Tea Party in the House: A Hierarchical Ideal Point Topic Model and Its Application to Republican Legislators in the 112$^{\text{th}}$ Congress

Viet-An Nguyen
Computer Science
University of Maryland
College Park, MD
vietan@cs.umd.edu

Jordan Boyd-Graber
Computer Science
University of Colorado
Boulder, CO
Jordan.Boyd.Graber@colorado.edu

Philip Resnik
Linguistics & UMIACS
University of Maryland
College Park, MD
resnik@umd.edu

Kristina Miler
Government and Politics
University of Maryland
College Park, MD
kmiler@umd.edu

Abstract

We introduce the *Hierarchical Ideal Point Topic Model*, which provides a rich picture of policy issues, framing, and voting behavior using a joint model of votes, bill text, and the language that legislators use when debating bills. We use this model to look at the relationship between Tea Party Republicans and "establishment" Republicans in the U.S. House of Representatives during the 112$^{\text{th}}$ Congress.

1 Capturing Political Polarization

Ideal-point models are one of the most widely used tools in contemporary political science research (Poole and Rosenthal, 2007). These models estimate political preferences for legislators, known as their *ideal points*, from binary data such as legislative votes. Popular formulations analyze legislators' votes and place them on a one-dimensional scale, most often interpreted as an ideological spectrum from liberal to conservative.

Moving beyond a single dimension is attractive, however, since people may lean differently based on policy issues; for example, the conservative movement in the U.S. includes fiscal conservatives who are relatively liberal on social issues, and vice versa. In *multi-dimensional* ideal point models, therefore, the ideal point of each legislator is no longer characterized by a single number, but by a multi-dimensional vector. With that move comes a new challenge, though: the additional dimensions are often difficult to interpret. To mitigate this problem, recent research has introduced methods that estimate multi-dimensional ideal points using both voting data and the texts of the bills being voted on, e.g., using topic models and associating each dimension of the ideal point space with a topic. The words most strongly associated with the topic can sometimes provide a readable description of its corresponding dimension.

In this paper, we develop this idea further by introducing HIPTM, the *Hierarchical Ideal Point Topic Model*, to estimate multi-dimensional ideal points for legislators in the U.S. Congress. HIPTM differs from previous models in three ways. First, HIPTM uses not only votes and associated bill text, but also the *language* of the legislators themselves; this allows predictions of ideal points from politicians' writing alone. Second, HIPTM improves the interpretability of ideal-point dimensions by incorporating data from the Congressional Bills Project (Adler and Wilkerson, 2015), in which bills are labeled with major topics from the Policy Agendas Project Topic Codebook.[1] And third, HIPTM discovers a *hierarchy* of topics, allowing us to analyze both agenda issues and issue-specific frames that legislators use on the congressional floor, following Nguyen et al. (2013) in modeling framing as second-level agenda setting (McCombs, 2005).

Using this new model, we focus on Republican legislators during the 112$^{\text{th}}$ U.S. Congress, from January 2011 until January 2013. This is a particularly interesting session of Congress for political scientists, because of the rise of the Tea Party, a decentralized political movement with populist, libertarian, and conservative elements. Although united with "establishment" Republicans against Democrats in the 2010 midterm elections, leading to massive Democratic defeats, the Tea Party was—and still is—wrestling with establishment Republicans for control of the Republican party.

The Tea Party is a new and complex phenomenon for political scientists; as Carmines and D'Amico (2015) observe: "Conventional views of ideology as a single-dimensional, left-right spectrum experience great difficulty in understanding or explaining the Tea Party." Our model identifies legislators who have low (or high) levels of "Tea Partiness" but are (or are not) members of the Tea Party Caucus, and providing insights into the na-

[1] http://www.policyagendas.org/

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1438–1448,
Beijing, China, July 26-31, 2015.

ture of polarization within the Republican party. HIPTM also makes it possible to investigate a number of questions of interest to political scientists. For example, are there Republicans who identify themselves as members of the Tea Party, but whose votes and language betray a lack of enthusiasm for Tea Party issues? How well can we predict from someone's language alone whether they are likely to associate themselves with the Tea Party? Our computational modeling approach to "Tea Partiness", distinct from self-declared Tea Party Caucus membership, may have particular value in understanding Republican party politics going forward because, despite the continued influence of the Tea Party, the official Tea Party Caucus in the House of Representatives is largely inactive and its future uncertain (Fuller, 2015).

2 Polarization across Dimensions

Ideal point models describe probabilistic relationships between observed responses (votes) on a set of items (bills) by a set of responders (legislators) who are characterized by continuous latent traits (Fox, 2010). A popular formulation posits an *ideal point* u_a for each lawmaker a, a *polarity* x_b, and *popularity* y_b for each bill b, all being values in $(-\infty, +\infty)$ (Martin and Quinn, 2002; Bafumi et al., 2005; Gerrish and Blei, 2011). Lawmaker a votes "Yes" on bill b with probability

$$p(v_{a,b} = \text{Yes} \mid u_a, x_b, y_b) = \Phi(u_a x_b + y_b) \quad (1)$$

where $\Phi(\alpha) = \exp(\alpha)/(1 + \exp(\alpha))$ is the logistic (or inverse-logit) function.[2] Intuitively, most lawmakers vote "Yes" on bills with high popularity y_b and "No" on bills with low y_b. When a bill's popularity is lower, the outcome of the vote $v_{a,b}$ depends more on the interaction between the lawmaker's ideal point u_a and the bill's polarity x_b.

Multi-dimensional ideal point models replace scalars u_a and x_b with K-dimensional vectors $\boldsymbol{u}_a$ and $\boldsymbol{x}_b$ (Heckman and Jr., 1997; Jackman, 2001; Clinton et al., 2004). Unfortunately, as Lauderdale and Clark (2014) observe, the binary data used for these models are "insufficiently informative to support analyses beyond one or two dimensions", and the additional dimensions are difficult to interpret. To address this lack of interpretability, recent work has proposed multi-dimensional ideal point models to jointly capture both binary votes and the associated text (Gerrish and Blei, 2012; Gu et al., 2014; Lauderdale and Clark, 2014; Sim et al., 2015).

3 Hierarchical Ideal Point Topic Model

Bringing topic models (Blei and Lafferty, 2009) into ideal-point modeling provides an interpretable, text-based foundation for political scientists to understand why the models make the predictions they do. However, both the *topic*—what is discussed—and the *framing*—how it is discussed—also reveal political preferences. We therefore introduce *frame-specific* ideal points, using a hierarchy of topics to model issues and their issue-specific frames. Although the definition of "frame" is itself a moving target in political science (Entman, 1993), we adopt the theoretically motivated but pragmatic approach of Nguyen et al. (2013): just as agenda-issues map naturally to topics in probabilistic topic models (e.g., Grimmer (2010)), the frames as second-level agenda-setting (McCombs, 2005) map to second-level topics in a hierarchical topic model.

Our model's inputs are votes $\{v_{a,b}\}$, each the response of legislator $a \in [1, A]$ to bill $b \in [1, B]$. Two types of text supplement the votes: floor speeches (documents) $\{\boldsymbol{w}_d\}$ from legislator a_d, and the text $\boldsymbol{w}'_b$ of bill b. While congressional debates are typically about one piece of legislation, we make no assumptions about the mapping between $\boldsymbol{w}_d$ and $\boldsymbol{w}'_b$. In principle this allows $\boldsymbol{w}_d$ to be *any* text by legislator a_d (e.g., not just floor speeches about this bill, but blogs, social media, press releases) and—unlike Gerrish and Blei (2011)—this permits us to make predictions about individuals even without vote data for them. Figure 1 shows the plate notation diagram of HIPTM, which has the following generative process:

1. For each issue $k \in [1, K]$
    (a) Draw k's associated topic $\phi_k \sim \text{Dir}(\beta, \phi^\star_k)$
    (b) Draw issue-specific distribution over frames $\psi_k \sim \text{GEM}(\lambda_0)$
    (c) For each frame $j \in [1, \infty)$ (specific to issue k)
        i. Draw j's associated topic $\phi_{k,j} \sim \text{Dir}(\beta, \phi_k)$
        ii. Draw j's regression weight $\eta_{k,j} \sim \mathcal{N}(0, \gamma)$
2. For each document $d \in [1, D]$ by legislator a_d
    (a) Draw topic (i.e., issue) distribution $\theta_d \sim \text{Dir}(\alpha)$
    (b) For each issue $k \in [1, K]$, draw frame distribution $\psi_{d,k} \sim \text{DP}(\lambda, \psi_k)$
    (c) For each token $n \in [1, N_d]$
        i. Draw an issue $z_{d,n} \sim \text{Mult}(\theta_d)$
        ii. Draw a frame $t_{d,n} \sim \text{Mult}(\psi_{d,z_{d,n}})$
        iii. Draw word $w_{d,n} \sim \text{Mult}(\phi_{z_{d,n},t_{d,n}})$
3. For each legislator $a \in [1, A]$ on each issue $k \in [1, K]$
    (a) Draw issue-specific ideal point $u_{a,k} \sim \mathcal{N}(\sum_{j=1}^{J_k} \hat{\psi}_{a,k,j}\eta_{k,j}, \rho)$ weighting $\eta_{k,j}$ by how much the legislator talks about that frame

[2] A probit function is also often used where $\Phi(\alpha)$ is instead the cumulative distribution function of a Gaussian distribution (Martin and Quinn, 2002).

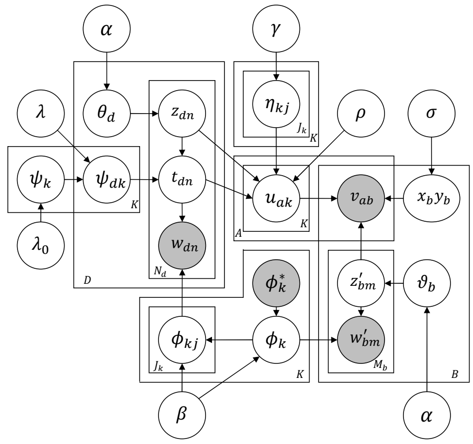

Figure 1: Plate notation diagram of HIPTM.

Agriculture: food; agriculture; loan; farm; crop; dairy; rural; conserve; commodity; eligible; farmer; margin; milk; contract; nutrition; livestock; plant

Health: drug; medicine; coverage; disease; public_health; hospital; social_security; health_insurrance; patient; application; treatment; payment; physician; nurse; clinic

Labor, Employment, and Immigration: employment; immigration; labor; paragraph; eligible; status; compensation; application; wage; homeland_security; unemployment; board; violation; file; perform; mine

Table 1: Examples of informed priors $\phi_k^\star$ for issues.

4. For each bill $b \in [1, B]$
   (a) Draw polarity $x_b \sim \mathcal{N}(0, \sigma)$
   (b) Draw popularity $y_b \sim \mathcal{N}(0, \sigma)$
   (c) Draw topic (i.e., issue) proportions $\vartheta_b \sim \text{Dir}(\alpha)$
   (d) For each token $m \in [1, M_b]$ in the text of bill b
      i. Draw an issue $z'_{b,m} \sim \text{Mult}(\vartheta_b)$
      ii. Draw a word type $w'_{b,m} \sim \text{Mult}(\phi_{z'_{b,m}})$
5. For each vote $v_{a,b}$ of legislator a on bill b
   (a) $p(v_{a,b} \mid \boldsymbol{u}_a, x_b, y_b, \hat{\vartheta}_b) = \Phi\left(x_b \sum_k \hat{\vartheta}_{b,k} u_{a,k} + y_b\right)$

Topic Hierarchy. With the goal of analyzing agendas and frames in mind, our topic hierarchy has two levels: (1) *issue nodes* and (2) *frame nodes*. (Look ahead to Figure 6 for an illustration.) More specifically, there are K issue nodes, each with a topic ϕ_k drawn from a Dirichlet distribution with concentration parameter β and a prior mean vector $\phi_k^\star$, i.e., $\phi_k \sim \text{Dir}(\beta, \phi_k^\star)$. In this hierarchical structure, first-level nodes map to agenda issues, which we treat as non-polarized, and second-level nodes map to issue-specific frames, which we assume polarize on the issue-specific dimension.[3]

To improve topic interpretability, issue nodes have an informed prior from the Congressional Bills Project $\{\phi_k^\star\}$ (Table 1).[4] The frame topic $\phi_{k,j}$ at each frame node is a Dirichlet draw centered at the corresponding (parent) issue node. While the number of issues is fixed *a priori*, the number of second-level frames is unbounded. We also associate each second-level frame node with an ideal point $\eta_{k,j} \sim \mathcal{N}(0, \gamma)$. This resembles how supervised topic models (Blei and McAuliffe, 2007; Nguyen et al., 2015) discover polarized topics' associated response variables.

Generating Text from Legislators. One of our model's goal is to study how legislators *frame* policy agenda issues. To achieve that, we analyze congressional speeches (documents) $\{\boldsymbol{w}_d\}$, each of which is delivered by a legislator a_d. To generate each token $w_{d,n}$ of a speech d, legislator a_d will (1) first choose an issue $z_{d,n} \in [1, K]$ from a document-specific multinomial θ_d, then (2) choose a frame $t_{d,n}$ from the set of infinitely many possible frames of the given issue $z_{d,n}$ using the frame proportion $\psi_{d,k}$ drawn from a Dirichlet process, and finally (3) choose a word type from the chosen frame's topic $\phi_{z_{d,n}, t_{d,n}}$. In other words, our model generates speeches using a mixture of K HDPs (Teh et al., 2006).[5]

Generating Bill Text. The bill text provides information about the policy agenda issues that each bill addresses. We use LDA to model the bill text $\{\boldsymbol{w}'_b\}$. Each bill b is a mixture ϑ_b over K issues, which is drawn from a symmetric Dirichlet prior, i.e., $\vartheta_b \sim \text{Dir}(\alpha)$. Each token $w'_{b,m}$ in bill b is generated by first choosing a topic $z'_{b,m} \sim \text{Mult}(\vartheta_b)$, and then choosing a word type $w'_{b,m} \sim \text{Mult}(\phi_{z'_{b,m}})$, as in LDA.

Generating Roll Call Votes. Following recent work on multi-dimensional ideal points (Lauderdale and Clark, 2014; Sim et al., 2015), we define the probability of legislator a voting "Yes" on bill b as $p(v_{a,b} = \text{Yes} \mid \boldsymbol{u}_a, x_b, y_b, \hat{\vartheta}_b) =$

$$\Phi\left(x_b \sum_{k=1}^{K} \hat{\vartheta}_{b,k} u_{a,k} + y_b\right) \quad (2)$$

where $\hat{\vartheta}_b$ is the empirical distribution of bill b over the K issues and is defined as $\hat{\vartheta}_{b,k} = \frac{M_{b,k}}{M_{b,\cdot}}$. Here, $M_{b,k}$ is the number of times in which tokens in b

[3] Nguyen et al. (2013) allow first-level nodes to polarize but find first-level nodes are typically neutral.

[4] The Congressional Bills Project provides a large collection of labeled congressional bill text. We compute $\{\phi_k^\star\}$ as the empirical word distribution from all bills labeled with k. $K = 19$, corresponding to 19 major topic headings in the Policy Agendas Project Topic Codebook.

[5] If we abandoned the labeled data from the Congressional Bills Project to obtain the prior means $\phi_k^\star$, it would be relatively straightforward to extend to a fully nonparametric model with unbounded K (Ahmed et al., 2013; Paisley et al., 2014).

are assigned to issue k and $M_{b,\cdot}$ is the marginal count, i.e., the number of tokens in bill b.

The ideal point of legislator a specifically on issue k is $u_{a,k}$ and comes from a normal distribution

$$\mathcal{N}(\hat{\psi}_{a,k}^T \boldsymbol{\eta}_k, \rho) \equiv \mathcal{N}\left(\sum_{j=1}^{J_k} \hat{\psi}_{a,k,j} \eta_{k,j}, \rho\right) \quad (3)$$

where J_k is the number of frames for topic k, which is unbounded. The mean of the Normal distribution is a linear combination of the ideal points $\{\eta_{k,j}\}$ of all issue k's frames, weighted by how much time legislator a spends on each frame when talking about issue k, i.e., $\psi_{a,k,j} = \frac{N_{a,k,j}}{N_{a,k,\cdot}}$. Here, $N_{a,k,j}$ is the number of tokens authored by a that are assigned to frame j of issue k, and $N_{a,k,\cdot}$ is the marginal count. When $N_{a,k,\cdot} = 0$, which means that legislator a does not talk about issue k, we back off to an uninformed zero mean.

Equation 3 resembles how supervised topic models (SLDA) link topics with a response, in that the response—the issue-specific ideal point $u_{a,k}$—is latent. It is similar to how Gerrish and Blei (2011) use the bill text to regress on the bill's latent polarity x_b and popularity y_b. In this paper, we only use text from congressional speeches for regression, as these can capture how legislators frame specific topics. Incorporating the bill text into the regression is an interesting direction for future work.

4 Inference

Given observed data of (1) votes $\{v_{a,b}\}$ by A legislators on B bills, (2) speeches $\{\boldsymbol{w}_d\}$ from legislators, and (3) bill text $\{\boldsymbol{w}'_b\}$, we estimate the latent variables using stochastic EM. In each iteration, we perform the following steps: (1) sampling issue assignments $\{z'_{b,m}\}$ for bill text tokens, (2) sampling the issue assignments $\{z_{d,n}\}$ and frame assignments $\{t_{d,n}\}$ for speech tokens, (3) sampling the topics at first-level issue nodes $\{\phi_k\}$, (4) sampling the distribution over frames $\{\psi_k\}$ for all issues, (5) optimizing frames' regression parameters $\{\eta_{k,j}\}$ using L-BFGS (Liu and Nocedal, 1989), and (6) updating legislators' ideal points $\{u_{a,k}\}$ and bills' polarity $\{x_b\}$ and popularity $\{y_b\}$ using gradient ascent.

Sampling Issue Assignments for Bill Tokens The probability of assigning a token $w'_{b,m}$ in the bill text to an issue k is

$$p(z'_{b,m} = k \,|\, \text{rest}) \propto \frac{M_{b,k}^{-b,m} + \alpha}{M_{b,\cdot}^{-b,m} + K\alpha} \cdot \hat{\phi}_{k,w'_{b,m}} \quad (4)$$

where $M_{b,k}$ denotes the number of tokens in bill text b assigned to issue k. The current estimated probability of word type v given issue k is $\hat{\phi}_{k,v}$ (Equation 7). Marginal counts are denoted by $\cdot$ and the superscript $^{-b,m}$ excludes the assignment for token $w'_{b,m}$ from the corresponding count.

Sampling Frame Assignments in Speeches To sample the assignments for tokens in the speeches, we first sample an issue using

$$p(z_{d,n} = k \,|\, \text{rest}) \propto \frac{N_{d,k}^{-d,n} + \alpha}{N_{d,\cdot}^{-d,n} + K\alpha} \cdot \hat{\phi}_{k,w_{d,n}} \quad (5)$$

where $N_{d,k}$ similarly denotes the number of times that tokens in d are assigned to issue k. Given the sampled issue k, we sample the frame as

$$p(t_{d,n} = j \,|\, z_{d,n} = k, a_d = a, \text{rest}) \propto$$
$$\begin{cases} \mathcal{N}(u_{a,k}; \mu_{a,k,j}, \rho) \cdot \left(\frac{N_{d,k,j}^{-d,n}}{N_{d,k,j}^{-d,n} + \lambda} + \frac{\lambda \cdot \hat{\psi}_{k,j}}{N_{d,k,j}^{-d,n} + \lambda}\right), \\ \mathcal{N}(u_{a,k}; \mu_{a,k,j^{\text{new}}}, \rho) \cdot \frac{\lambda}{N_{d,k,j}^{-d,n} + \lambda} \cdot \hat{\psi}_{k,j^{\text{new}}}, \end{cases} \quad (6)$$

where $\mu_{a,k,j} = (\sum_{j'=1}^{J_k} \eta_{k,j'} N_{d,k,j'}^{-d,n} + \eta_{k,j}) / N_{d,k,\cdot}$ for an existing frame j, and for a newly created frame j^{new}, we have $\mu_{a,k,j^{\text{new}}} = (\sum_{j'=1}^{J_k} \eta_{k,j'} N_{d,k,j'}^{-d,n} + \eta_{k,j^{\text{new}}}) / N_{d,k,\cdot}$, where $\eta_{k,j^{\text{new}}}$ is drawn from the Gaussian prior $\mathcal{N}(0, \gamma)$. Here, the estimated global probability of choosing a frame j of issue k is $\hat{\psi}_{k,j}$.

Sampling Issue Topics In the generative process of HIPTM, the topic ϕ_k of issue k (1) generates tokens in the bill text and (2) provides the Dirichlet priors of the issue's frames. Rather than collapsing multinomials and factorizing (Hu and Boyd-Graber, 2012), we follow Ahmed et al. (2013) and sample

$$\hat{\phi}_k \sim \text{Dir}(\boldsymbol{m}_k + \tilde{\boldsymbol{n}}_k + \beta\phi_k^{\star}) \quad (7)$$

where $\boldsymbol{m}_k \equiv (M_{k,1}, M_{k,2}, \cdots, M_{k,V})$ is the token count vector from the bill text assigned to each issue. The vector $\tilde{\boldsymbol{n}}_k \equiv (\tilde{N}_{k,1}, \tilde{N}_{k,2}, \cdots, \tilde{N}_{k,V})$ denotes the token counts propagated from words assigned to topics that are associated with frames of issue k, approximated using minimal or maximal path assumptions (Cowans, 2006; Wallach, 2008).

Sampling Frame Proportions Following the *direct assignment* method described in Teh et al. (2006), we sample the global frame proportion as

$$\hat{\psi}_k \equiv (\hat{\psi}_{k,1}, \hat{\psi}_{k,2}, \cdots, \hat{\psi}_{k,j^{\text{new}}})$$
$$\sim \text{Dir}(\hat{N}_{\cdot,k,1}, \hat{N}_{\cdot,k,2}, \cdots, \hat{N}_{\cdot,k,J_k}, \lambda_0) \quad (8)$$

where $\hat{N}_{\cdot,k,j} = \sum_{d=1}^{D} \hat{N}_{d,k,j}$ and $\hat{N}_{d,k,j}$ can be sampled effectively using the Antoniak distribution (Antoniak, 1974).

Optimizing Frame Regression Parameters We update the regression parameters $\boldsymbol{\eta}_k$ of frames under issue k using L-BFGS (Liu and Nocedal, 1989) to optimize $\mathcal{L}(\boldsymbol{\eta}_k)$

$$-\frac{1}{2\rho}\sum_{a=1}^{A}(u_{a,k} - \boldsymbol{\eta}_k^T \hat{\boldsymbol{\psi}}_{a,k}) - \frac{1}{2\gamma}\sum_{j=1}^{J_k}\eta_{k,j}^2 \quad (9)$$

Updating Ideal Points, Polarity and Popularity We update the multi-dimensional ideal point $\boldsymbol{u}_a$ of each legislator a and the polarity x_b and popularity y_b of each bill b by optimizing the log likelihood using gradient ascent.

5 Data Collection

What makes a Tea Partier? To address that question, we use *key votes* identified by Freedom Works as the most important votes on issues of economic freedom. Led by former House Majority Leader Dick Armey (R-TX), Freedom Works is a conservative non-profit organization which promotes "Lower Taxes, Less Government, More Freedom".[6] Karpowitz et al. (2011) report that Freedom Works endorsements are more effective than other Tea Party organizations at getting out votes for Republican candidates in the 2010 midterms.

For the 112th Congress, Freedom Works selected 60 key votes, 40 in 2011 and 20 in 2012. We are interested in ideal points with respect to the Tea Party movement, i.e., on the anti-pro Tea Party dimension: whether a legislator agrees with Freedom Works on a bill. More specifically, we assign $v_{a,b}$ to be 1 if legislator a agrees with Freedom Works on bill b, and 0 otherwise. In addition to the votes, we obtained the bill text with labels from the Congressional Bills Project[7] and the congressional speeches from GovTrack.[8] In total, we have 240 Republicans, 60 who self-identify with the Tea Party Caucus, and 13,856 votes.

6 Predicting Tea Party Membership

To quantitatively evaluate the effectiveness of HIPTM in capturing "Tea Partiness", we predict Tea Party Caucus membership of legislators given their votes and text. This examines (1) how effective the baseline features extracted from the votes and text are in predicting the Caucus membership, and (2) how much prediction improves using features extracted from HIPTM. For baselines, we consider simple feature sets where each legislator is represented by their speeches as either (1) a TF-IDF vector (Salton, 1968), (2) a normalized TF-IDF vector, or (3) a binary vector containing their votes.

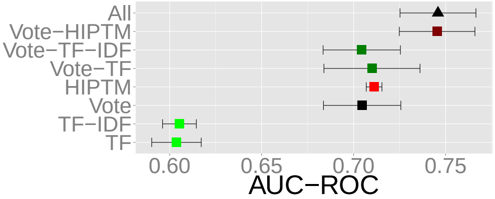

Figure 2: Tea Party Caucus membership prediction results over five folds using AUC-ROC (higher is better, random baseline achieves 0.5). The features extracted from our model are estimated using both the votes and the text.

Our dataset for binary prediction comprises a set of 60 Republican representatives who self-identify as Tea Party Caucus members and 180 who do not. These are divided using 5-fold cross-validation with stratified sampling, which preserves the ratio of the two classes in both the training and test sets. We report performance using AUC-ROC (Lusted, 1971) using SVM^{light} (Joachims, 1999).[9] After preprocessing, our vocabulary contains 5,349 unique word types.

Membership from Votes and Text. First, given the votes and text of all the legislators, we run HIPTM for 1,000 iterations with a burn-in period of 500 iterations. After burning in, we keep the sampled state of the model after every fifty iterations. The feature values are obtained by averaging over the ten stored models as suggested in Nguyen et al. (2014). Each legislator a is represented by a vector concatenating:

- K-dimensional ideal point vector estimated from both votes and text $u_{a,k}$
- K-dimensional vector, estimating the ideal point using only text $\boldsymbol{\eta}_k^T \hat{\boldsymbol{\psi}}_{a,k}$
- B probabilities estimating a's votes on B bills $\Phi(x_b \sum_{k=1}^{K} \hat{\vartheta}_{b,k} u_{a,k} + y_b)$

Figure 2 shows AUC-ROC results for our feature sets. VOTE-based features clearly outperform text-based features like TF and TF-IDF. Combining VOTE with either TF or TF-IDF does not improve the prediction performance much (i.e., VOTE-TF and VOTE-TF-IDF). Features extracted from our

[6] http://congress.freedomworks.org/

[7] http://congressionalbills.org/

[8] https://www.govtrack.us/data/us/112/

[9] We use the default settings of SVM^{light}, except that we set the cost-factor equal to the ratio between the number of negative examples (i.e., number of non-Tea Party Caucus members) and the number of positive examples (i.e., number of Tea Party Caucus members).

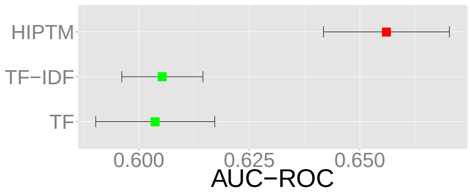

Figure 3: Tea Party Caucus membership prediction results over five folds using AUC-ROC (higher is better, random baseline achieves 0.5). The features extracted from our model for unseen legislators are estimated using their text only.

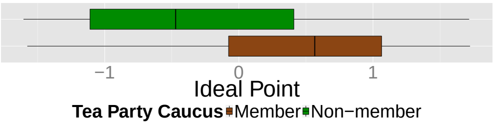

Figure 4: Box plots of the one-dimensional Tea Party ideal points, estimated as a baseline in Section 7.1, for members and non-members of the Tea Party Caucus among Republican Representatives in the 112th U.S. House. The median of members' ideal points is significantly higher than that of non-members'.

model, HIPTM, also outperform TF and TF-IDF significantly, but only slightly better than VOTE. However, HIPTM and VOTE together significantly outperform VOTE alone.

Membership Prediction from Text Only. The results in Figure 2 require both votes and legislators' language. This is limiting, since it permits predictions of "Tea Partiness" only for people with an established congressional voting record. A potentially more interesting and practical task is prediction based on language alone.

Thus, we first run our inference algorithm on the training data, which includes both votes and text. After training, using multiple models, we sample the issue and frame assignments for each token of the text authored by test lawmakers. Since the votes are not available, HIPTM's extracted features here only consist of (1) the K-dimensional vector $\boldsymbol{\eta}_k^T \hat{\boldsymbol{\psi}}_{a,k}$ estimating legislators' ideal point using text alone, and (2) the B probabilities $\Phi(x_b \sum_{k=1}^{K} \hat{\vartheta}_{b,k} u_{a,k} + y_b)$ estimating the votes.

Figure 3 compares this approach with the two baselines capable of using text alone, TF and TF-IDF. Since HIPTM can no longer access the votes in the test data, its performance drops significantly compared with VOTE. However, it still quite strongly outperforms the two text-based baselines, showing that jointly modeling the voting behavior improves the text-based elements of the model.

7 How the Tea Party Votes

In this section, we examine legislators' ideal points. We first expose Tea Party-specific ideal points by examining one-dimensional ideal points and then move on to the issue-specific ideal points that HIPTM enables.

7.1 One-dimensional Ideal Points

First, as a baseline, we estimate the one-dimensional ideal points of the legislators in our dataset.[10] Figure 4 shows the box plots of estimated Tea Party ideal points for both members and non-members.[11] The Tea Party ideal points correlate with DW-NOMINATE ($\rho = 0.91$), and the median ideal point of Tea Party Caucus members is higher than non-members. This confirms that Tea Partiers are more conservative than other Republicans (Williamson et al., 2011; Karpowitz et al., 2011; Gervais and Morris, 2012; Gervais and Morris, 2014).

Divergences involving these ideal points help demonstrate the face validity of our approach. For example, the model gives Jeff Flake (R-AZ) the second highest ideal point; he only disagrees with Freedom Works position on one of 60 Freedom Works key votes, but he is not a member of the Tea Party Caucus. Another example is Justin Amash (R-MI), who founded and is the Chairman of the Liberty Caucus. Its members are conservative and libertarian Republicans, and Amash has agreed with Freedom Works on every single key vote selected by Freedom Works since 2011.

Conversely, some self-identified Tea Partiers often disagree with Freedom Works and thus have relatively low ideal points. For example, Rodney Alexander (R-LA) only agrees with Freedom Works 48% of the time, and was a Democrat before 2004. Alexander and Ander Crenshaw (R-FL, 50% agreement) are categorized as "Green Tea" by Gervais and Morris (2014), i.e. Republican legislators who are "associated with the Tea Party on their own initiative" but lack support from Tea Party organizations.

[10] We use gradient ascent to optimize the likelihood of votes whose probabilities are defined in Equation 1. We also put a Gaussian prior $\mathcal{N}(0, \sigma)$ on u_a, x_b, and y_b.

[11] Estimated ideal point signs might be flipped, as $u_a x_b = (-u_a)(-x_b)$, which makes no difference in Equation 1. To ensure that higher ideal points are "pro-Tea Party", we first sort the legislators according to the fraction of votes for which they agree with Freedom Works and initialize the ideal points of the top and bottom five legislators with +3σ and -3σ, where σ is the variance of u_a's Gaussian prior.

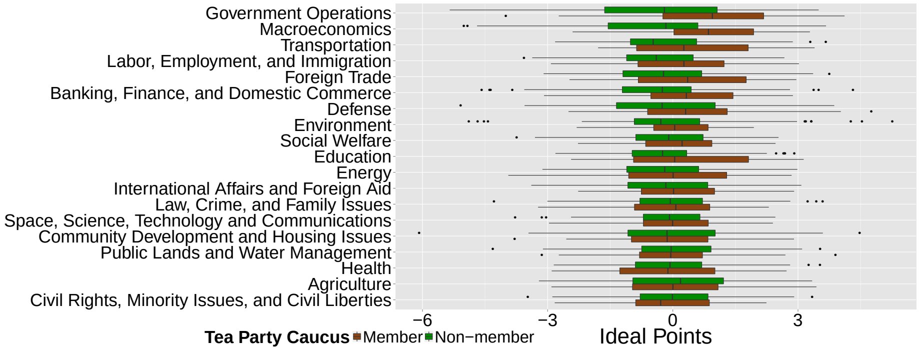

Figure 5: Box plots of ideal points dimensions, each corresponding to a major topic in the Policy Agendas Topics Codebook estimated by our model. On most issues the ideal point distributions over the two Republican groups (member vs. non-member of the Tea Party Caucus) overlap. The most polarized issues are Government Operations and Macroeconomics, which align well with the agenda of the Tea Party movement supporting small government and lower taxes.

7.2 Multi-dimensional Ideal Points

While it is interesting to compare holistic measures of Tea Partiness, it doesn't reveal *how* legislators conform or deviate from what defines a mainstream Tea Partier. In this section, HIPTM reveals how *issue-specific* ideal points of the two groups of Republican representatives differ.

Figure 5 shows estimated ideal points for each policy agenda issue, sorted by the difference between the median of the two groups' ideal points. On most issues, the ideal point distributions of the two Republican groups are nigh identical.

On several issues, though, the ideal point distributions of the two groups of legislators diverge. In the remainder of this section, we consider the Government Operation, Macroeconomics, and Transportation topics, and look at why HIPTM estimates these issues as the most polarized.

Government operations Tea Partiers differ from their Republican colleagues on reducing government spending on the Economic Development Administration, the Energy Efficiency and Renewable Energy Program and Fossil Fuel Research and Development. More specifically, for example, on the key vote *to eliminate the Energy Efficiency and Renewable Energy Program*, nearly 80% (41 out of 53) of Tea Partiers vote "Yea" (with Freedom Works) but only 43% of non-Tea Partiers agree.

Macroeconomics Our model estimates Macroeconomics policies as being the second most polarizing topic for House Republicans, which is consistent with the emphasis that the Tea Party places on issues like a balanced budget and reduced federal spending. Indeed, we see that Tea Party Republicans have distinct preferences on these types of issues as compared to more mainstream Republican legislators. An illustration of this polarization can be seen in the intra-party fight over the budget. Roll call vote 275 in 2011 and roll call vote 149 in 2012 both would have replaced Paul Ryan's budget (the "establishment" Republican budget) with the Republican Study Committee's (RSC) "Back to Basics" budget that would cut spending more aggressively and balance the budget in a decade. In 2011, non-Tea Party Republicans were evenly split in their budget preferences, but three quarters of the Tea Party Caucus supported it, which illustrates the difference between the two factions of the Republican party. Similarly, in 2012, more than 80% of Tea Partiers voted for the the RSC budget, but fewer than half of non-Tea Party Republicans did. Other polarizing votes in the Macroeconomics topic include votes to raise the debt ceiling and to avert the "fiscal cliff". In these cases, support for these votes was 25 percentage points higher among Tea Partiers than non-Tea Party Republicans, which again illustrates their distinct policy preferences.

Transportation Transportation is the third most polarized issue estimated by our model, with two key votes focusing on federal spending on transportation that illustrate some polarization, but also some shared preferences among Republicans. Consistent with the Tea Party's emphasis on reducing government spending, Tea Party Republicans voted differently from their non-Tea Party colleagues on these issues. The first key vote, roll call vote 378 in 2012, caps highway spending at the amount taken in by the gas tax. More than half of Tea Party Caucus members (32 out of 55) voted in favor, while non-members voted against it by a greater than 2:1 margin (122 of 172). Conversely, the second key vote (roll call vote 451 in 2012) authorizes fed-

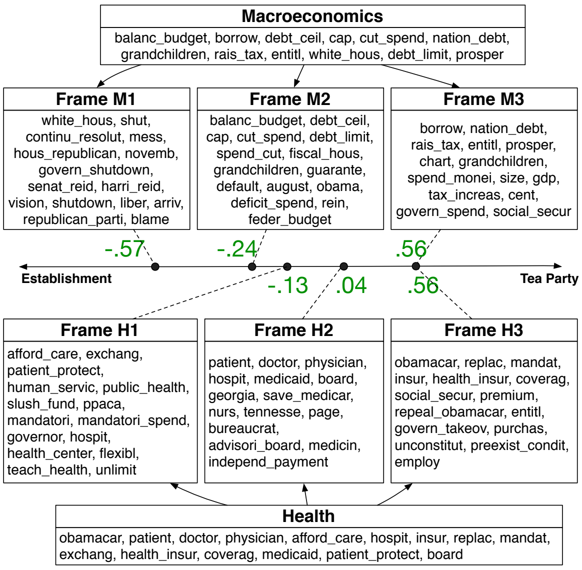
Macroeconomics
balanc_budget, borrow, debt_ceil, cap, cut_spend, nation_debt, grandchildren, rais_tax, entitl, white_hous, debt_limit, prosper
Frame M1
white_hous, shut, continu_resolut, mess, hous_republican, novemb, govern_shutdown, senat_reid, harri_reid, vision, shutdown, liber, arriv, republican_parti, blame
Frame M2
balanc_budget, debt_ceil, cap, cut_spend, debt_limit, spend_cut, fiscal_hous, grandchildren, guarante, default, august, obama, deficit_spend, rein, feder_budget
Frame M3
borrow, nation_debt, rais_tax, entitl, prosper, chart, grandchildren, spend_monei, size, gdp, tax_increas, cent, govern_spend, social_secur
-.57
-.24
.56
Establishment
Tea Party
-.13
.04
.56
Frame H1
afford_care, exchang, patient_protect, human_servic, public_health, slush_fund, ppaca, mandatori, mandatori_spend, governor, hospit, health_center, flexibl, teach_health, unlimit
Frame H2
patient, doctor, physician, hospit, medicaid, board, georgia, save_medicar, nurs, tennesse, page, bureaucrat, advisori_board, medicin, independ_payment
Frame H3
obamacar, replac, mandat, insur, health_insur, coverag, social_secur, premium, repeal_obamacar, entitl, govern_takeov, purchas, unconstitut, preexist_condit, employ
Health
obamacar, patient, doctor, physician, afford_care, hospit, insur, replac, mandat, exchang, health_insur, coverag, medicaid, patient_protect, board

Figure 6: Framing of Macroeconomics (top) and Health (bottom) among House Republicans, 2011-2012. Higher ideal point values are associated with the Tea Party.

eral highway spending at a level that far exceeds its revenue from the gas tax, which was opposed by Freedom Works. This measure was broadly popular with Republicans regardless of Tea Party affiliation and a majority of both Tea Partiers and non-Tea Partiers opposed it.

8 How the Tea Party Talks

Looking at HIPTM's induced topic hierarchy, using labeled data to create informative priors produces highly interpretable topics at the agenda-issue level; e.g., see the first-level nodes in Figure 6, which capture key issue-level debates. For example, one major event during the 112th Congress was the 2011 debt-ceiling crisis, which dominates discussions in Macroeconomics. Similarly, Defense is dominated by withdrawing U.S. troops from Iraq.

Turning to framing, recall that second-level nodes of the hierarchy capture issue-specific frames of parent issues, each one associated with a frame-specific ideal point. To analyze intra-Republican polarization, we first compute, for each issue k, the *span* of ideal points the frames associated with k, i.e., the difference between the maximum and the minimum ideal points for frames under that issue.[12] We then consider several issues with a large span, i.e. whose frames are highly polarized.

Macroeconomics. The HIPTM subtree for Macroeconomics, in Figure 6 (top), foregrounds Republican polarization related to budget issues. The most Tea Party oriented frame node, M3, focuses on criticizing government overspending, a recurring Tea Party theme.[13] In contrast, Frame M1, least oriented toward the Tea Party, focuses on the downsides of a government shutdown, highlighting establishment Republican concerns about being held responsible for the political and economic consequences.

Health. Healthcare was a central issue during the 112th Congress, particularly the Affordable Care Act (Obamacare). Although all Republicans voted to repeal Obamacare, Figure 6 (bottom) highlights intra-party differences in framing the issue. Frame H1 leans strongest toward the establishment Republican end of the spectrum, and frames opposition in terms of the implementation of health care exchanges and the mandatory costs of the program. In contrast, H3 captures the more strident Tea Party framing of Obamacare as an unconstitutional government takeover. More neutral from an intra-party perspective, Frame H2 emphasizes Medicare, Medicaid, and the role of health care professionals within these systems.[14]

Labor, Employment and Immigration. The discussion of this issue illustrates how HIPTM sometimes captures frames that are distinct from Tea Partiness, *per se*. For example, it discovered a strongly Tea Party oriented frame that focused on "union, south carolina, nlrb, boeing". On inspection, this frame reflects a controversy in which the National Labor Relations Board accused airline manufacturer Boeing of violating Federal labor law by transferring production to a non-union facility in South Carolina "for discriminatory reasons",[15] and surfaces mainly in speeches by four legislators from South Carolina, three of whom are from the Tea Party Caucus. This second-level topic illustrates a limitation of HIPTM; it does not formally distinguish frames from other kinds of subtopics. We observe that modeling polarization on other kinds of sub-issues is nonetheless valuable: here it highlights a geographic locus of conflict involv-

[12] The frame proportions Dirichlet process ψ_k creates many frames with one or two observations (Miller and Harrison, 2013). We ignore those with posterior probability $\psi_{k,j} < 0.1$.

[13] E.g., Scott Garrett (R-NJ): "We will not compromise on our principles; our principles of defending the Constitution and defending Americans and making sure that our posterity does not have this excessive debt on it."

[14] This does not mean that discussions using this frame lacked combative or partisan elements. For example, Glenn Thompson (R-PA) argues that "on the Democratic side, they're just willing to pull the plug and let [Medicare] die".

[15] http://www.nlrb.gov/news-outreach/fact-sheets/fact-sheet-archives/boeing-complaint-fact-sheet

Ideal Point Distributions		Not	Polarized
Distribution of Issue Frames	Not	Civil Rights, Minority Issues, Civil Liberties	Banking and Finance; Transportation
	Polarized	Health; Public Lands and Water Management	Macroeconomics; Government Operations

Table 2: Examples of agenda issues classified by polarization of ideal points and issue frames within the Republican party.

ing South Carolina, where many representatives are Tea Party Caucus members. This may provide insight into how geography shapes Tea Party membership (Gervais and Morris, 2012).

9 Latent and Visible Disagreement

Our analyses suggest a novel framework for understanding the political and policymaking implications of the Tea Party in the 112th Congress, illustrated in Table 2. Each issue can be characterized by two features: (1) the degree to which ideal points among Republican legislators are polarized, and (2) the degree to which the frames used are polarized. From these two assessments, we can organize all policy issues into four categories that have meaningful implications for congressional politics and policy outcomes. At upper left we will find issues where HIPTM indicates low intra-party polarization between Tea Party and non-Tea Party, and all Republicans tend to frame the issue in similar ways; e.g., Civil Rights, Minority Issues, and Civil Liberties. In such cases, we expect cooperation among Republicans regardless of Tea Party status, therefore a greater likelihood of bill passage in a majority-Republican House. In stark contrast, issues at lower right involve polarized ideal points and polarized framing, e.g., the budget crisis, where many establishment Republicans balked at a government shutdown but hard-line Tea Party legislators did not. These issues pose the greatest challenge to Republican party leaders.

Between these extremes are the issues in which *either* Republicans' ideal points *or* their policy frames are polarized. Our model suggests that on issues at upper right, with similar framing, the Tea Party and establishment Republicans will *appear* to be in sync, and therefore it may seem to voters that legislative progress is likely, but the underlying issue polarization will make it hard to find policy common ground, potentially increasing public frustration. Last, at lower left are issues where Republicans generally share similar ideal points and vote similarly, but frame the issue in distinct ways, e.g., Obamacare. Here legislative success may come despite the appearance that Republican factions are talking past each other, because the distribution of their ideal points on the policy is actually quite similar. Put differently, Republicans share policy goals on issues in this quadrant even if they frame those preferences differently, and this underlying agreement on the ideal point may allow Republicans to reach consensus even when the political rhetoric suggests otherwise.

10 Conclusion

We introduce HIPTM, which integrates hierarchical topic modeling with multi-dimensional ideal points to jointly model voting behavior, the text content of bills, and the language used by legislators. HIPTM is more effective than previous methods on the task of predicting membership in the Tea Party Caucus. This improvement is especially consequential as the formal organization of the Tea Party Caucus is now defunct in the House, yet Tea Party legislators remain both numerous and influential in Congress. In addition, unlike previous ideal-point methods, HIPTM makes it possible to make predictions for members of Congress who have not yet established a voting record. More intriguingly, this also suggests the possibility of assessing the "Tea Partiness" of candidates (or, anyone else, e.g., media outlets) based on language.

It is political conventional wisdom that the influx of Tea Party legislators in the 112th Congress complicated the task of governance and policymaking for Republican leaders. By looking at issue-level ideal points and issue-specific framing using our model, we begin to address the complexity of this relationship, finding the model successful both in establishing face validity and in suggesting novel insights into the dynamics of a Republican Congress. In future work, we plan to pursue the new framework suggested by our analyses, investigating the interaction of issue polarization and framing-based polarization. With the help of these new tools, we aim to both understand and predict substantive policy areas in which the Tea Party is likely to be most successful working with the Republican party, and, conversely, to flag ahead of time policy areas in which we can expect to see legislative gridlock and grandstanding.

Acknowledgements

This work was supported by the NSF Grant IIS-1211153 and IIS-1409287. Boyd-Graber is also supported by NSF grants IIS-1320538 and NCSE-1422492. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

E. Scott Adler and John Wilkerson. 2015. Congressional bills project. NSF 00880066 and 00880061.

Amr Ahmed, Liangjie Hong, and Alexander Smola. 2013. Nested Chinese restaurant franchise process: Applications to user tracking and document modeling. In *Proceedings of the International Conference of Machine Learning*, pages 1426–1434.

Charles E. Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174.

Joseph Bafumi, Andrew Gelman, David K Park, and Noah Kaplan. 2005. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13(2):171–187.

David M. Blei and John Lafferty, 2009. *Text Mining: Theory and Applications*, chapter Topic Models. Taylor and Francis, London.

David M Blei and Jon D McAuliffe. 2007. Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 121–128.

Edward G Carmines and Nicholas J D'Amico. 2015. The new look in political ideology research. *Annual Review of Political Science*, 18(4).

Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review*, 98(02):355–370.

Philip J Cowans. 2006. *Probabilistic Document Modelling*. Ph.D. thesis, University of Cambridge.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Jean-Paul Fox. 2010. *Bayesian item response modeling: Theory and applications*. Springer.

Matt Fuller. 2015. New Tea Party Caucus chairman: DHS fight could break the GOP. *Roll Call*, February. http://blogs.rollcall.com/218/new-tea-party-caucus-chairman-dhs-fight-could-break-the-gop/.

Sean Gerrish and David M. Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of the International Conference of Machine Learning*, pages 489–496.

Sean Gerrish and David M. Blei. 2012. How they vote: Issue-adjusted models of legislative behavior. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2753–2761.

Bryan T Gervais and Irwin L Morris. 2012. Reading the tea leaves: Understanding Tea Party Caucus membership in the US House of Representatives. *PS: Political Science & Politics*, 45(02):245–250.

Bryan T Gervais and Irwin L Morris. 2014. Black Tea, Green Tea, White Tea, and Coffee: Understanding the variation in attachment to the Tea Party among members of Congress. In *Annual Meeting of the American Political Science Association*.

Justin Grimmer. 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1):1–35.

Yupeng Gu, Yizhou Sun, Ning Jiang, Bingyu Wang, and Ting Chen. 2014. Topic-factorized ideal point estimation model for legislative voting network. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 183–192.

James J. Heckman and James M. Snyder Jr. 1997. Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *The RAND Journal of Economics*, 28:142–189.

Yuening Hu and Jordan Boyd-Graber. 2012. Efficient tree-based topic modeling. In *Association for Computational Linguistics*.

Simon Jackman. 2001. Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis*, 9(3):227–241.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - SVM*. Universität Dortmund.

Christopher F Karpowitz, J Quin Monson, Kelly D Patterson, and Jeremy C Pope. 2011. Tea time in America? The impact of the Tea Party movement on the 2010 midterm elections. *PS: Political Science & Politics*, 44(02):303–309.

Benjamin E. Lauderdale and Tom S. Clark. 2014. Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, 58(3):754–771.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Lee B. Lusted. 1971. Signal Detectability and Medical Decision-Making. *Science*, 171:1217–1219, March.

Andrew D Martin and Kevin M Quinn. 2002. Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999. *Political Analysis*, 10(2):134–153.

Maxwell McCombs. 2005. A look at agenda-setting: Past, present and future. *Journalism Studies*, 6(4):543–557.

Jeffrey W Miller and Matthew T Harrison. 2013. A simple example of dirichlet process mixture inconsistency for the number of components. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 199–206. Curran Associates, Inc.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1106–1114.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2014. Sometimes average is best: The importance of averaging for prediction using MCMC inference in topic modeling. In *Proceedings of Emperical Methods in Natural Language Processing*, pages 1752–1757.

Thang Nguyen, Jordan Boyd-Graber, Jeff Lund, Kevin Seppi, and Eric Ringger. 2015. Is your anchor going up or down? Fast and accurate supervised topic models. In *North American Association for Computational Linguistics*.

John Paisley, Chong Wang, David M Blei, and Michael I Jordan. 2014. Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Keith T Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384.

Keith T Poole and Howard L Rosenthal. 2007. *Ideology and Congress*. New Brunswick, NJ: Transaction Publishers.

Gerard. Salton. 1968. *Automatic Information Organization and Retrieval.* McGraw Hill Text.

Yanchuan Sim, Bryan Routledge, and Noah A Smith. 2015. The utility of text: The case of Amicus briefs and the Supreme Court. In *Association for the Advancement of Artificial Intelligence*.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476).

Hanna M Wallach. 2008. *Structured Topic Models for Language*. Ph.D. thesis, University of Cambridge.

Vanessa Williamson, Theda Skocpol, and John Coggin. 2011. The Tea Party and the remaking of Republican conservatism. *Perspectives on Politics*, 9(01):25–43.