Budget Justifications and Relational Topic Modeling

Devin Judge-Lord

May 3, 2017

[Early Draft for Workshop Participants][1]

This paper introduces a new dataset of over 100,000 pages of discussion about how federal budget line items ought to be used. With these texts, I hope to test theories about the role of the executive branch in setting the policy agenda for congressional appropriations, the role of Congress in setting the agenda of executive agencies, and how congressional attention relates to budget outlays.

I propose a relational topic modeling approach. The key difference between this and most prior applications is that my central unit of analysis is itself the relationship between two texts, which raises unique methodological challenges. I implement a basic version of this method that measures the distribution of issue attention across the portions of texts that were added and subtracted each year from agency and committee budget justifications. More advanced versions may include text reuse algorithms and the information contained in verb tense, sentiment, or citations. These can enter at the text preprocessing stage or be built into a more sophisticated topic model.

1 Introduction

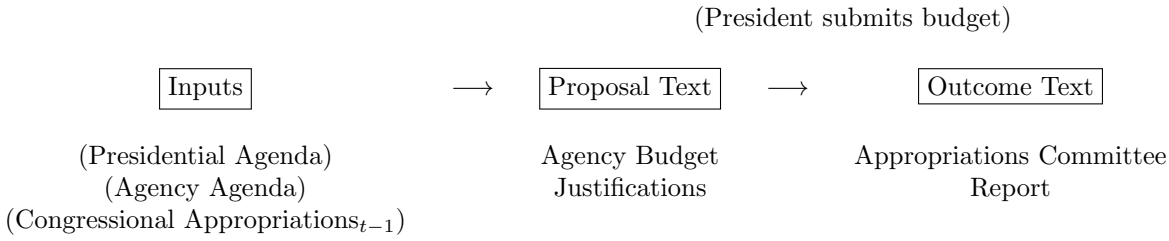
On July 1, 2016, the Director of the Obama Administration's Office of Management and Budget circulated a short memo to all federal agencies: "The FY 2018 Budget will be submitted by the next President, and agencies are not required to submit a formal budget request in September" (OMB 2016). The allocation of federal funds is widely recognized as an important part of politics and has been well studied in political science. Equally important, but much more difficult to study is what government agencies do with these resources. The short answer is that government does too many things to properly catalog and many specifics are unknown at the time the budget is proposed. Nevertheless, one place to start is to look at what the executive branch says it will do with the funds requested and what appropriations committees in Congress say the funds are for. Who drives this more qualitative aspect of budgeting and what can it tell us about executive and legislative power?

How budgets are justified matters. From the executive side, what agencies highlight in their budget requests can be seen as a statement of presidential or agency priorities, which may or may not align.

[1] In this draft I outline the approach but do not attempt to estimate interpretable models. Intermediate results are presented for illustrative purposes. Thanks to Ellie Powell and Alex Tahk for generous guidance.

On the legislative side, we may expect that what an agency proposes to do with its budget affects appropriation of funds by Congress. Completing the circuit, agencies may expect that the detailed appropriations committee reports that accompany their budget are what Congress expects them to do. I visualize these intuitions in Figure 1 and formalize them in Hypotheses 1.1 and 1.2 below.

Figure 1: The Textual Record of US Federal Budgeting



In this paper I introduce a new dataset of the budget justification texts published by agencies and appropriations committees. These data include ten years of budget justifications from 70 federal agencies falling under the authority of three departments (the US Department of Agriculture, Department of Health and Human Services, and Department of the Interior) and one independent agency (the Environmental Protection Agency). These agencies constitute the vast bulk of the jurisdiction of two Senate and two House appropriations subcommittees (the Agriculture, Rural Development, Food and Drug Administration, and Related Agencies Subcommittees and the Interior, Environment, and Related Agencies Subcommittees). Thus, for each year 2008-2017 there are roughly 71 agency documents and exactly 4 appropriations committee reports. Documents range from several dozen to over one thousand pages, most being between 100 and 600 pages. The result is well over 100,000 pages of discussion about how federal funds ought to be used.

These data may shed new light on at least four things of interest to political scientists: (1) the extent to which agencies or the president set the congressional agenda in budgeting, (2) how responsive agencies are to Congress, (3) how much presidential transitions, party control, and committee membership matter for policy content, and (4) whether congressional attention (or lack thereof) to an issue or agency is a good or bad sign for the size of corresponding appropriations.

2 Theory

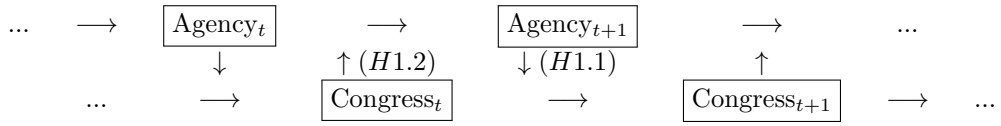
I hypothesize that what the executive branch asks for in discretionary budgeting affects what Congress appropriates and that, in turn, agencies respond to congressional priorities. We know that agencies and

their allies lobby Congress effectively (Carpenter 2014, 2001), but budgeting and committee budget reports are also seen as mechanisms of congressional control (Bolton and Thrower 2015, McCubbins and Schwartz 1984, Yackee and Yackee 2009). We also know that the attention of policymakers over any set of issues is limited (Jones and Baumgartner 2005). A cursory reading of appropriations committee reports aligns with this scholarship. They appear to be partially paraphrasing agencies' own justifications for their budget requests, partially constraining budgetary discretion, and partially copied from the previous year's report. I thus expect these congressional budget reports to discuss a topic more or less when agencies do so first. I also expect agencies to respond to what Congress told them to do the previous year. Figure 2 presents these two agenda setting hypotheses.

Much attention has been paid to the agenda setting power of the president's budget (Berry, Burden and Howell 2010, Brady, Neihesel and Stout 2016, Fisher 2015, Whittington and Carpenter 2003, Wildavsky 1964). Like the above quote from the former Office of Management and Budget Director, this literature leads us to expect that budgets will highlight different priorities under new administrations. Yet the budget blueprints that originate in the White House are sparse. Agencies prepare their budget justifications and the Office of Management and Budget reviews and amends them to the extent the White House is attentive and able to amend. Starting this process after the January inauguration rather than September of the previous year means a shorter window of time to change the text of budget justifications. Additionally, lame duck presidents continue to wield power and may focus on things like administrative policy texts (Howell and Mayer 2005). Given this constraint, I expect new administrations to be much more likely to modify the numbers in a budget than the descriptive details of what the funds are for.

I hope to test these three agenda setting hypotheses: (1.1) appropriations committees tend to shift emphasis from year to year in the same direction as agency budget justifications, (1.2) agency budget justifications tend to shift their emphasis in the direction that the appropriations committee did the previous year, and (1.3) presidential transitions have a significantly greater effect on the magnitude of budget numbers than what agencies propose to do with the funds they are allocated. A stronger version of (1.3) would be that agency budget justification texts do not change significantly due to presidential transitions.

Figure 2: Agenda Setting



For agency j and committee c each discussing topic τ with proportion $\pi_{j,k}$ and $\pi_{c,k}$ at time t:[2]

$$\begin{aligned} \textit{Hypothesis } 1.1: &\quad \pi_{c,\tau,t_1-t_2} \sim \pi_{j,\tau,t_1-t_2} \\ 1.2: &\quad \pi_{j,\tau,t_2-t_3} \sim \pi_{c,\tau,t_1-t_2} \\ 1.3: &\quad \textit{Budget Proposed}_{j,\tau,t_1-t_2} > \pi_{j,\tau,t_1-t_2} \end{aligned}$$

The second class of theories I aim to test relates to the outcomes of politics of attention. When an appropriations committee does attend to an issue significantly more or less than an agency, does this bode well or poorly for the funding of this part of the president's budget? Political science literature points in both directions. Theories that focus on positive agenda control find that coalitions compete for the attention of policymakers, suggesting that attention is often good for their issue (Baumgartner and Jones 1991, Jones and Baumgartner 2005, Kingdon 1995).[3] Conversely, principal agent theories of Congress and the bureaucracy (e.g. "fire-alarm-control") suggest that attention often means sanction, opposition, or constraint (Bolton and Thrower 2015, McCubbins and Schwartz 1984).

Scholarship on Congressional committees and appropriations has focused on party agendas and the politics of attention.

While the act of passing a budget may be seen as a fairly non-ideological sign of party competence (Butler and Powell 2014, Lee 2016), the content of a budget report likely reflects party and committee member agendas (Adler and Lapinski 1997, Lee 2000, Shepsle and Weingast 1987). Even if achieving the party agenda is seen as a general sign of competence (Cox and McCubbins 2005), the content of that agenda is ideological and we expect committees to pay more attention to issues they care about. Yet (Berry, Burden and Howell 2010) find that some program budgets are affected by partisan control in Congress and others are not. Importantly, like Lee (2000) they do not find that committee membership drives program spending toward their districts. These new data allow new tests of the extent to which committee chairs and partisan control matter by looking at the text added, deleted, and edited on issues where the old and new committee members disagree. We may see partisan control effects, committee chair effects, both, or neither. I expect committees to attend to issues raised in agency justifications of

[2]Agency budget justification texts are indexed to the year t that the corresponding appropriations committee report is published, i.e. the year the budget is passed and the year before the fiscal year it funds. Agency budget justifications are published between 9 and 18 months before the fiscal year begins. Appropriations reports are published prior to the budget going to the floor, generally 2 to 6 months before the fiscal year begins.

[3]Groups may also organize to advocate for a part of the budget to be cut or to restrict what funds can be used for, but most scholarship points to examples of groups organizing to push policymakers to attend to problems that require more spending on their issue rather than less. Jones and Baumgartner (2005) find that attention is more punctuated around budget and policy expansion than retrenchment. In the specific case of budget justifications, it seems unlikely that agencies frequently advocate for smaller budgets.

the president's budget proportionally to the chair's ideological alignment with the executive branch on each issue.

Corollary to the politics of attention is the politics of inattention. Members of Congress lack the time to read many bills (Curry 2015), much less thousands of pages of budget justifications from administrative agencies. It may not be surprising that a substantial amount of text that appropriations committees send to the floor is copied either from the previous budget report or agency justifications. The stable core of discretionary budgeting likely represents non-salient issues that are simply ignored due to the limits of information processing (Jones and Baumgartner 2005) or issues of broad agreement (Adler and Wilkerson 2012, Lowi 1967). Additionally, industry groups target relevant committees members of both parties with campaign contributions in order to get their attention to industry-specific problems and redirect government spending in their direction (Powell and Grimmer 2016). The stable core of congressional budget justifications that I observe could thus represent low-salience issues, bipartisan issues, and issues on which committee members of both parties are consistently captured by industry.

In contrast, scholarship on Congress and the bureaucracy has focused on budgeting, and appropriations committee budget reports specifically, as a mechanism of sanction and constraint on delegation.

Wildavsky (1964) proposed a model of the budgeting process relating to how much agencies ask for, how much Congress allocates, and the strategies actors employ to get what they want. More rigorous modeling in the law and economics tradition takes seriously that bureaucrats may have political agendas as well as resource objectives. McCubbins, Noll and Weingast (1987) offer a framework with two general types of control: oversight and administrative procedures. Budgeting is seen a form of oversight . Yackee and Yackee (2009) use the discretionary share of budget as a measure of the strength of an agency's relationships with elected officials. Bolton and Thrower (2015) use the length of committee budget reports as a measure of how much Congress constrains different agencies.[4] Bendor, Glazer and Hammond (2001) suggest that rational principals will delegate to agents with similar goals, repeated interactions, and when they are able to overcome commitment and information problems.

[4]Bolton and Thrower (2015) use the length of these committee reports relative to the size of the budget to measure how much Congress constrains agencies. Using this as a measure of executive discretion, they find that Congress gives greater discretion to ideologically aligned presidents. However, my finding that committee budget reports are highly stable from year to year (even more stable in length than in content), suggests that this is primarily driven by more more funding going to issues on which Congress and the president agree while the length of budget reports stays the same. By looking at how the words of these reports change rather than just their length, we may be able to better measure the relationship between budget appropriations and these texts. Furthermore, as Fisher (2015) notes, committees have been know to sanction agencies for using funds for purposes not specified in their budget justification. This aligns with my intuition that sometimes additional attention from Congress is beneficial (e.g. when an agency adds to the scope of its request).

By measuring the textual alignment on specific topics, we can know when a committee is expanding and when it is constraining what an agency can do with its budget, giving scholars a broadly useful new measure of the discretion granted to executive agencies by Congress. For example, studies of bureaucratic policymaking could benefit from this policy-specific measure of congressional opposition.

The repeated interactions of annual budget cycles has been a core subject of principal agent scholarship. This scholarship assumes that budget authority is a form of discretion and that accompanying texts like appropriations committee budget reports are a form of constraint, but this relationship remains largely untested. Importantly, because of the focus on text as means of constraint, the assumption in this literature is that attention from Congress in these texts bodes poorly for agency budgets.

Drawing on these two literatures, I consider two general types of attention: good attention and bad attention. I hypothesize that good attention (attention associated with increased budgets) is more likely when political allies control the appropriations committee. Bad attention (associated decreased budgets) is more likely when political opponents control the appropriations committee. Combining these intuitions, I expect the magnitude and direction of budget change on an issue to correlate with the interaction of ideological agreement and the difference between the how the agency discusses the topic and how the committees discuss the topic. This means that no significant change in how a topic is discussed suggests no disproportionate change in the budget on that issue.

For agency i and committee c each discussing topic τ with proportion $\pi_{j,k}$ and $\pi_{c,k}$ at time t:

$$Hypothesis\ 2: \quad Budget\ Allocated_{j\tau,t_1-t_2} \sim Ideological\ Alignment_{c,j} * (\pi_{c,\tau,t_1-t_2} - \pi_{j,\tau,t_1-t_2}),$$

The dominant referents for ideological alignment are partisanship and liberal-conservative ideology scores. To assess alignment, scholars have coded agencies as liberal, conservative, or neutral as well as more nuanced scores based on surveys (Clinton and Lewis 2008). Taking a different approach, Benoit and Herzog (2015) estimate ideological positions of lawmakers using transcripts of budget debates, which could be extended to include budget justification texts and testimony.

The next section describes the modeling approach, section 4 presents intermediate results, and section 5 discusses next steps.

3 Modeling Changes in Justifications: A Relational Approach

Like Grimmer (2013), I attend to the "quality" aspects of what government officials do. Votes and budgets may be more easily quantified, but discourse and qualitative policy content is also a key policy output (Mansbridge 2003). Policy texts give meaning to line items and votes. In the case of budget justifications, this meaning is especially interesting because it can be exactly matched with numeric

budget allocations.

This is the first study of which I am aware to look systematically at budget justifications across multiple agencies and committees over time and the first to analyze the text of either. Scholars have applied text analysis methods to similar questions, though none with exactly the same methodological challenges presented by budget justification documents. The general task is to identify similarities and differences between groups of texts (in my case, those written by agencies and by congressional committees). The quantitative approach best suited to this task is the *Latent Dirichlet Allocation* (LDA) model (Blei, Ng and Jordan 2003), specifically a used in the Structural Topic Model (Roberts et al. N.d.). This a mixture model, meaning that documents are assumed to be a mixture of topics, rather than assuming that each document comes from one only one topic (Grimmer 2010).

My methodological contribution is to combine topic modeling approaches with text reuse methods, allowing scholars to better understand not just what is discussed but the topic distributions of what is being added, cut, copied from others, or otherwise receiving special attention. I call this a relational topic modeling approach. Of course, all topic models focus on the relationship between text, but by making the text units being modeled itself a relationship between texts, my version of STM takes a "difference in difference" (e.g. what was added or deleted) or "difference in similarity" (e.g. what was copied) form. This paper focuses on the case where a series of documents retain content from one version to the next, as is common in policy texts. In the discussion, I suggest additional applications of a relational topic modeling approach.

Question	Approach	Estimates distribution of
What topics are discussed?	LDA Topic Model	Words over T topics
Who is discussing what?	Structural Topic Model	J document types over T topics
What is copied from where?	Text Reuse	Tokens in document j copied from j'

This section first reviews the Blei, Ng and Jordan (2003) *Latent Dirichlet Allocation* (LDA) model, then the unique text preparation and effect estimation steps necessary to address my questions, and finally additional steps and extensions to improve topic and effect estimation. In the discussion section, I suggest additional questions that may be addressed using the text analysis approach advanced here.

3.1 LDA: Estimation of the Distribution of Words over Topics

In LDA, each word in a document is assigned to exactly one topic and each document is represented as a vector of topic proportions, i.e. what fraction of the words in that document belong to each topic (Blei, Ng and Jordan 2003). For example, in a model of the Environmental Protection Agency's budget

justifications, "climate," "adaptation," "carbon," and a dozen other words may co-occur and indicate a topic about climate change. The words, "clean", "air," and "health" may also co-occur and have relatively high frequencies in a topic that seems to be about air quality. Each document, representing an agency or subcommittee in a single year, would have a π proportion of words belonging to the *climate change* topic ($\%\tau_{Climate} = \pi_{Climate}$). This may be a relatively high portion for Environmental Protection Agency documents and a low portion for the House Appropriations Subcommittee on the Environment, Interior, and Related Agencies ($\pi_{Climate,EPA} > \pi_{Climate,House}$) compared to the air quality topic which may be more equal.

Importantly, these proportions vary in each document type from year to year. This offers a new way to capture what Jones and Baumgartner (2005) call *attention allocation*, the change weights on policy images and issues: in this case, what the Environmental Protection Agency ought to do.

Figure 3: The Latent Dirichlet Model (LDA)

Text:	President's Budget	$\longrightarrow$ $H1.1$	Appropriations Committee
Dist. over T topics (e.g. $\tau_{1:3}$):	Agency$_j$+ $\sim\{\%\tau_1, \%\tau_2, \%\tau_3\}$ Agency$_j$, $\sim\{\%\tau_1, \%\tau_2, \%\tau_3\}$		Senate$_c$+ $\sim\{\%\tau_1, \%\tau_2, \%\tau_3\}$ Senate$_c$− $\sim\{\%\tau_1, \%\tau_2, \%\tau_3\}$ House$_c$+ $\sim\{\%\tau_1, \%\tau_2, \%\tau_3\}$ House$_c$− $\sim\{\%\tau_1, \%\tau_2, \%\tau_3\}$

More formally, the percent of each topic τ within each document j is estimated as $\pi_{j,\tau}$ where:

$$\tau_{i,j}|W_{i,j} \sim Multinomial(\pi_{w_{i,j}}) \tag{1}$$

$$\pi_j \sim Dirichlet(\alpha) \tag{2}$$

$$W_{i,j} \sim Multinomial(\rho_{\tau,w}) \tag{3}$$

$$\rho_{\tau,w} \sim Dirichlet(\beta) \tag{4}$$

We observe the total number of unique words $(w_1, ..., w_W)$ in the vocabulary of all documents and $w_{i,j}$ is the word observed at the ith token in document j. All texts are "tokenized" by giving each word[5] a unique index i. If token i belongs to topic τ, then the probability that the token is word w is the topic-specific probability $\pi_{\tau,w}$. At the document level, $\pi_{\tau,j}$ is the estimated proportion of topic τ for document j.

[5]For topic estimation, I use single words, but tokenizing may be done by sentence or by any n-gram string of characters or words.

T, α, and β are defined. T is the number of topics $(\tau_1, ...\tau_T)$ where $\tau_{i,j}$ is the topic of the ith token in document j. Each token comes from exactly one topic. α is the parameter of the prior on the per-document topic distributions, and β is the parameter of the prior on the per-topic word distributions. $\rho_{\tau,w}$ is the distribution over w words in each topic τ, i.e. the probability of drawing the wth word of the vocabulary for topic τ.

Figures 5-9 show credible intervals based on the posterior distribution of π.

3.2 Selecting and Preprocessing Texts

To measure the relationship between how agency budget justifications and committee reports change from year to year I employ several extra preprocessing steps.

As each committee oversees more than one agency, the first step was to select the relevant portions of committee reports. I select all pages that cite the agency's name or abbreviation.[6]

Budget justifications repeat about 25-50% of their text verbatim from year to year and more if lightly edited text is included. In one respect, this fact is interesting, as is the content that remains stable in budget justifications. This remarkable stability likely reflects inertia, limits on political attention, and areas of broad agreement. With respect to identifying agenda setting, however, we are also interested in what changes and who is driving these changes. If a topic model is estimated on the full budget justification texts, the stable core of these texts leads models to attribute almost all of the variation to the type of text and none to the year.

I focus on year-to-year change by selecting only the text that was added or subtracted. This can be thought of as a versioning problem where the agency updates their budget justification each year and then the appropriations committee updates theirs. To focus on what changed, I excluded sentences that appear verbatim in the previous year's version. I do this by tokenizing documents by sentence, marking sentence tokens that also appear in the previous version, and retokenizing by word, retaining the information about which word tokens are from copied sentences.[7] This allows me to estimate an

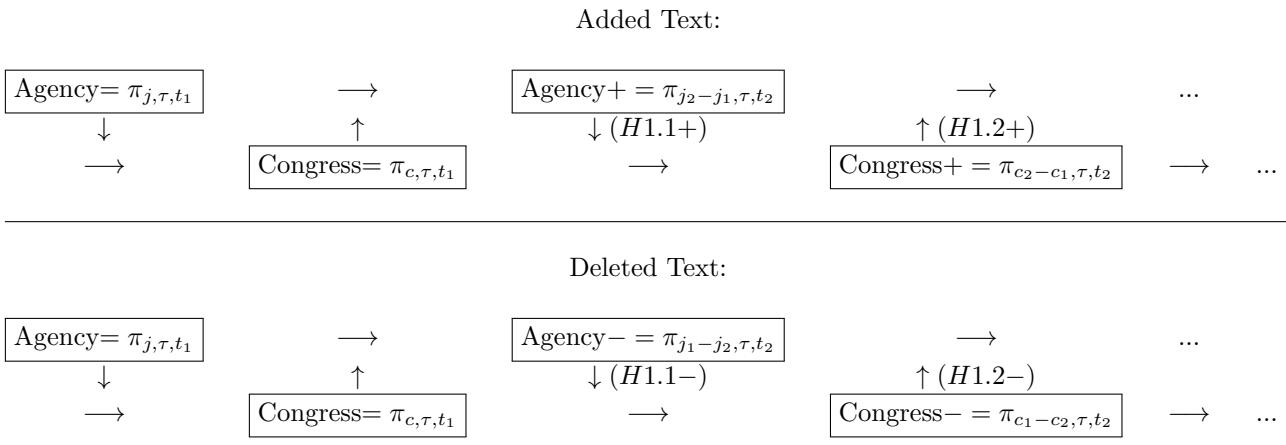
[6] This is a simplistic use of citation information, which is a potentially powerful class of attributes for identifying the relationship between texts. For example, citations may identify more meaningful relationships between texts and text fragments in court cases (Cross et al. 2010).

[7] By only identifying sentences that match verbatim, this approach fails to account for content copied with minor changes. In the future, I hope to improve this by using the Smith-Waterman alignment algorithm (developed for identifying DNA matches and commonly used in plagiarism software) to identified sections of text that are close matches. Wilkerson, Smith and Stramp (2015) successfully employs this approach to identify content copied from various bills in the legislative processes leading to the Affordable Care Act.

I also tested an alternative approach, identifying changes by word instead of by sentence. Given that most commonly used words appear in any given 100 page document, this approach effectively selected extremely rare or new words which,

LDA topic model (identifying which words belong to which topics) with only non-copied text and then to identify distributions of topics among the portions of each text that were added and those portions that were deleted.

Figure 4: Agenda Setting: Text Added and Subtracted



After dividing the text corpus into copied, added, and deleted sentences, I perform additional pre-processing steps, now tokenizing by word. This included removing symbols and numbers, words that appeared in less than 10 percent of the documents, and common filler words with no topical meaning.[8] I also "stem" words, removing suffixes such as -s, es, -ed, and -ing. so that conjugations all contribute to the frequency of the same root concept rather than treating them as separate.

3.3 Assumptions

I do not parse budget justification text by program or sub-agency. Most documents are over 100 pages and could be parsed into smaller units by section headings or page number, but finding consistent parsing methods for inconsistent texts would be difficult and does not have clear value for topic estimation. It may, however, be helpful in the future to parse down to finer levels of bureaucracy to increase the number of budget line item subtotal observations.

I assume topics have the same word distributions from 2008 to 2017. Rather than trying to measure change the content of topics or the number of documents assigned to a single topic, I assume that the content of topics is the same over my ten-year time period and measure the distribution of these topics in agency and congressional texts over time. This is reasonable because this time period is too short for

while doing a poor job of capturing broad budgetary priorities, may have applications for identifying new programs, court cases, or technologies.

[8]The R package *pretext* (Denny and Spirling 2017) provides estimates of the effect of these and other preprocessing decisions on topic model results.

major linguistic change. If linguistic change is occurring, this would likely show up as a slow shift in topic proportion across the time period and not significantly affect year-to-year differences. Thus, there is no need to estimate topics dynamically as developed by Blei and Lafferty (2006) and Brookhart and Tahk (2015).

3.3.1 Post-hoc Analysis of Variation in Topic Distribution

To assess my hypotheses, I need to know if changes in how agencies justify their budgets are associated with changes in how the appropriations committee justifies it and if this relates to changes in budget allocation.

The most straightforward approach to identifying alignment between what agencies and appropriations committees discuss is to first estimate the topics naively (i.e. without attention to what we already know about documents). Given the topics, we know the proportion of words in each document from each topic (recall Figure 3). A time series model can then be used to estimate if changes in these proportions are correlated between two document types, for example the change in a proportion of a topic for an agency justification and appropriation committee report. However, this approach neglects information that could be used to improve topic estimation. Knowing which agency or committee published the document can inform our prior expectations about the distribution of words over topics.

3.4 Improvements Upon LDA

While the Structural Topic Models (STMs) have great potential to estimate differences in policy content across types of texts, they have not been widely applied in this area. STMs are most commonly used to estimate treatment effects on open-ended survey responses (Fong and Grimmer 2016, Mildenberger and Tingley 2015, Roberts et al. 2014). More similar to my approach, Bagozzi and Berliner (2016) use an STM to examine attention to different issues over time in State Department Reports. While their data allow them use the whole text, the versioning nature of budget justifications require the additional preprocessing discussed above.

Instead of covariates only being used post-hoc (estimating effects *after* naively estimating topics), STMs bring the information contained in covariates into the topic model by (1) assigning unique priors by covariate value, (2) allowing topics to be correlated, (3) allowing word use within a topic to vary by covariate values. For STM, instead of $\pi \sim Dirichlet(\alpha)$, topic proportions can be influenced by

covariates X through a regression model, $\pi \sim LogisticNormal(X\gamma, \Sigma)$. This helps the model avoid having to develop a categorization scheme from scratch (Grimmer and King 2011) and improves the consistency of estimated covariate effects Roberts et al. (2014).

In my case, it is likely that budget reports for each agency and each committee are generated by a very similar process. Thus I will give each agency and committee a unique prior rather than assuming that all documents arise from the same distribution of words. The rate of use of each word in a topic is allowed to vary by the agency or committee who wrote it.

Instead of estimating added, deleted, and copied text as separate documents, I aim to bring this information into the model. For example tokens could be marked as old, new, or copied. If old (from the previous year's document) or copied (from another document type), they could be required to be assigned to the same topic. Alternatively, change over time and issue adoption from other document types could be estimated by having the topic for each token be drawn with some estimated probability from the distribution of words in the previous version and of words in other document types.

3.5 Selecting Topics (or the lack thereof)

Topic models generally require one to set the number of topics. There are a number of approaches to topic selection, including algorithms[9], but none are objective (Roberts et al. N.d.). Furthermore, with respect to budgeting Jones and Baumgartner (2005) find that no single explanation can account for U.S. public expenditures in all policy areas. Selecting topics requires attention to the histories of the policy area and there may be no simple interpretation of topic model results. I plan to select topics based on what is more interpretable, especially with respect to budget line items, and then report results with a wide variety of other topic choices in an appendix. This conforms with the principle that selecting the number of topics based on what offers the most sensible interpretation is acceptable as long as no other choice contradicts the conclusions.

In this draft, I present models with 6 topics for each agency (Figures 5-9). It is unlikely that this number of topics best fits the policy debate and thus effects are likely muddled.

[9]Models can also be estimated using the methods described in Lee and Mimno (2014), which selects the number of topics based on t-distributed stochastic neighbor embedding.

4 Intermediate Results

For the purposes of getting feedback on this working paper, Figures 5-9 offer visualization to illustrate my approach. These figures can be seen as intermediate results or summary statistics of the key part of the model. They show the distribution of topics over time by agency and committee with respect to what each agency and committee is adding and subtracting from the previous year.

Having estimated topics, I examine how the topics are distributed across agencies and Congress over time. Each topic has a distribution over words and each document has a distribution over topics. Aggregating, I estimate the expected proportion of topic in a given document type, for example house appropriation committee reports. Specifically, I extract interaction effects (i.e. the expected proportion of each topic) for a given author in a given year. The result is an estimate and credible intervals for how much more or less each actor discusses each topic each year.

To assess hypothesis 1.1, I will estimate the correlation between agency and committee topic proportion means in the same period. To assess 1.2 I will estimate the agency mean as a function of the lag of the committee mean. To assess 1.3 I will compare the magnitude of change in presidential transition years to non presidential years and the difference in difference with respect to budgetary change in presidential years.

To assess Hypothesis 2, I will examine the difference in difference with respect to agency and committee topic distribution and agency and committee budget numbers.

5 Discussion

5.1 Next Steps

There are three primary next steps: (1) gather budget number data (2) test hypotheses with a time-series model and (3) improve how I capture text reuse.

Detailed proposed and granted budget authority numbers are available in, respectively, agency budget justifications and appropriations bills. The specific numbers needed depend on whether my unit of analysis is the issue, program, or the agency. If I focus on issues, I could measure budgetary change as an average proposed and granted budget authority weighted by the portion of an agency's texts on that topic. Alternatively, I could attempt to parse line items within each agency's budget by topic. If I focus on

Figure 5: Proportion of Environmental Protection Agency Budget Justification Emphasizing on Each of Six Topics 2008-2017. Dotted vertical lines indicate a new House Appropriations Chair. Topics are labeled by the six most frequent and exclusive words as identified by the FREX algorithm. Shaded region is the 95% credible interval. Congressional texts are Appropriations Subcommittee Budget Justification Report pages that contain the agency's name or abbreviation.

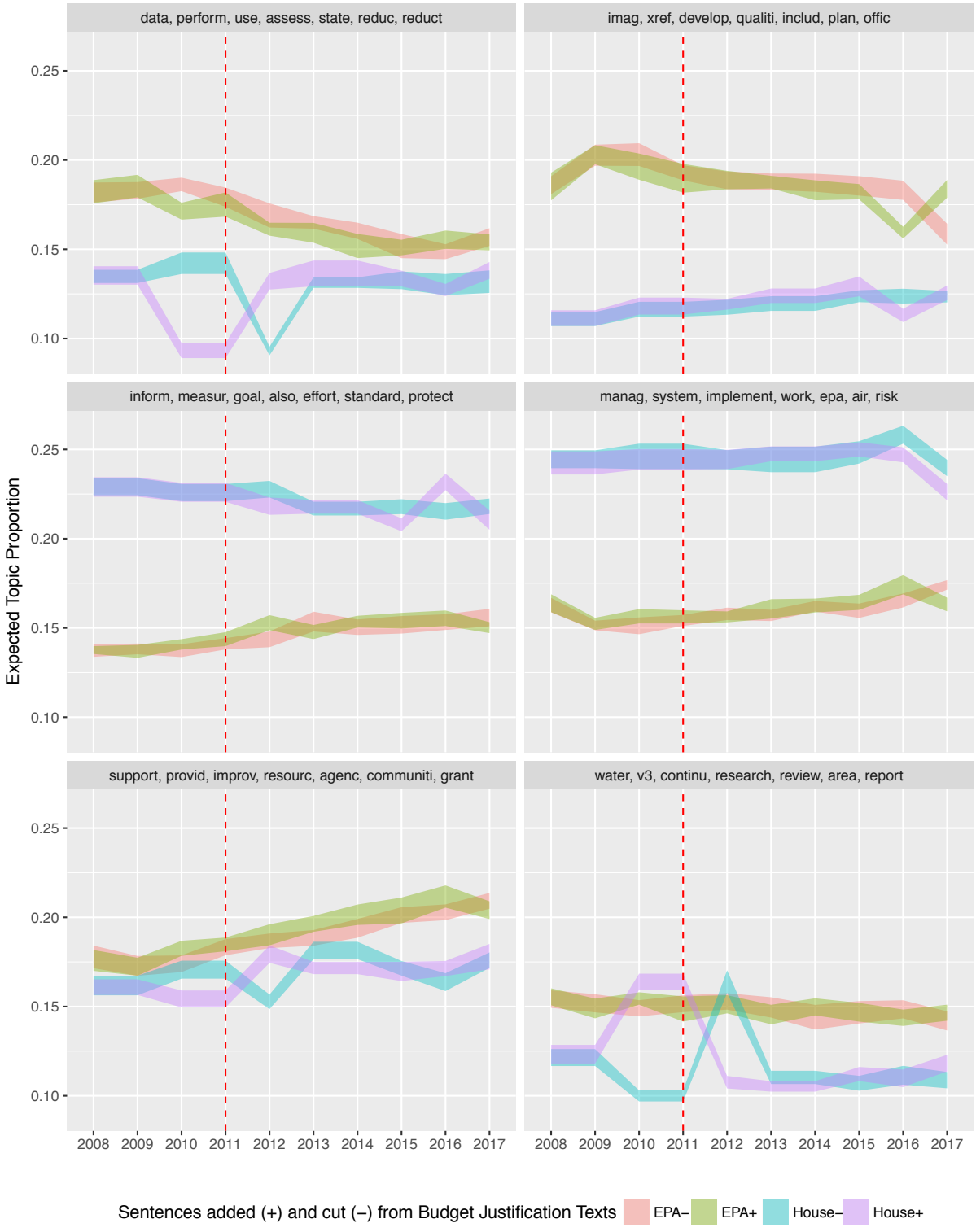


Figure 6: Proportion of Forest Service Budget Justification Emphasizing on Each of Six Topics 2008-2017. Dotted vertical lines indicate a new House Appropriations Chair. Topics are labeled by the six most frequent and exclusive words as identified by the FREX algorithm. Shaded region is the 95% credible interval. Congressional texts are Appropriations Subcommittee Budget Justification Report pages that contain the agency's name or abbreviation.

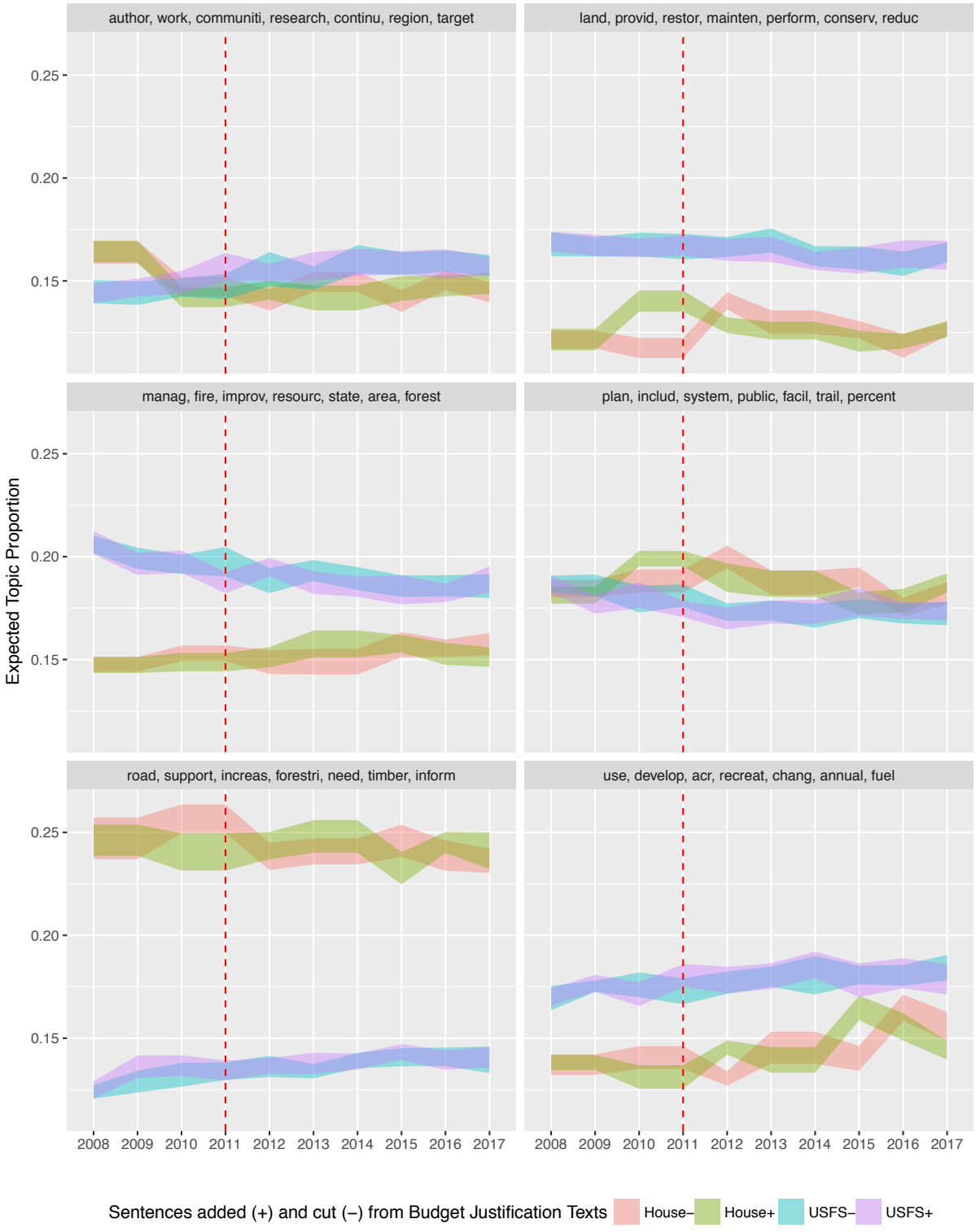


Figure 7: Proportion of Department of Interior Budget Justifications Emphasizing on Each of Six Topics 2008-2017. Dotted vertical lines indicate a new House Appropriations Chair. Topics are labeled by the six most frequent and exclusive words as identified by the FREX algorithm. Shaded region is the 95% credible interval. Lines are portion means. Congressional texts are Appropriations Subcommittee budget justification report pages that contain the agency's name or abbreviation.

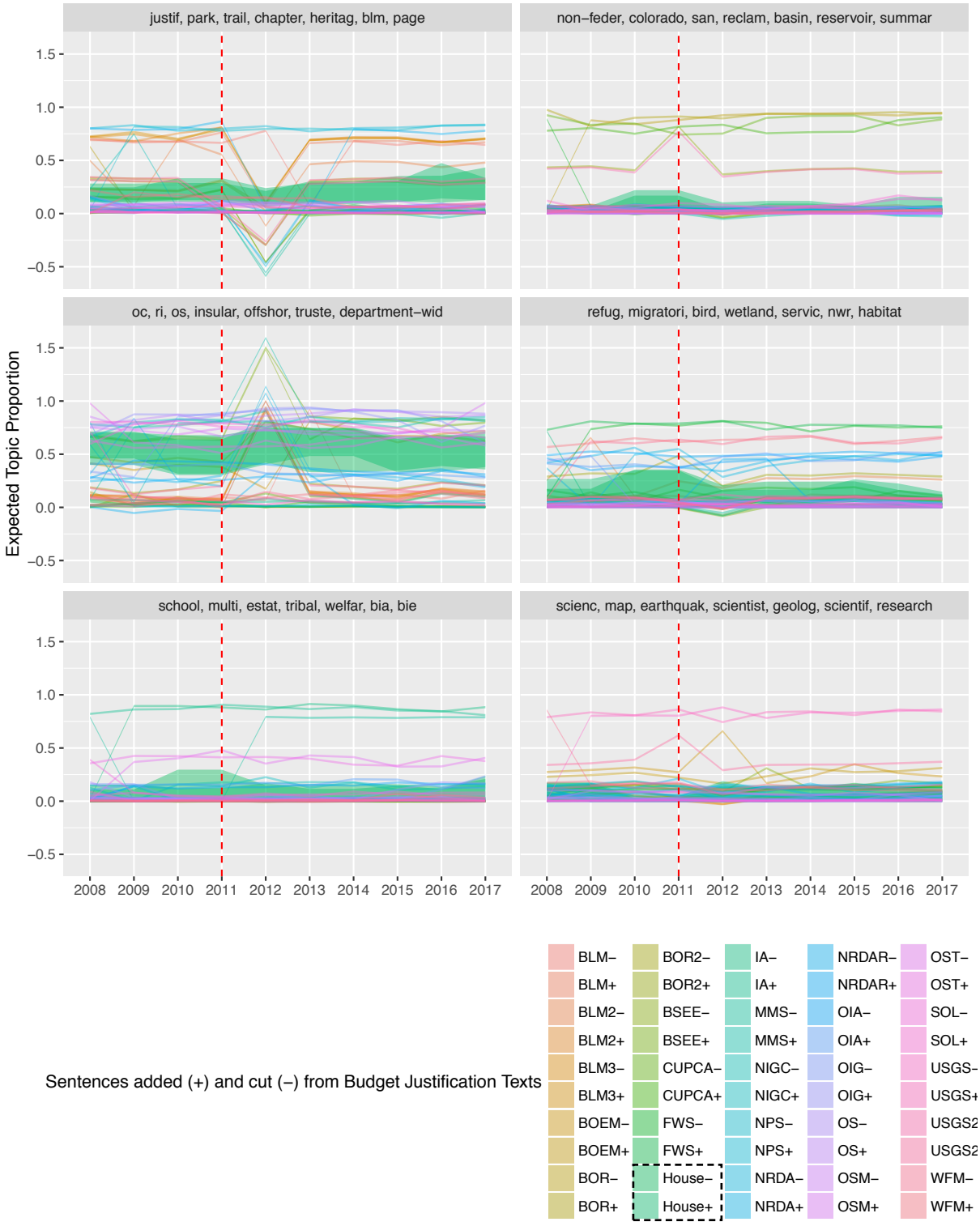
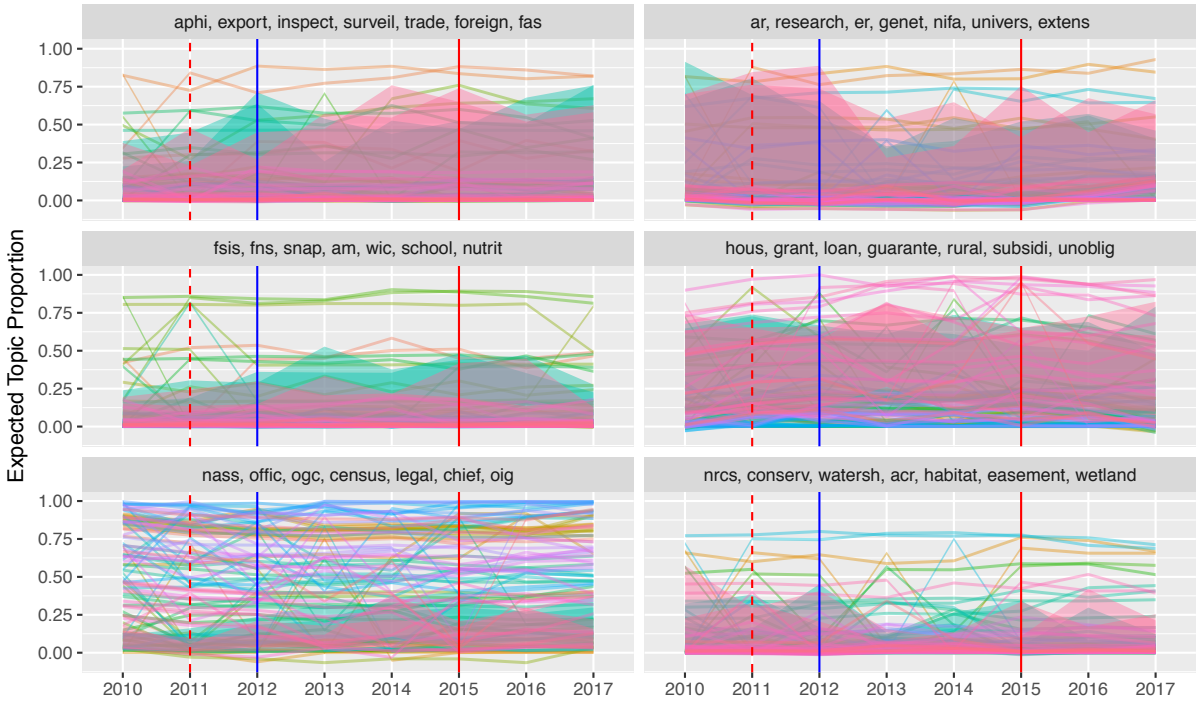


Figure 8: Proportion of Food and Drug Administration Budget Justifications Emphasizing on Each of Six Topics 2010-2017. Dotted vertical lines indicate a new House Appropriations Chair. Solid lines are a new Senate Appropriations Chair. Topics are labeled by the six most frequent and exclusive words as identified by the FREX algorithm. Shaded region is the 95% credible interval. Lines are portion means. Congressional texts are Appropriations Subcommittee budget justification report pages that contain the agency's name or abbreviation.

Figure 9: Proportion of Department of Agriculture Budget Justifications Emphasizing on Each of Six Topics 2010-2017. Dotted vertical lines indicate a new House Appropriations Chair. Solid lines are a new Senate Appropriations Chair. Topics are labeled by the six most frequent and exclusive words as identified by the FREX algorithm. Shaded region is the 95% credible interval. Lines are portion means. Congressional texts are Appropriations Subcommittee budget justification report pages that contain the agency's name or abbreviation.



agencies or programs as the unit of analysis, measuring budgetary change is clear, and topic change could be the average of topic change weighted by the average portion of each topic in that agency's budget justification texts.

The most straightforward way to test my hypothesis would be to take topic model outputs and run a separate time series model. A better way may be to build into the topic model parameters for the relationship between topic change in agency texts and appropriations committee texts and include budget numbers as covariates.

Here, text reuse is measured by nearly verbatim sentence matches and only measured from one year to the next within the same document type. I plan to improve text reuse detection by using Smith-Waterman alignment algorithm that identifies matching text fragments despite minor editing. This will improve the identification of what was copied, what was deleted, and what is new from one year to the next. More importantly, it will be able to detect fragments copied across document types which is too rarely exactly verbatim to be measured with my current approach.

Two additional innovations could improve this type of analysis:

First, texts could be parsed by verb-tense to identify sentences that are discussing an agency's past work from sentences discussing future plans. Budget justification texts are a mix of both types of speech, but we may expect closer agreement between agencies on Congress on how funds were spent in the past than how they are to be spent in the future.

Second, modeling the relationship between texts could be informed by sentiment analysis. Dictionaries that code words as positive, negative, or neutral may help us discriminate between good and bad attention. This can be done in at least four ways: Texts could be subset or sorted into positivity, negativity, or neutral components and modeled as separate observations. Alternatively, sentiment analysis could be applied after topic modeling to measure the extent to which each document tends to use the words in each topic in a positive or negative light. Perhaps the best approach would be to build sentiment analysis into the model. This could be down with either a lexical prior that topics will be either positive or negative type or by estimating a parameter for the sentiment of each document on each topic.

5.2 Broader Applications

Combining classification algorithms like topic models with matching algorithms that identify text reuse has broad potential in political science. To illustrate how this approach be adapted different questions in

political science, I briefly discuss two such potential applications: administrative rulemaking and court cases.

The Administrative Procedures Act often requires administrative agencies to solicit public comments on proposed regulations. Rich data on several decades of rulemaking are available but have yet to be fully utilized by scholars. Agencies publish draft rules, and thousands of comments received by interest groups, experts, and citizens. This offers leverage to identify the players, winners, and losers and to track those participating in the policy process over time.

A relational topic modeling approach could be used model the distribution of issues over comments compared to the change from draft to final rule. Text reuse methods could also detect specific changes in rule text attributable to certain comments. The topic distribution of modified or copied rule text would indicate the issues to which agencies attend and the type of commenter raising those issues.

High-profile court cases, especially those at the Supreme Court are receiving a growing number of *amicus curiae* briefs from third parties. These briefs often include carefully crafted suggestions for judges to consider and political scientists are interested the role briefs may play in court decisions and opinions (Collins 2008, Corley, Collins and Calvin 2011, Kearney, Merrill and Kearnet 2000). Like budget justifications, the content of Supreme Court opinions matters. Justices devote significant time negotiating over the content of majority opinions (Wahlbeck, Spriggs and Maltzman 1998). Political and private actors in society look to opinions to determine what actions are legal (Spriggs and Hansford 2001). "[S]cholars, practitioners, lower court judges, bureaucrats, and the public closely analyze judicial opinions, dissecting their content in an endeavor to understand the doctrinal development of the law" (Corley et al., 2011, pg. 31).

A relational topic modeling approach could be used to model the distribution of issues over court opinions and briefs. Unlike budget justifications and rulemaking, draft opinions are not written before briefs are received. Nevertheless, a Structural Topic Model could be used to identify coalitions and text fragments reused in opinions can identify the most influential members of winning coalitions. In the case of court opinions, citations to a brief represent a special kind of text reuse that indicates a special level of attention (Cross et al. 2010). Topic distributions could be estimated for cited text, text copied from petitioner, respondent, and *amicus curiae* briefs.

Words give meaning to political ideas. The methods advanced here offer scholars tools to model relationships across policy texts to identify what is flying under the radar, what is copied from elsewhere or otherwise receiving special attention, and ultimately who is driving the substance of policy change.

References

Adler, E. Scott and John D. Wilkerson. 2012. Congress and the Politics of Problem Solving. In *Part 1*. pp. 3–18.

Adler, ES and JS Lapinski. 1997. "Demand-side theory and congressional committee composition: A constituency characteristics approach." *American Journal of Political Science* 41(3):895–918.

Bagozzi, Benjamin E. and Daniel Berliner. 2016. "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* pp. 1–17.

Baumgartner, F R and B D Jones. 1991. "Agenda Dynamics and Policy Subsystems." *Journal of Politics* 53(4):1044–1074.

Bendor, Jonathan, Amihai Glazer and Thomas Hammond. 2001. "Theories of delegation." *Annual review of political science* 4(1):235–269.

Benoit, Kenneth and Alexander Herzog. 2015. "Text Analysis: Estimating Policy Preferences From Written and Spoken Words.".

Berry, Christopher R, Barry C. Burden and William G Howell. 2010. "The President and the Distribution of Federal Spending." *American Political Science Review* 104(04):783–799.

Blei, D, AY Ng and MI Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3(1):993–1022.

Blei, David M and John D Lafferty. 2006. "Dynamic Topic Models." *International Conference on Machine Learning* pp. 113–120.

Bolton, Alex and Sharece Thrower. 2015. "The Constraining Power of the Purse: Executive Discretion and Legislative Appropriations *.".

Brady, Jacob, Robert Neihesel and Kevin Richard Stout. 2016. "Stimulating Presidential Support: The American Recovery and Reinvestment Act, Presidential Pork, and Vote-buying in Congress.".

Brookhart, Jennifer L and Alexander Tahk. 2015. "The Origin of Ideas Department of Political Science.".

Butler, Daniel M. and Eleanor Neff Powell. 2014. "Understanding the Party Brand: Experimental Evidence on the Role of Valence." *The Journal of Politics* 76(02):492–505.

Carpenter, Daniel. 2014. *Reputation and power: organizational image and pharmaceutical regulation at the FDA*. Princeton University Press.

Carpenter, Daniel P. 2001. *The forging of bureaucratic autonomy: Reputations, networks, and policy innovation in executive agencies, 1862-1928*. Princeton University Press.

Clinton, Joshua D. and David E. Lewis. 2008. "Expert Opinion, Agency Characteristics, and Agency Preferences." *Political Analysis* 16(01):3–20.

Collins, Paul M. 2008. *Friends of the Supreme Court : interest groups and judicial decision making*. Oxford University Press.

Corley, Pamela C., Paul M. Collins and Bryan Calvin. 2011. "Lower Court Influence on U.S. Supreme Court Opinion Content." *The Journal of Politics* 73(01):31–44.

Cox, Gary W. and Mathew Daniel McCubbins. 2005. *Setting the Agenda: Responsible Party Government in the U.S. House of Representatives*. Cambridge [England] ; New York, N.Y.: Cambridge University Press.

Cross, Frank B, James F Spriggs II, Timothy R Johnson and Paul J Wahlbeck. 2010. "Citations in the Supreme Court: An Empirical Study of their Use and Significance." *University of Illinois Law Review* pp. 489–575.

Curry, James M. 2015. *Legislating in the dark : information and power in the House of Representatives*.

Denny, Matthew J. and Arthur Spirling. 2017. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It.".

Fisher, Louis. 2015. *Presidential Spending Power*. Princeton, NJ: Princeton University Press.

Fong, Christian and Justin Grimmer. 2016. "Discovery of Treatments from Text Corpora." pp. 1–20.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18:1–35.
URL: *http://web.stanford.edu/ jgrimmer/ExpAgendaFinal.pdf*

Grimmer, Justin. 2013. "Appropriators not position takers: The distorting effects of electoral incentives on congressional representation." *American Journal of Political Science* 57(3):624–642.

Grimmer, Justin and Gary King. 2011. "General purpose computer-assisted clustering and conceptualization." *Proceedings of the National Academy of Sciences of the United States of America* 108(7):2643–2650.

Howell, William G and Kenneth R Mayer. 2005. "The Last One Hundred Days." *Presidential Studies Quarterly* 35(3).
URL: *http://williamghowell.com/wp-content/uploads/2015/02/TheLast100.pdf*

Jones, Bryan D and Frank R Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems*. Chicago, IL: University of Chicago Press.

Kearney, Joseph D, Thomas W Merrill and Joseph D Kearnet. 2000. "The Influence of Amicus Curiae Briefs on the Supreme Court." *University of Pennsylvania Law Review Pa. L. Rev* 148(743).

Kingdon, John W. 1995. *Agendas, alternatives, and public policies*. 2nd ed. New York: HarperCollins College Publishers.

Lee, Frances E. 2000. "Senate Representation and Coalition Building in Distributive Politics." *American Political Science Review* 94(1):59–72.

Lee, Frances E. 2016. *Insecure majorities : Congress and the perpetual campaign*.

Lee, Moontae and David Mimno. 2014. "Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference." *Empirical Methods in Natural Language Processing* pp. 1319–1328.

Lowi, Theodore. 1967. "The Public Philosophy : Interest-Group Liberalism." *The American Political Science Review* 61(1):5–24.

Mansbridge, Jane. 2003. "Rethinking Representation." *American Political Science Review* 97(4).
URL: *http://eurogender.eige.europa.eu/sites/default/files/3593021.pdf*

McCubbins, Mathew D, Roger G Noll and Barry R Weingast. 1987. "Administrative procedures as instruments of political control." *Journal of Law, Economics, & Organization* 3(2):243–277.

McCubbins, Mathew D and Thomas Schwartz. 1984. "Congressional oversight overlooked: police patrols versus fire Alarms." *American Journal of Political Science* 28(1).

Mildenberger, Matto and Dustin Tingley. 2015. "Beliefs about Climate Beliefs: Second-Order Opinions in the Climate Domain." pp. 0–38.

OMB. 2016. "Circular No. A-11 Revised.".
URL: *https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/a11_current_year/a11_2016.pdf*

Powell, Eleanor Neff and Justin Grimmer. 2016. "Money in Exile: Campaign Contributions and Committee Access." *The Journal of Politics* 78(4):974–988.
URL: *http://www.journals.uchicago.edu/doi/10.1086/686615*

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58(4):1064–1082.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley and Edoardo M Airoldi. N.d. "The Structural Topic Model and Applied Social Science *." . Forthcoming.
URL: *https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf*

Shepsle, Kenneth A. and Barry R. Weingast. 1987. "The Institutional Foundations of Committee Power." *American Political Science Review* 81(1):85–104.

Spriggs, James F. and Thomas G. Hansford. 2001. "Explaining the Overruling of U.S. Supreme Court Precedent." *Journal of Politics* 63(4):1091–1111.

Wahlbeck, Paul J., James F. Spriggs and Forrest Maltzman. 1998. "Marshalling the Court: Bargaining and Accommodation on the United States Supreme Court." *American Journal of Political Science* 42(1):294–315.
URL: *http://www.jstor.org/stable/2991757?origin=crossref*

Whittington, Keith E and Daniel P Carpenter. 2003. "Executive power in American institutional development." *Perspectives on Politics* 1(03):495–513.

Wildavsky, Aaron B. 1964. *Politics of the budgetary process*. Boston, MA: Little Brown.

Wilkerson, John, David Smith and Nicholas Stramp. 2015. "Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach." *American Journal of Political Science* 59(4):943–956.

Yackee, J W and S W Yackee. 2009. "Divided government and US federal rulemaking." *Regulation & Governance* 3(2):128–144.