom 2007. ''Lawmakers make renewable fuels availability, energy efficiency a top priority for Congress.'' http://harkin.senate.gov/ (accessed January 1, 2008).
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The elements of statistical learning*. New York: Springer.
Hill, Kim Quaile, and Patricia Hurley. 2002. ''Symbolic speeches in the US Senate and their representational implications.'' *Journal of Politics* 64(1):219–31.
Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2008. ''Computer-assisted topic classification for mixed-methods social science research.'' *Journal of Information Technology and Politics* 4(4):31–46.
Hopkins, Daniel, and Gary King. Forthcoming. ''Extracting systematic social science meaning from text.'' *American Journal of Political Science.*
Jordan, Michael, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. 1999. ''An Introduction to variational methods for graphical models.'' *Machine Learning* 37:183–233.
King, Gary. 1991. ''Constituency service and the incumbency advantage.'' *British Journal of Politics* 21(1): 119–28.
Kingdon, John. 1989. *Congressmen's voting decisions*. Ann Arbor: University of Michigan.
Kyl, Sen. John. 2007. ''Senate approves Kyl Feinstein provision adding judgeship.'' http://kyl.senate.gov/ (accessed January 1, 2008).
Lautenberg, Sen. Frank 2007. ''Lautenberg Bill to reverse Bush administration's weakening of toxic releases reporting,'' Press Release.
Lee, Frances. 2008. ''Dividers, not uniters: Presidential leadership and Senate partisanship, 1981–2004.'' *Journal of Politics* 70(4):914–28.
Lipinski, Daniel. 2004. *Congressional communication: Content and consequences*. Ann Arbor: University of Michigan Press.
Lugar, Sen. Richard 2007. ''Biotown awarded 1.71 million USDA grant.'' http://lugar.senate.gov/ (accessed January 1, 2008).
MacKay, David. 2003. *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.
Manning, Christopher, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.
Mansbridge, Jane. 2003. ''Rethinking representation.'' *American Political Science Review* 97(4):515–28.
Martin, Andrew, and Kevin Quinn. 2008. ''Markov chain Monte Carlo package (MCMCpack).'' Software, R Package.

Mayhew, David. 1974. *Congress: The electoral connection*. New Haven, CT: Yale University Press.
McCombs, Maxwell. 2004. *Setting the agenda: The mass media and public opinion*. Cambridge: Polity.
McLachlan, Geoffrey, and David Peel. 2000. *Finite mixture models*. New York: John Wiley & Sons.
McLachlan, Geoffrey, and Thriyambakam Krishnan. 1997. *The EM algorithm and extensions*. New York: Wiley.
Menendez, Sen. Robert, 2007. "Lautenberg Bill to reverse Bush administration's weakening of toxic releases reporting," Press Release.
Mimno, David, and Andrew McCallum. 2008. "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression." *Conference on Uncertainty in Artificial Intelligence.* Plenary Presentation, Helsinki, Finland.
Ng, Andrew, Michael Jordan, and Yair Weiss. 2002. "On spectral clustering: Analysis and an algorithm." *Advances in Neural Information Processing Systems 14: Proceedings of the 2002 Conference,* Vancouver, Canada.
Petrocik, John. 1996. "Issue ownership in presidential elections, with a 1980 case study." *American Journal of Political Science* 40(3):825–50.
Porter, Martin. 1980. "An algorithm for suffix stripping." *Program* 14(3):130–7.
Quinn, Kevin, Burt Monroe, Michael Colaresi, Michael Crespin, and Dragomir Radev. Forthcoming. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science.*
Schaffner, Brian. 2006. "Local news coverage and the incumbency advantage in the US house." *Legislative Studies Quarterly* 31(4):491–511.
Schiller, Wendy. 2000. *Partners and rivals: Representation in US Senate delegations*. Princeton, NJ: Princeton University Press.
Sigelman, Lee, and Emmitt Buell. 2004. "Avoidance or engagement? Issue convergence in US presidential campaigns, 1960–2000." *American Journal of Political Science* 48(4):650–61.
Simon, Adam. 2002. *The winning message: Candidate behavior, campaign discourse, and democracy* Cambridge, UK: Cambridge University Press.
Staff, 2007. "Sens. Snowe, Collins announce NEG Funding." *Bangor Daily News,* November 2, 2007 (accessed June 15, 2008).
Sulkin, Tracy. 2005. *Issue politics in congress*. Cambridge: Cambridge University Press.
Teh, Y., M. Jordan, M. Beal, and D. Blei. 2006. "Hierarchical Dirichlet processes." *Journal of the American Statistical Association* 101(476):1566–81.
Vinson, Danielle. 2002. *Through local eyes: Local media coverage of congress*. Creskill, NJ: Hampton.
Watanabe, Satosi. 1969. *Knowing and guessing: A quantitative study of inference and information*. New York: Wiley.
Webb, Sen. Jim 2007. "Senators Warner and Webb announce recommendations for judgeships." http://webb.senate.gov (accessed January 1, 2008).
Wolpert, D. H., and W. G. Macready. 1997. "No free lunch theorems for optimization." *IEEE Transactions on Evolutionary Computation* 1(1):67–82.
Yiannakis, Diana Evans 1982. "House members' communication styles: Newsletter and press releases." *Journal of Politics* 44(4):1049–71.
Zhong, Shi, and Joydeep Ghosh. 2003. "A unified framework for model-based clustering." *Journal of Machine Learning* 4(Nov.):1001–37.