A Robust Latent Dirichlet Allocation Approach for the Study of Political Text

Andreu Casas
New York University

Tianyi Bi
New York University

John Wilkerson
University of Washington

Abstract

A rapidly growing body of research in political science uses unsupervised topic modeling techniques such as Latent Dirichlet Allocation (LDA) models to construct measures of topic attention for the purpose of hypothesis testing. A central advantage of unsupervised topic modeling, compared to supervised approaches, is that it does not require advance knowledge of the topics to be studied, or a sizable set of training examples. However, the topics discovered using these methods can be unstable. This is potentially problematic to the extent that researchers report results based on a single topic model specification. We propose an approach to using topic model results for hypothesis testing that incorporates information from multiple specifications. We then illustrate this robust approach by replicating an influential political science study. An `R` package (`ldaRobust`) for its implementation is provided.[1]

[1]Prepared for the 2018 Conference of the American Political Science Association, Boston Aug 29-Sept 2.
We appreciate all comments and suggestions (`andreucasas@nyu.edu`). Please do not cite without permission.

1 Introduction

Topics have always been important variables in many areas of political science research. The discipline has also seen considerable innovation in machine learning methods for automatically assigning topics to documents.[2] Supervised machine learning methods use labeled example documents provided by the researcher to train algorithms to predict the topics of other unlabeled documents (Collingwood and Wilkerson, 2012; Workman, 2015; Casas and Wilkerson, 2017). Unsupervised learning methods are designed to help researchers 'discover' themes or topics in large volumes of documents (Quinn et al., 2010; Grimmer, 2013). In computer science, unsupervised methods are frequently used as a precursor to supervised approaches. In political science, unsupervised methods have become popular tools for assigning labels in part because they do not require large numbers of labeled examples.

However, unsupervised methods have also come under increasing scrutiny as they have become more popular. In particular, researchers have found that modeling decisions can have important implications for the topics that are discovered. These findings are concerning because the common practice in political science research is to select a single topic model specification as the best representation of the data. Researchers have proposed tools and practices to support this single model selection process (Roberts et al., 2013; Wallach et al., 2009), but there is little evidence to indicate that these interventions resolve concerns about topic instability.

The alternative appproach we propose here is to incorporate information about topic instability into hypothesis testing practices. Researchers commonly report alternative model specification in quantitative research. In this paper we propose a *Robust Latent Dirichlet Allocation* approach for testing the sensitivity of empirical findings to alternative topic model specifications. We focus on one important parameter - the number of topics k that the

[2]See Grimmer and Stewart (2013) and Wilkerson and Casas (2016) for overview.

researcher must designate *ex ante*. We draw on established practices to first identify a single "best" topic model. We then estimate additional models with fewer and more topics. We use document clustering techniques to align similar topics across models. Finally, we replicate each hypothesis test using the information from each topic model, and report average effects, with confidence bounds corresponding to the extreme bounds of all the models. Our `R` package (`ldaRobust`) facilitates implementation of this robust approach.

This paper is organized as follows. We begin by presenting the Robust LDA approach in more detail. We introduce LDA topic models and our approach to specifying additional models after a researcher has selected a single best model. We then describe the methods for computing topic similarities and for clustering similar topics. In the second half of the paper, we apply this method to Justin Grimmer's 2013 study of U.S. Senators' press releases (Grimmer, 2013). We first replicate his original 44 topic model and then estimate an additional ensemble of 6 other k=41-47 models from the same data. After using clustering methods to align topics across these 7 models, we compare the topics of the k=44 model with those of its k neighbors. This examination underscores the topic instability concern that motivates the project. We then estimate 7 regressions predicting differences in topic emphasis among senators, and compute average results and uncertainty bounds across the 7 estimations. Finally we compare the results to the original results of the Grimmer study.

2 Unsupervised Topic Modeling in Political Science

In one of the earliest political science studies applying automatic text classification methods, Quinn et al. (2010) demonstrated how unsupervised methods could be used to discover policy topics in the floor speeches of by U.S. Senators. Since then, unsupervised models have become increasingly accessible, thanks in part to exceptionally helpful software packages also developed by political scientists (Roberts et al., 2013).

Political scientists in a wide range of fields are now using unsupervised topic models in their research. For example, Roberts et al. (2013) uses them to estimate treatment effects in open-ended surveys. Lucas et al. (2015) use them to examine the topics of fatwas and social media messages in different languages (after first translating the messages to english). Bagozzi and Berliner (2016) apply topic models to U.S. State Department Country Reports on Human Rights Practices to track human rights violations attention over time. Berliner et al. (2018) use them to study government accountability in Mexico by examining information requests filed with Mexican federal government agencies. Farrell (2016) and Boussalis and Coan (2016) use them to study the arguments of climate change countermovement organizations. Jacobi et al. (2016) explore the issues discussed in the New York Times across time. Barbera et al. (2018) apply topic modeling to 50 million tweets to examine the responsiveness of members of the U.S. Congress.

In some but not all of these studies, the authors used unsupervised topic modeling not only for discovery but also to test hypotheses. In one of the first studies of this kind, Grimmer (2013) classified 60,000 senator press releases into 44 topics derived from a topic model. The question of interest was whether Senators from electorally competitive states were more likely to emphasize credit claiming (topics) and less likely to express specific policy positions than their safer counterparts. He found that this was indeed the case.

2.1 The Elephant in the Room: Topic instability

The common practice in all of these studies is to selecting a single topic model as best representing the data after exploring the result for several different models. In terms of model choice, computer scientists recommend a number of goodness of fit measures for selecting the best model (Wallach et al., 2009). However, this approach has been criticized because experiments demonstrate that humans do not always agree that the best model in terms of objective fit measures is also the best from a semantic perspective (Chang et al.,

2009).

Other scholars recommend focusing on whether a model does a good job of capturing what the researcher is trying to capture. Quinn et al. (2010) and Grimmer and Stewart (2013) recommend that researchers estimate several models by varying the number of topics k before using "human judgment" to "assess the quality of the clusters" (Grimmer and Stewart, 2013, 20). Quinn et al. (2010) propose five quality assessment criteria: semantic validity (are the most predictive features in each topic about the same substantive topic?), convergent construct validity (do the resulting topics correlate with other measures capturing the same construct?), discriminant construct validity (are the resulting topics negatively correlated to measures capturing distinct constructs?), predictive validity (do the resulting topic classifications predict external real-world events one would expect?), and hypothesis validity (does the nature and scope of the topics serve the researcher's substantive goals?). In their STM package, Roberts et al. (2013) offer objective metrics to help evaluate the quality of a model by (among other things) giving preference to models with more semantically coherent and discriminating topics. However, they too emphasize that these metrics should not be treated as substitutes for human judgment.

An important limitation of this best single-model approach is that information is unavoidably lost in choosing one model at the expense of other. This was less of a concern when topic models were used for initial discovery and as the starting point for supervised approaches. It become more important when topic model results are used as measures themselve. And only recently have scholars begun to assess the potential costs of committing to a single model (Chuang et al., 2015; Roberts et al., 2016; Wilkerson and Casas, 2016; Denny and Spirling, 2018).

Figure 1: Example of Topic Instability: Comparing a model with 41 and 42 topics fit to one minute floor speeches by members of the U.S. Congress

Figure 1 presents an example. Here we fit two topic models k=41 and k=42 to a corpus of congressional one minute floor speeches. We then use cosine similarity to compare the similarity of the topics of each model. In the figure, darker shading indicates greater similarity: *Very High* ($> .95$), *High* ($> .90$), *Medium* ($> .80$), or *Low* ($< .80$).

The diagonal in figure 1 indicates topics that are present in both models. The far right column indicates the additional topic of the 42 topic model. There is a lot to the figure but the important point for our purposes is that there are six topics along the diagonal that - while present in both models - are less similar across the two models. When the goal is to simply discover topics, these differences may not be remarkable. But when the goal is to use the results to estimate topic attention across a set of documents, or to assign documents to topics based on word co-occurence, then these differences could very well lead to different findings. Whether this is the case cannot be known when researchers commit *ex ante* to a single topic model specification.

3 The Robust Latent Dirichlet Allocation Approach

We propose a method for incorporating information from multiple topic model specifications into a hypothesis test. We assume that the researchers has already settled on an *original* LDA model m_O with k_O topics that, in their view and based on available metrics, best captures their subject of interest.

As described in Blei et al. (2003), LDA is a mixed membership topic modeling approach. An indexed vector of unique words V in the corpus is the starting point.[3] Each latent topic is a probability distribution over these words. And each document is assumed to be a random mixture of these latent topics. More precisely the generative model assumes:

1 The number of words in the document is determined by: $N \sim Poisson(\xi)$

[3]N-gram is the more precise unit.

2 The topic distribution for the documents is determined by: $\theta \sim Dir(\alpha)$

3 The word distribution for topic k is determined by: $\beta \sim$ Dirichlet(δ)

4 For each N word in document w the algorithm assigns:

(a) A topic $z_s \sim$ Multinomial(θ)

(b) A word w_n from $p(w_n|z_s, \beta)$: a multinomial probability conditioned on z_s.

The next step is to estimate an ensemble of Z (LDA) models denoted by $\mathbf{m} = \{m_1, m_2, ..., m_Z\}$, where each model m_j has a different k number of topics extracted from a vector of length Z, denoted by $\mathbf{k} = (k_1, k_2, ..., k_Z)$, where the initial model k_O is at the median. This first step produces a list $\mathbf{L}$ of size $\sum_{i=1}^{Z} k_i$ containing each topic s (β_s) of each model j (m_j) in $\mathbf{m}$: $\{\beta_{m_1,1}, \beta_{m_1,2}, ..., \beta_{m_2,1}, \beta_{m_2,2}, ..., \beta_{m_j,s}, ..., \beta_{m_Z,k_Z}\}$. We then calculate pairwise similarity scores for all possible topic pairings. This produces a similarity matrix $\mathbf{s}$ of size $(\sum_{i=1}^{Z} k_i) \times (\sum_{i=1}^{Z} k_i)$.

The information from this similarity matrix is then used to align topics across the different models. We start by attempting to align each topic of the original model $\beta_{m_O,s}$ with topics of the alteernative models ($\beta_{m_j s}$) using a similarity threshold ϕ that is chosen by the researcher. We then use ϕ to align additional topics of others models that were not part of the original model. This yields a topic alignment matrix $\mathbf{q}$ of size $(\sum_{i=1}^{Z} k_i) \times (>= k_O)$ that can be used to ask whether results (based on the topics of the original model) change if the same topics are drawn from a somewhat different model.

In the remainder of this section we describe these steps in more detail. In the section that follows we demonstrate the method by applying it to Grimmer (2013)'s analysis of senators' press releases.

3.1 Computing Topic Similarities

To reiterate, each topic β_{js} is a vector of indexed word probabilities expressing the likelihood that each word w in V is associated with a particular topic. This indexed structure means that: a) all topics β_{js} are vectors of the same size V, and b) any n parameter in any topic vector β_{js} contains information about the same word w_n.

A variety of similarity algorithms can be used to compute the similarity of these indexed vectors (such as Euclidean distance, Manhattan distance, Jaccard similarity, and Cosine similarity) and to construct a topic similarity matrix $\mathbf{s}$ of size $(\sum_{i=1}^{Z} k_i) \times (\sum_{i=1}^{Z} k_i)$.

3.2 Aligning topics across models

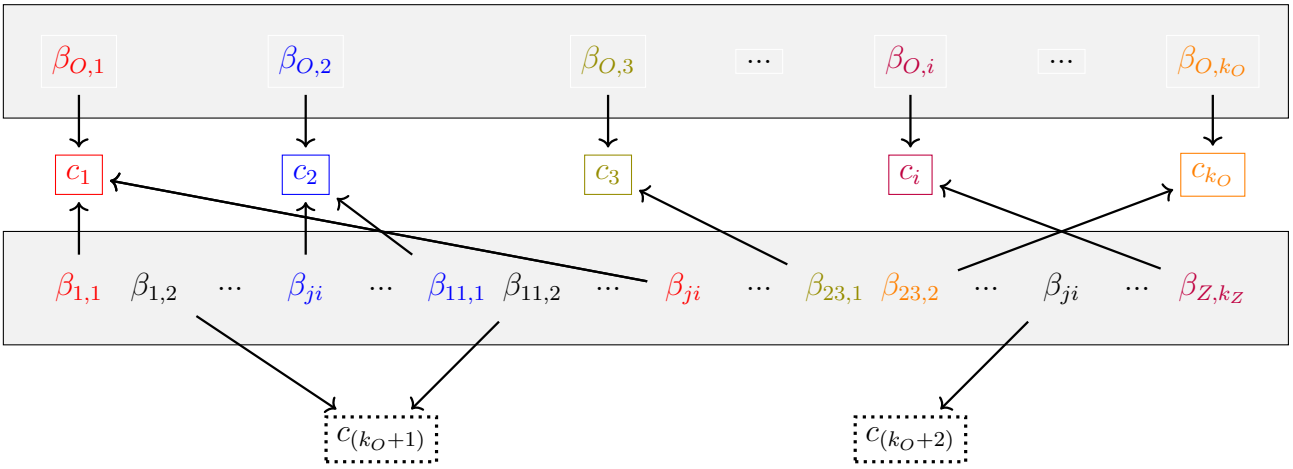
The process of clustering and aligning topics across models is summarized in figure 2. For each topic of the original model β_{Oi}, we build a topic cluster c_i that includes all topics from alternative models β_{js}[4] that exceed a topic similarity $T(\beta_{js}, \beta_{Os})$[5] that exceeds a pre-determined similarity threshold ϕ. To avoid double counting topics in the event that an alternative topic exeeded these thresholds for two of the original topics, we only align alternative model topics β_{js} to a single most similar original model topic β_{Oi}.

$$c_i = \beta_{js} \in \mathbf{L} \mid \big(max(T(\beta_{js}, \beta_O)) = T(\beta_{js}, \beta_{Oi}) > \phi\big) \tag{1}$$

Setting this threshold is an interesting question in itself. In general, a finding that holds up under a lower threshold is more robust. A very high similarity threshold means that it will be difficult to compare results for different models because they will have few shared topics. A low threshold risks an apples to oranges comparison - topics do align but they are semantically different. One option would be to start with a high threshold and examine the

[4]Where $j \neq O$.
[5]Which we can extract from $\mathbf{s}$.

Figure 2: Visual representation of the clustering process.



topics that do not align with any original topics (for example by examining the top loading terms). If some topics of alternative model topics seem like they should align but do not, then the threshold should probably be lowered.

We then have a group of topic clusters $\mathbf{c}$ for the k_O topics of the original model. We may or may not have additional clusters that do not include topics from the original model denoted by $P = \beta_{js} \notin \mathbf{c}$. Because there is no such thing as *the* correct topic model, we also want to consider how including these topics impacts our findings. We therefore compare each t potential remaining topic in P to every other t potential remaining topic. Any additional topic clusters, based on the threshold, are then added to $\mathbf{c}$.

$$c_t = \beta_{js} \in P \mid \left(\beta_{js} \notin \mathbf{c} = T(\beta_{js}, \beta_t) > \phi\right) \tag{2}$$

The final product of this process is a group of topic clusters $\mathbf{c}$ of size $>= k_O$ that provides insights into how varying k alters topic substance beyond simply adding or subtracting topics from the model. As illustrated in Figure 1, we may also discover that varying k alters the word probabilities of the original model topics in ways that impact research findings.

A *beta* version of our (`ldaRobust`) package can be found at: https://github.com/CasAndreu/ldaRobust. The package includes functions to implement the steps described above as well as to replicate a quantitative analysis using the topics of each model in the ensemble. We suggest reporting average effects and confidence bounds that reflect the extreme upper and lower bounds of the models.

4 The Robust Approach in Practice

In *Appropriators not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation*, Grimmer (2013) uses unsupervised topic modeling to discover the topics addressed in over 60,000 press releases by U.S. senators. The question of interest from that project that we examine here is: do senators from electorally competitive states communicate differently than senators from safer states? Grimmer finds that Senators from competitive states (those with a higher proportion of other party supporters) are more likely to clam credit for the things they have brought to the district and less likely to express policy positions. Grimmer further argues that this dynamic contributes to an 'artificial' polarization of the national political debate because the views of senators with more moderate policy views tend to be underrepresented in the public dialogue.

Grimmer develops an unsupervised mixed membership Bayesian topic model to *simultaneously* estimate four quantities of interest: the topics discussed in press releases; the dominant topic of each press release; the topic proportions of each senator's press releases in each year, and each Senator's "home style" (the proportion of their press releases that are primarily about credit-claiming *versus* position-taking). He finds that a 44 topic model best captures the distribution of topics across of the press releases.

4.1 Topic Clusters in Senators' Press Releases

We begin by fitting an LDA model with 44 topics (m_O).[6] We then fit six additional nearby models (in all $\mathbf{k} = \{41, 42, 43, 44, 45, 46, 47\}$). We then align each of the 44 original topics to topics from the 6 alternative models based upon a cosine similarity threshold of .95.[7] In addition, we align alternative topics that did not align with any of the original 44 topics using the .95 threshold.

Figure 3 displays the results of this alignment process. In addition to the original 44 topics we add two more topics. The topic clusters are labeled on the left y-axis and the most predictive words for each are displayed on the right y-axis. Each column represents one of the alternative models (the model with 43 topics is on the far left; 47 on the far right). With the exception of the bottom two rows, the cell colors indicate whether an alternative model includes an aligning topic (green) or not (grey).

Most of the topics of the original model are also found in all of the the alternative models. Some others are not however. Child safety, Regional Development and Procedural are only found in one of the alternative models. Budget and Urban Development are only found in two of them. In other words, five of the original 44 topic do not appear in a majority of the nearby models. In addition, two of the models with a larger k include a Retirement topic that is not a topic of the original model.

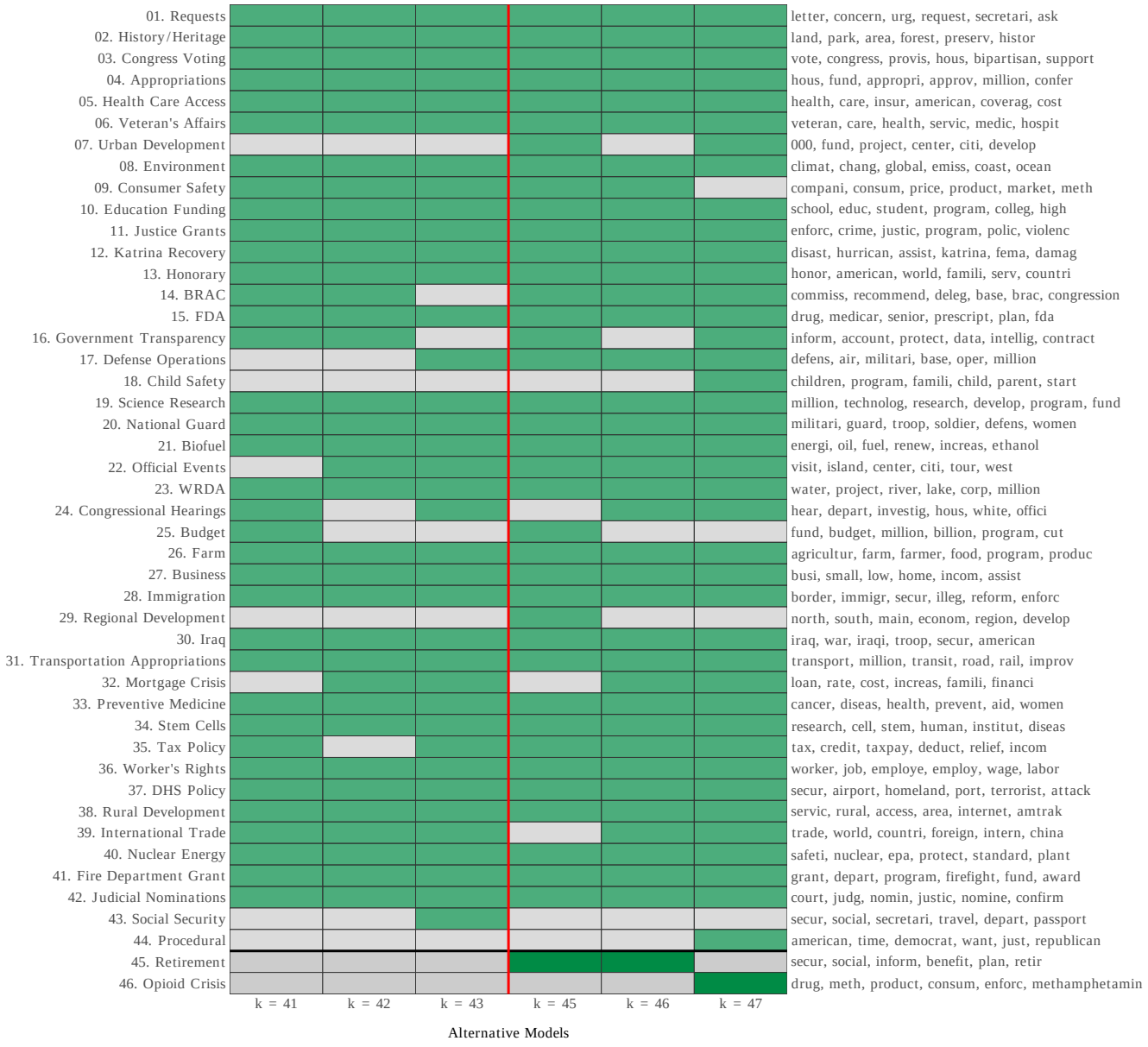
4.1.1 Topic Instability and Document Classification

Would we arrive at different conclusions about the impact of electoral competitiveness on senators' home styles if we used one of these other models? In Figure 4 we begin exploring the impat of topic instability by examining the proportion of press releases that are primarily

[6] Because pre-processing decisions can also have important effects Denny and Spirling (2018) we fit the 44-topic model to a pre-processed Document Term Matrix that Grimmer includes in the replication materials of his article: https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/18602

[7] We also explored lower thresholds but found that we were aligning (in our view) substantially different topics.

Figure 3: Discovered topic cluster and whether they are present in the topic models explored.



about each topic according to each of the models. Once again, each topic is described in the left and right y-axes. Each cross for each topic indicates the proportion of press releases deemed to be about that topic by a given model; the dots represent the average proportion

across all of the models models[8]; and the lines represent the 95% confidence intervals of the averages.[9]

Some topic proportions such as Honorary and Stem Cell Research are very stable across model specifications. The averages for these two topics are very similar to the numbers in Grimmer (2013)'s analysis (see Table 1 in p.629). But the more important point of the figure from our perspective is the broad range of estimated topic proportions for many topics (the x's) and the very wide confidence intervals in many cases. This is even the case for many topics that appear in all of the models. The estimated proportions for Appropriations, for example, range from 1.75% to 7.75%. We can definitively say that there are differences in topic attention in some cases (e.g. War versus Water resources). But in most cases we cannot have much confidence that their proportions differ. Recall that a single model approach would not let us evaluate such questions.

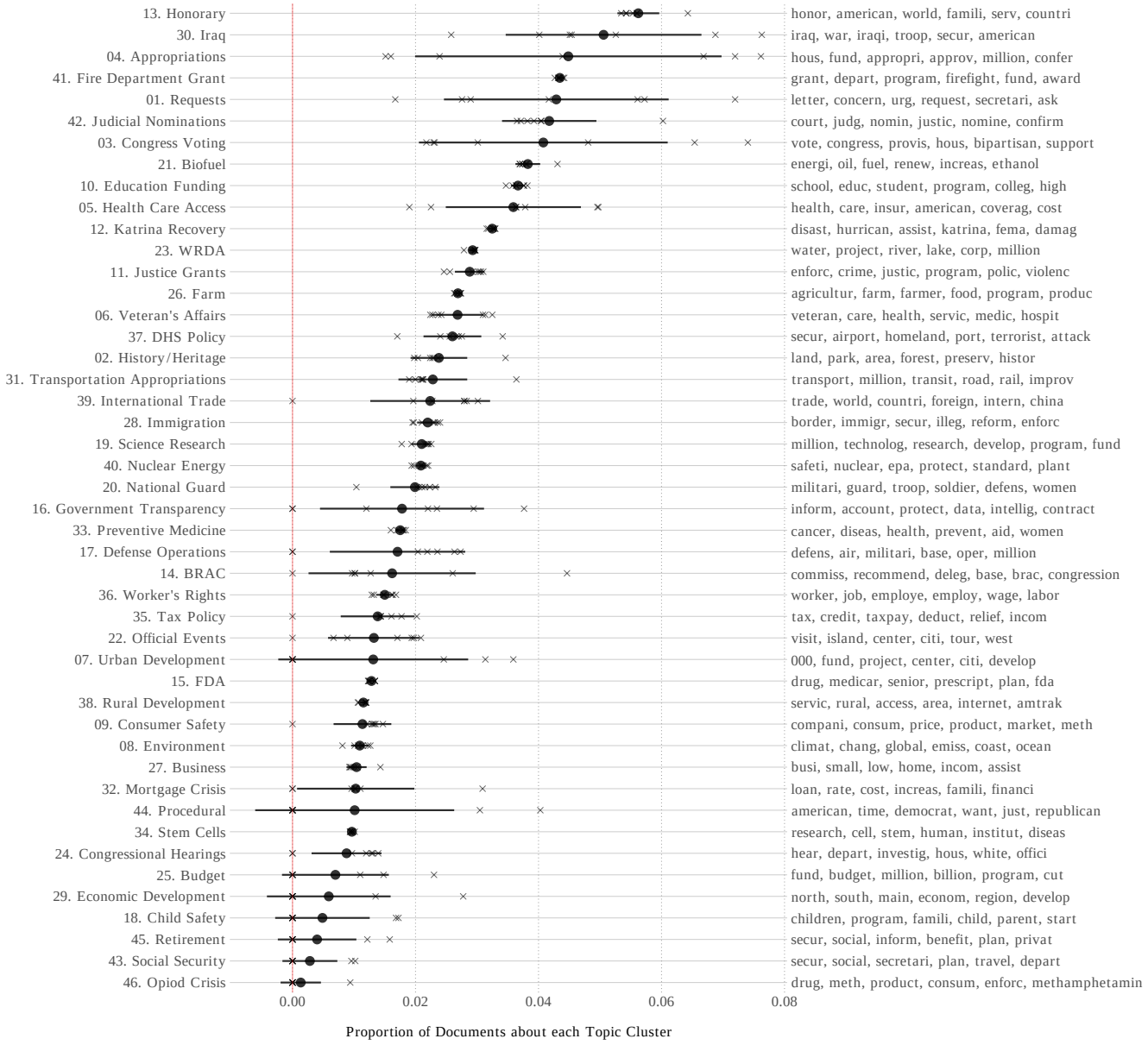
4.1.2 Topic Instability and Covariate Effects

Which senators dedicated more attention to the different topics? In Figure 5 we use the same robust approach to ask whether ideology and party correlate with topic attention. For each covariate of interest and for each topic cluster, we estimate two sets of seven logistic regressions where the dependent variable is topic attention and the independent variable is the senator's ideology or party. For each topic cluster we then compute the average effect π_i. More specifically, let topic models with at least one topic belonging to the cluster i be denoted by $\mathbf{r}_i = m_j \in \mathbf{m} \mid (\beta_{js} \in c_i) > 0$, and let the length of $\mathbf{r}_i$ be expressed by R_i. We can then calculate π_i as follows:

[8] $\frac{\sum_{j=1}^{Z} \left(\frac{\sum_{g=1}^{M} (\mathbf{w}_{g z_{m_j s}} | \beta_{m_j s} \in c_i)}{M} \right)}{Z}$

[9] We calculate confidence intervals for the average by assuming a t distribution. This confidence interval takes into account the spread of the 7 observations to provide information about how confident we are that the particular average is the true-correct value.

Figure 4: Clusters of Topics in Senators' press releases and proportion of documents dominated by each cluster.



$$\pi_i = \frac{\sum_{j=1}^{R_i} \left(\frac{1}{1 + exp(-x_d \delta_{dji})} \right)}{R_i} \tag{3}$$

where x_d is a covariate with document, author, or cluster-level information. The statis-

tical model then allows us to estimate a set of cluster-level parameters δ_{di} for any covariate of interest by calculating the average parameter across the models in $\mathbf{r}_i$: $\delta_{di} = \frac{\sum_{j=1}^{R_i}(\delta_{dji})}{R_i}$.

Figure 5: Bivariate relationship between Senator-level covariates and the topic clusters discussed in press releasses.

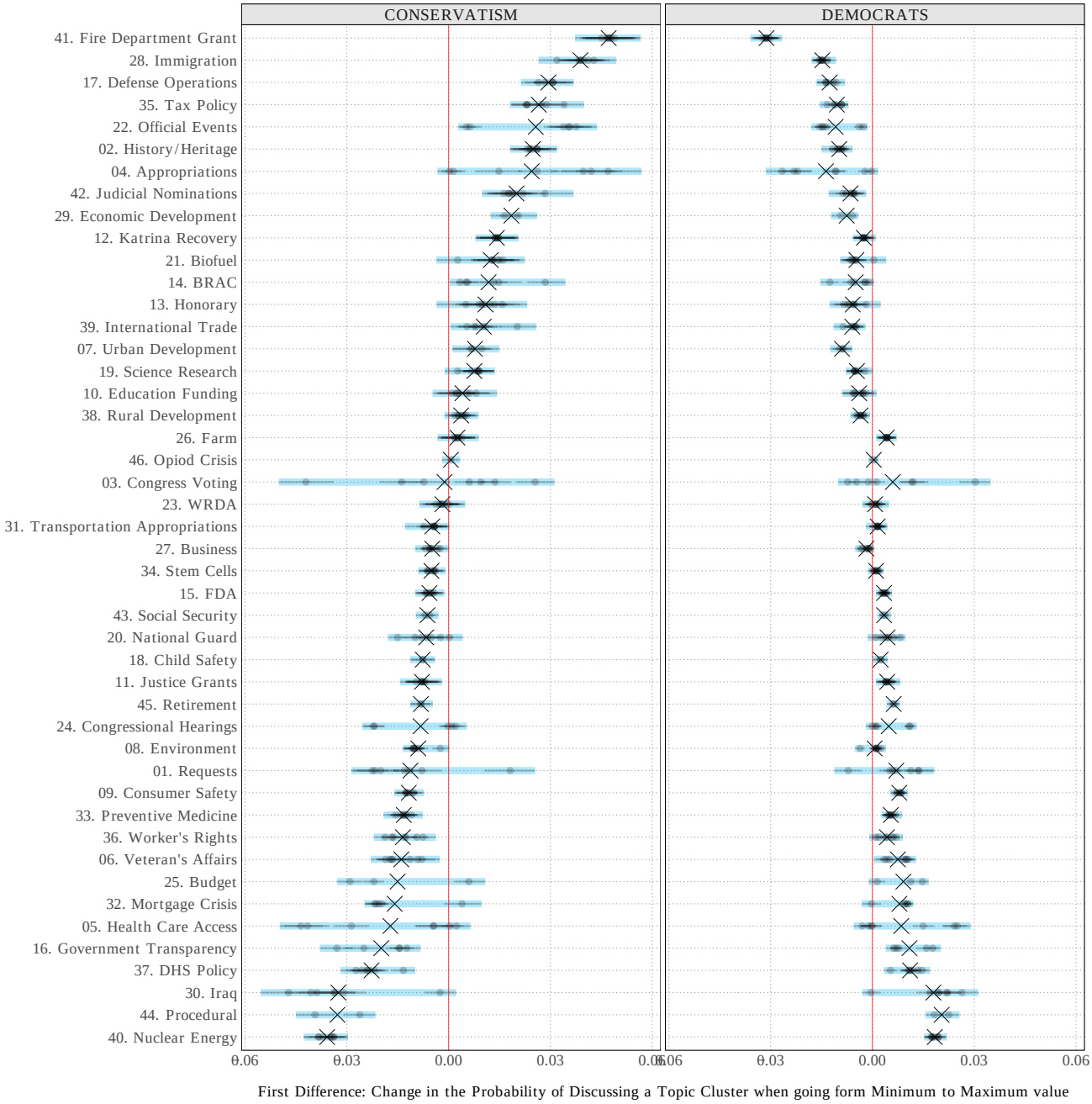


Figure 5 reports the differences in predicted topic attention for the seven different es-

timations (based on the 7 different topic models). The parameters in this case have been standardized by calculating first differences: the estimated effect on the outcome of going from the minimum to the maximum value of the covariate (DW-NOMINATE or Party). The cross indicates the average effect while the dots indicate the (standardized) parameters for each topic model that is part of the cluster for that topic. The lines around them represent 95% bootstrapped confidence intervals.

Once again, the remarkable feature of the figure is how incorporating results from multiple topic models provides additional information about variable effects. If we had not done so, we would not appreciate that we can have more confidence in the effects of ideology and party on messaging about some topics than others.

4.1.3 Topic Instability and Inference

In this final section we illustrate how the method can be used to assess the robustness of covariate effects while controlling for alternative explanations. To do this we replicate some key statistical models in Grimmer (2013). Recall that Grimmer was less interested in the specific topics of senators' press releases than in broader groupings of topics: credit claiming *versus* position-taking press releases. Following Grimmer (2013), we first manual label each discovered topic for whether it is an example of credit-claiming or position-taking. We then calculate the proportion of documents that each Senator dedicated to position-taking and credit-claiming in each of the three years of the study, *using the results from each of the seven topic models.*

Our 44 topic model uses the same data but the results are somewhat different because he estimates the four effects simultaneously while we do not. 32 of our 44 topics are virtually that same as the one's he discovers. For the remaining 12 we assigned our own labels based on the most predictive words (see Appendix A for a list of the topic clusters, their grouping assignments, and whether they match Grimmer (2013)'s original topics). We then

create three author-level outcomes: the proportion of press releases each senator dedicates to credit claiming (y_{credit}); position taking ($y_{position}$); and the differences between the two ($y_{balance}$). Finally, we fit seven linear models predicting each outcome as a function of author-level covariates, and then average the resulting parameters.[10] We use the same author-level covariates from the original study. These include: *Alignment* (higher values indicate that a Senator's district has a larger proportion of copartisans based on presidential vote share);[11] *Years/100* (the senator's tenure in the institution); *Democrat* (or a Republican); *Former House Representative*; *Freshman*; *Majority* party; *In Cycle* (up for election); *State Population in Millions*.

Figure 6 presents three sets of results for three models correspondng to those found in Grimmer's analysis. The red estimates (+ 95% confidence intervals) are the results reported in Grimmer's original study that is based on a single topic model. The black estimates come from our results and are based the results of a single topic model ($k = 43$). We present these results as a baseline of comparison. The 43 topic model produces the *Alignment* estimate that is closest to the one in the original study.[12] Finally, the blue estimates are based on estimates drawn from seven different topic models (as discussed earlier, where the dot represents the average effect).

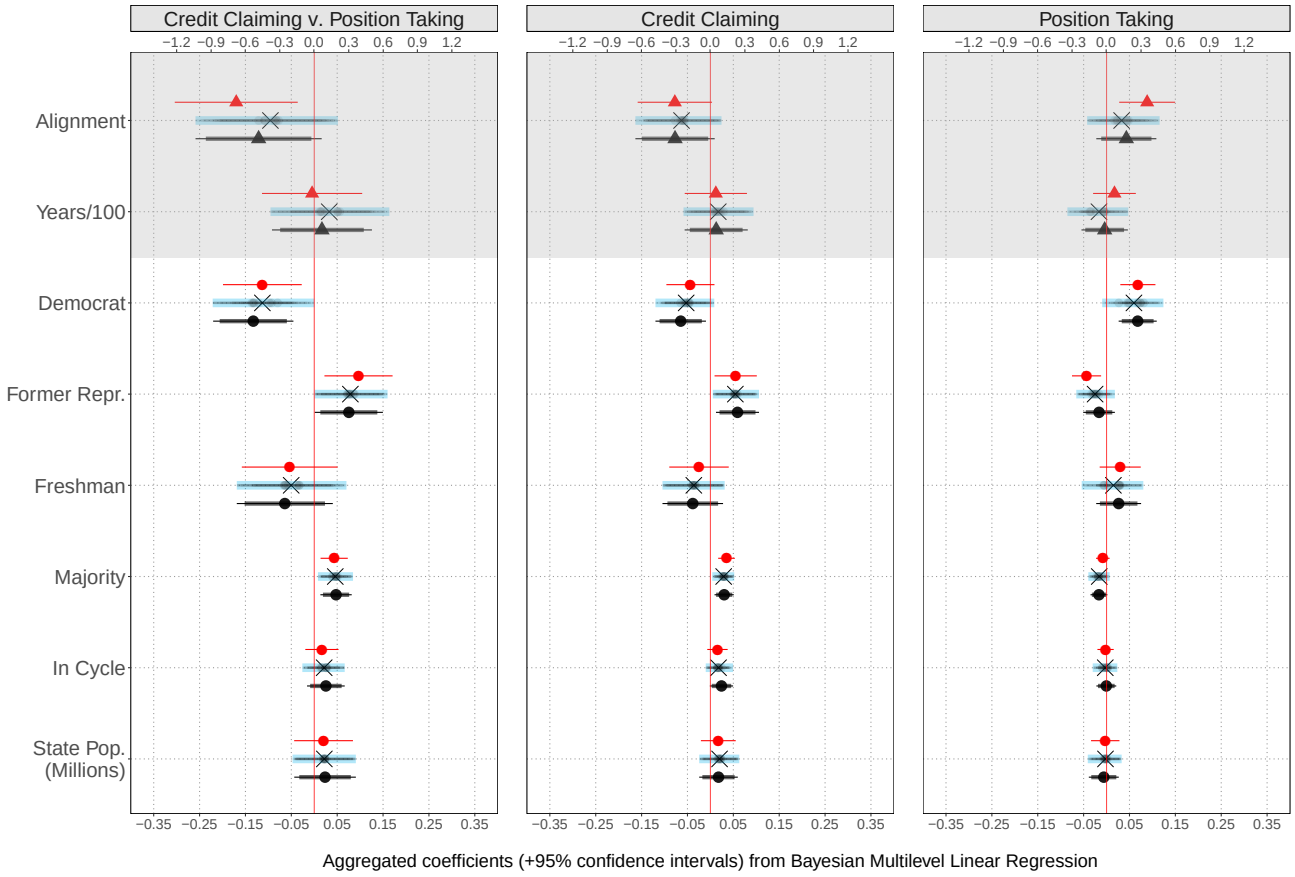
The two top covariates in the figure are much larger than the others, and are represented on a different scale (the top axis ranged from -1.2 to 1.2 whereas the bottom axis ranges from -.35 to .35). Several key points stand out in Figure 6. First, our point estimates (in black) are very similar to the estimates of the original study (red triangles and dots): they are always in the same direction and of very similar magnitude. The estimate that turns to be the most different is the *Alignment* covariate in the *Position Taking* model (which of course carries over to the *Credit Claiming v. Position Taking* model). This could be simply a

[10] $y = \frac{\sum_{j=1}^{Z}(X\delta_j + \epsilon_j)}{Z}$

[11] This is the key electoral competitiveness variable in the original study

[12] The *Alignment* covariate is the key explanatory variable in that article.

Figure 6: Comparing our results to Grimmer (2013)'s original findings



function of our original topic model being slightly different than Grimmer (2013)'s. However, the black *Alignment* estimates indicate that at least one of our alternative models finds thw *Alignment* score to be negatively related to *Credit Claiming* at the .1 level of confidence (first and second models from the left in Figure 6).

Overall, we are able to confirm many of the findings of the original study, *Former House Representative* and members of the *Majority* are more likely to focus their communications on *Credit Claiming*. However, our results seems to question the robustness of the core finding of the study - that Senators from electorally competitive states are more likely to focus on credit claiming in their press releases.

5 Discussion

A rapidly growing literature in political science uses unsupervised topic modeling techniques to test hypothesis about how political actors and the public distribute topic attention. When using unsupervised models, researchers must make modeling decisions that can have important implications for their findings. We focus on one of these decisions - choosing the number of topics. Instead of reporting results for a single best model, as is common practice, we recommend incorporating the results of multiple models into the hypothesis testing process.

Our Robust LDA approach starts with a single best model, but then incorporates additional information from nearby models (models with fewer or more topics). After using clustering methods to align topics across these models we replicate each hypothesis test using the result of each model and report average effects confidence intervals corresponding to the extreme bounds of all the models tested. Applying this approach to Grimmer's seminal study we are able to confirm that many of his findings are robust. However, we did not find this to be the case for his key independent variable - state electoral competitiveness.

The advantage of the method is its simplicity and flexibility. We decided in this paper to address the instability of a particular type of unsupervised topic model, LDA, and to calculate topic similarity using cosine similarity. However, the logic of the method can also be adapted for other types of unsupervised algorithms such as STM and other well-known similarity methods such as Euclidian or Manhattan distance.

Researchers must still make *ex ante* decisions that can impact their findings, such as choosing a similarity threshold or the similarity algorithm used to cluster topics. Nevertheless, we believe that this approach represents a substantial improvement over the current practice within political science of reporting results based on a single best model selected by the researcher.

References

Bagozzi, Benjamin E. and Daniel Berliner. The politics of scrutiny in human rights monitoring: Evidence from structural topic models of us state department human rights reports. *Political Science Research and Methods*, page 117, 2016.

Barbera, Pablo, Andreu Casas, Jonathan Nagler, Patrick Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. 2018.

Berliner, Daniel, Benjamin E. Bagozzi, and Brian Palmer-Rubin. What information do citizens want? evidence from one million information requests in mexico. *World Development*, 109:222 – 235, 2018.

Blei, David M, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Boussalis, Constantine and Travis G. Coan. Text-mining the signals of climate change doubt. *Global Environmental Change*, 36:89 – 100, 2016.

Casas, Andreu and John D. Wilkerson. A Delicate Balance: Party Branding During the 2013 Government Shutdown. *American Politics Research*, pages 1–23, 2017.

Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, 2009.

Chuang, Jason, Margaret E. Roberts, Brandon M. Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. Topiccheck: Interactive alignment for assessing topic model stability. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 175–184, 01 2015.

Collingwood, Loren and John Wilkerson. Tradeoffs in accuracy and efficiency in supervised learning methods. *Journal of Information Technology & Politics*, 9(3):298–318, 2012.

Denny, Matthew J. and Arthur Spirling. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168189, 2018.

Farrell, Justin. Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences*, 113(1):92–97, 2016.

Grimmer, Justin. Appropriators not position takers: the distorting effects of electoral incentives on congressional representation. *American Journal of Political Science*, 2013.

Grimmer, Justin and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, (3):267–297, jan 2013.

Jacobi, Carina, Wouter van Atteveldt, and Kasper Welbers. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1):89–106, 2016.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2):254277, 2015.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228, 2010.

Roberts, Margaret, Brandon Stewart, and Dustin Tingley. *Navigating the Local Modes of Big Data: The Case of Topic Models*. Cambridge University Press, New York, 2016.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58 (4):1064–1082, 2013.

Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, (4):1–8, 2009.

Wilkerson, John D. and Andreu Casas. Large-scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science*, pages 1–18, 2016.

Workman, Samuel. *The Dynamics of Bureaucracy in the US Government: How Congress and Federal Agencies Process Information and Solve Problems*. Cambridge University Press, 2015.

Appendix A The discovered topic clusters, their topic type, and whether they match topics in Grimmer (2013) original model

* The *Match* column indicates whether the topics discovered in our original model m_O with $k = 44$ match the topics discovered by the 44-topic model in Grimmer (2013)'s study.

#	Label	Top Predictive Features	Type	Match
1	Requests	letter, concern, urg, request, secretari, ask	other	
2	History/Heritage	land, park, area, forest, preserv, histor	other	Yes
3	Congress Voting	vote, congress, provis, hous, bipartisan, support	other	
4	Appropriations	hous, fund, appropri, approv, million, confer	credit	
5	Health Care Access	health, care, insur, american, coverag, cost	position	Yes
6	Veteran's Affairs	veteran, care, health, servic, medic, hospit	other	Yes
7	Urban Development	000, fund, project, center, citi, develop	credit	
8	Environment	climat, chang, global, emiss, coast, ocean	position	Yes
9	Consumer Safety	compani, consum, price, product, market, meth	position	Yes
10	Education Funding	school, educ, student, program, colleg, high	credit	Yes
11	Justice Grants	enforc, crime, justic, program, polic, violenc	credit	Yes
12	Katrina Recovery	disast, hurrican, assist, katrina, fema, damag	other	Yes
13	Honorary	honor, american, world, famili, serv, countri	other	Yes
14	BRAC	commiss, recommend, deleg, base, brac, congression	other	Yes
15	FDA	drug, medicar, senior, prescript, plan, fda	position	Yes
16	Gov. Transp.	inform, account, protect, data, intellig, contract	position	Yes
17	Defense Operations	defens, air, militari, base, oper, million	credit	Yes
18	Child Safety	children, program, famili, child, parent, start	other	Yes
19	Science Research	million, technolog, research, develop, program, fund	credit	Yes
20	National Guard	militari, guard, troop, soldier, defens, women	other	Yes
21	Biofuel	energi, oil, fuel, renew, increas, ethanol	other	Yes
22	Official Events	visit, island, center, citi, tour, west	other	
23	WRDA	water, project, river, lake, corp, million	credit	Yes
24	Cong. Hearings	hear, depart, investig, hous, white, offici	other	
25	Budget	fund, budget, million, billion, program, cut	position	Yes
26	Farm	agricultur, farm, farmer, food, program, produc	other	Yes
27	Business	busi, small, low, home, incom, assist	position	
28	Immigration	border, immigr, secur, illeg, reform, enforc	position	Yes
29	Economic Dev.	north, south, main, econom, region, develop	other	
30	Iraq	iraq, war, iraqi, troop, secur, american	position	Yes
31	Transp. Approp.	transport, million, transit, road, rail, improv	credit	Yes
32	Mortgage Crisis	loan, rate, cost, increas, famili, financi	position	Yes
33	Preventive Medicine	cancer, diseas, health, prevent, aid, women	other	Yes

34	Stem Cells	research, cell, stem, human, institut, diseas	position	Yes
35	Tax Policy	tax, credit, taxpay, deduct, relief, incom	position	Yes
36	Worker's Rights	worker, job, employe, employ, wage, labor	other	Yes
37	DHS Policy	secur, airport, homeland, port, terrorist, attack	other	Yes
38	Rural Development	servic, rural, access, area, internet, amtrak	other	
39	International Trade	trade, world, countri, foreign, intern, china	position	
40	Nuclear Energy	safeti, nuclear, epa, protect, standard, plant	position	
41	Fire Dept. Grant	grant, depart, program, firefight, fund, award	credit	Yes
42	Judicial Nom.	court, judg, nomin, justic, nomine, confirm	other	Yes
43	Social Security	secur, social, secretari, travel, depart, passport	position	Yes
44	Procedural	american, time, democrat, want, just, republican	other	

45	Retirement	secur, social, inform, benefit, plan, retir	position	
46	Opiod Crisis	drug, meth, product, consum, enforc, methamphetamin	position	