A Class of Bayesian Semiparametric Cluster-Topic Models for Political Texts *

Justin Grimmer † Rachel Shorey ‡ Hanna Wallach § Frances Zlotnick ¶

July 22, 2011

Abstract

We introduce a new Bayesian cluster-topic model for political texts and a novel methodology for model selection in statistical models for text. We first develop a fully parametric model that simultaneously estimates the topics articulated in texts as well as partitions of documents based on their attention to topics. In the context of this model we then outline a new and general methodology for model selection that combines the strengths of both statistical and substance based approaches. First, we extend our fully parametric model to a group of semiparametric models using three nonparametric priors: the Dirichlet process prior, the Pitman Yor process prior, and the uniform process prior. These models *estimate* the number of clusters used in the model, but we show the estimated number of clusters and the number of documents per cluster are heavily model dependent. Therefore, while the statistical guidance in selecting models is essential, human judgment is necessary to make a final model selection. To use human input to make this final selection, we introduce a battery of experimental methods that provide carefully-elicited subject expert evaluations of the models and explain how to extend these methods to new models and data sets. We implement our models and experiments on a new collection of over 19,000 House press releases from 2010. Using our results, we show that ideologically extreme representatives dominate policy debates—a finding with widespread consequences for policy deliberation and lawmaking.

*For helpful discussions and assistance in evaluations we thank Allison Anoll, Tabitha Bonilla, Kyle Dropp, Sasha Foo, Wendy Gross, Tyler Haddow, Mackenzie Israel-Trummel, John Kendall, Pat Kennedy, Greg Martin, and Emily Sydnor. This work was supported in part by the Center for Intelligent Information Retrieval, in part by a fellowship under NSF grant # DGE-0907995 , and in part by NSF grant # SBE-0965436 . Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

†Assistant Professor, Department of Political Science, Stanford University; Encina Hall West 616 Serra St., Palo Alto, CA, 94305

‡Ph.D. Candidate, Department of Computer Science, University of Massachusetts Amherst; 140 Governors Drive, Amherst, MA, 01003

§Assistant Professor, Department of Computer Science, University of Massachusetts Amherst; 140 Governors Drive, Amherst, MA, 01003

¶Ph.D. Student, Department of Political Science, Stanford University; Encina Hall West 616 Serra St., Palo Alto, CA, 94305

1 Introduction

We introduce a new statistical model for analyzing political texts and a new methodology for substance-driven model choice for social science research. Our work contributes a new and general model to the growing literature in political science on the statistical analysis of text (Laver, Benoit and Garry, 2003; Quinn et al., 2010; Hopkins and King, 2010), while also addressing unresolved model selection questions when applying models to large collections of texts. Increasingly, political scientists use statistical models to measure underlying traits from the content of documents. For example, scholars may estimate topics in floor speeches (Quinn et al., 2010) or the presentational styles of legislators in press releases (Grimmer, 2010). Each of the models used to measure the *latent* features requires assumptions about the number of underlying traits: each model requires a choice of *model complexity*. Both statistical and substance-based approaches to model complexity exist, but we show how both types of guidance can lead to poor model choices. We therefore propose a new approach to model selection that unifies statistical and human-based guidance. The result is a general approach to generating valid measures of the content that is useful for *social science* research.

We present our approach to social scientific model selection in the context of a new model for political text. Our model accomplishes two previously distinct tasks (Wallach, 2008, Chp. 5): measuring the topics discussed in a document (Blei, Ng and Jordan, 2003) and clustering documents based on their content (Fraley and Raftery, 2002). To do this, we build a model that simultaneously measures the topics discussed across texts, how much attention each document allocates to those topics, and then partitions documents based on their attention to topics. We show how measuring topics and clusters simultaneously ensures that our model can ignore idiosyncratic language and group together texts that focus on the same political concepts. The result is a model that is a general and useful tool for examining the content of political documents. But our model (and model structure) is also more general and applicable to other substantive problems. For example, the base model we introduce has a structure highly similar to that introduced in Grimmer (2011) to measure legislators' styles. The same model can even be applied to other data, such as roll call votes or survey responses.

We then use this model to present a new methodology for substance-driven model selection. Our

approach combines models that automatically learn model complexity and carefully elicited human judgement through a battery of experiments. To automatically determine model complexity we extend our statistical model to a class of semiparametric models using three Bayesian nonparametric priors: the Dirichlet process prior, the Pitman-Yor process prior, and the uniform process prior. Each of the priors allows us to simultaneously estimate our model and model complexity (Wallach et al., 2010). But emphasizing results from Wallach et al. (2010) we show that the choices about model complexity from each of the models are *necessarily* model dependent. Building on analytical results, we provide practical guidance on how the assumptions in each prior determine the number of clusters the model estimates. If strong *a priori* beliefs exist about the number of clusters in the data set, then this guidance will be sufficient to select the appropriate semiparametric model.

Social scientists, however, rarely have strong beliefs about the number of clusters in their data or the distribution of documents across a fixed number of clusters. This creates the need for a second round of model selection: one that is based on the substantive properties of the models. To perform this selection, we implement a battery of experiments to measure substantive model fit. We both introduce new experiments and build on existing work to measure the appropriateness of any two documents belonging to the same cluster (Grimmer and King, 2011), the coherence of a set of words describing a topic (Chang et al., 2009), the ability of a topic to characterize a document, and the ability of a topic to capture human generated key words. We then provide new graphical tools for comparing model fit across the candidate models, allowing the importance of each model feature to vary. This allows scholars to not only select a final model, but to infer the strengths and weaknesses of their models using substantive criteria.

We apply our statistical model for text and model selection strategy to a new collection of House press releases from 2010, constituting over 19,000 press releases. We show that each of our models provides conceptually valid and substantively interesting models of legislator speech. Then using the estimated topics and clusters from our final selected model, we show that ideologically extreme representatives dominate policy debate resulting in an artificial polarization in policy disputes.

Before proceeding to introduce our model and model selection strategy, we explain why Congressional press releases are an essential component of the representation process.

2 Political Representation and House Press Releases

While political representation is regularly conceptualized as a correspondence between representatives' roll call voting decisions and constituents' opinions, the actual representation process in America is much richer, involving many more activities than casting votes or even how members of Congress engage in the legislative process. The representation process also occurs outside of Congress, as representatives engage constituents in their district and constituents in other districts (Mansbridge, 2003).

Part of this engagement occurs as representatives return to their districts and directly engage their constituents (Fenno, 1978). But representatives also engage constituents through impersonal means. Representatives manage their reputation through carefully-crafted statements and interactions with the media. For example, legislators and their staff author newsletters, conduct press conferences, and participate in televised interviews to disseminate information to constituents. Among these impersonal tools, press releases are among the most valuable because legislators can issue them regularly and precisely control their content. This allows elected officials to carefully state policy positions or to draw attention to grant awards and other money for the district. Press releases are also an important tool because their content is likely to reach constituents. Press releases are regularly run verbatim in local papers and the contents of press releases are regularly used in news stories (Grimmer, 2010).

Press releases are important on their own as a conduit between representatives and constituents. But the content of press releases is valuable because it is a more general indicator of what representatives do in Washington and how representatives debate policy. Grimmer (2011) shows that there is a general correspondence between what senators say in their press releases and how senators invest their time in Washington. And what legislators discuss in their press releases is also a credible indicator of legislators' rhetorical strategies. Press secretaries regularly coordinate what legislators say on the floor and what they insert in newsletters and press releases (Lipinski, 2004).

We analyze all press releases from House members collected by the *US Federal News Service* in 2010. This results in a collection of 19,006 press releases or about 44 press releases issued *per representative*. Press releases are especially useful for analyzing discourse because they come from nearly all representatives and on nearly every day of the year. Further, both Republicans and

Democrats issue press releases at very similar rates: members of both parties use press releases to communicate with their constituents.

2.1 Preprocessing Press Releases

To represent the press releases quantitatively, we use a set of preprocessing steps standard in both political science and computer science (Quinn et al., 2010; Manning et al., 2008). We first discard word order, modeling each document as an exchangeable sequence of words. We then replace all upper case letters with lower case, and remove punctuation. We also remove a proprietary list of words that were either representative-specific or formatting-specific.[1] In addition, we discard the first fifty and last twenty words in each document, as these words tend to be standard official text added on by the representative's office or the publishing entity. Finally, we remove all words that do not occur at least five times in the corpus. Notice, that unlike many other statistical models for text, we *do not* stem the words (Porter, 1980), allowing for a richer set of features than other statistical models of text.

The result of our preprocessing steps is $N = 19,006$ press releases. Each press release i ($i = 1, \ldots, 19,006$) is represented as a sequence of J_i ($j = 1, \ldots, J_i$) word tokens, $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ_i})$. There are 1,939,838 total word tokens across the set of the press releases, with an average of about 102 word tokens per press release. The tokens span 18,836 unique word types.

3 A Parametric Cluster-Topic Model for Political Texts

Using the press releases, we demonstrate our approach to model selection with a hierarchical statistical model for texts. The model simultaneously measures the salient topics in each press release and then partitions press releases based on their attention to topics. In this section we develop a fully parametric version of the cluster-topic model. Our model describes a generative process for political texts. While real world documents are obviously not generated according to this process, by defining a generative process we can then use the machinery of statistical inference to learn about latent variables of interest from the observed text. In Section 4.1 we show how to extend

[1] To identify these words we used standard latent Dirichlet allocation (LDA) with 1,000 topics (Blei, Ng and Jordan, 2003). This approach allowed us to identify topics that were related to the formatting of the articles, information about press secretaries and contact information, and representative-specific information, such as the names of representatives or towns in their districts.

this baseline model to include a variety of nonparametric priors and in Section 4.2 we show how to perform final model selection with carefully elicited human guidance.

We simultaneously measure the topics and clusters of press releases by combining features from topic models and clustering methods into a hierarchical Bayesian model. At the lowest level of hierarchy, we model each text as a mixture of *topics* (Blei, Ng and Jordan, 2003). Substantively, we think of topics as the politically-relevant concepts discussed in the press releases. Formally, each topic is a discrete probability distribution over the $W = 18,836$ unique word types used in the press releases. Words with a high probability of use for a particular topic capture the language that representatives employ when discussing that topic. The key assumption is that there is a shared vocabulary across speakers when discussing each topic and that there is a largely separate (but possibly overlapping) vocabulary across political concepts.

Even though press releases are usually written with one focused point, we model each document as a mixture of multiple topics. From a substantive perspective, using multiple topics per press release ensures that we capture asides, nuances, or the occasional press release that raises several topics. Using multiple topics also helps create a cleaner model. If we were to model each document as falling into a single topic, words that are very common in many press releases, such as "bill", "law" and "vote" will have significant presence in most topics, making the topics look less distinct and perhaps impeding inference. One way to get around this issue would be to remove these domain-specific high frequency words . Such a procedure would be highly subjective and time-consuming, requiring multiple rounds of modeling followed by investigation by a subject expert to make sure all such words were removed. By allowing multiple topics to model each document, we let these extremely common words settle into a few very common topics, leaving the substantive issue topics free from these words. Clear and interpretable topics are extremely important since we perform human analysis tasks on the topics for model selection. For our model, we remove only unambiguous stopwords (i.e., non-content words) such as "the", "and" and "go" from a standard stopword list to speed inference. We also remove representative names and major cities in their districts to prevent the model from creating an individual topic for each representative, since we do not want the clustering portion of the model, described below, to cluster the documents according to these representative-specific topics.

Based upon these document-specific mixtures over topics, at the highest level of the hierarchy we *partition* the press releases into clusters, where press releases in the same cluster have similar topical composition. The clustering of press releases into groups of documents with similar focus is useful both substantively and statistically. Substantively, the clusters identify the salient features of press releases, facilitating interpretations that would be otherwise difficult. This is most useful when attempting to identify press releases associated with major policy conflicts—like the Iraq war or immigration—or associated with pork barrel politics—like airport and fire department grants. Without the clusters of documents, as is standard in topic models, identifying the primary theme of a press release would require arbitrary cutoffs based on each press release's mixture over topics. Statistically, grouping documents into similar clusters ensures that we only borrow information from similar documents during topic inference. This allows for a principled approach to smoothing.

The combination of topic models and clustering also results in more substantively meaningful clusters from the texts than if we performed just clustering on the raw word counts. Topic models facilitate the identification of groups of documents that use slightly different language to make the same substantive point. Topic models are able to do this because they identify groups of words that tend to be used together, though not necessarily all together in any one document. As a result, our model is able to group together two documents focused on the same topic even if slightly different language is used. This is a particularly useful for political documents, especially political documents from Congress. When political actors strategically communicate their views to constituents we might expect partisanship (or other speaker and temporal characteristics) to heavily influence word choice (Monroe, Colaresi and Quinn, 2008). Models based on raw word counts might separate Republican and Democrat speech if it uses a very different vocabulary to discuss the same key ideas. But, as long as strategic communication from Republicans and Democrats shares a basic vocabulary when discussing the same political concept, our model will avoid separating statements from opposing partisans on the same political subject.

To develop the model, we start at the bottom of our hierarchy and describe how we model press releases as a mixture of topics. We then move to the next level of the hierarchy to describe how we simultaneously cluster the press releases.

Topic Models of Documents Using Words We begin with the collection of 19,006 press releases. Suppose that each press release i ($i = 1, \ldots, 19,006$) is a mixture of K ($k = 1, \ldots, K$) topics. Recall that topics measure the rate political speakers use words when discussing a political concept. Call the proportion of press release i allocated to topic k, π_{ik} and let $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK})$ represent press release i's mixture over the K topics. Each document's mixture over topics affects the words that appear in the document by determining the rate words are drawn from the K topics. For the j^{th} word in press release i ($j = 1, \ldots, J_i$), suppose that we represent the topic that generated it with $\boldsymbol{\tau}_{ij}$, a $K \times 1$ indicator vector. Then $\boldsymbol{\tau}_{ij}$ is a draw from a Multinomial distribution,

$$P(\boldsymbol{\tau}_{ij}) \quad = \quad \text{Multinomial}(1, \boldsymbol{\pi}_i).$$

Conditional on the j^{th} word being generated by the k^{th} topic, we then draw the word y_{ij} from a topic-specific Multinomial distribution. Call θ_{kw} the probability of a speaker using word w when discussing topic k and collect this rate across all W words into $\boldsymbol{\theta}_k$. Then we draw y_{ij} from a Multinomial distribution.

$$P(y_{ij} \,|\, \tau_{ijk} = 1, \boldsymbol{\theta}_k) \quad = \quad \text{Multinomial}(1, \boldsymbol{\theta}_k).$$

Clustering of Documents Using Topic Mixtures At the next level of the hierarchy, we partition documents based on their attention to topics. To infer this partition, we model the documents' mixtures over topics using a mixture of hierarchical Dirichlet distributions. Suppose each document i is a member of 1-of-S clusters. We will represent the probability of a document belonging to the s^{th} cluster as β_s such that $\boldsymbol{\beta} = (\beta_1, ..., \beta_S)$. We then represent the cluster responsible for document i with an $S \times 1$ indicator vector $\boldsymbol{\sigma}_i$, drawn from a multinomial distribution,

$$P(\boldsymbol{\sigma}_i \,|\, \boldsymbol{\beta}) \quad = \quad \text{Multinomial}(1, \boldsymbol{\beta}).$$

Conditional on each press release's cluster, we draw $\boldsymbol{\pi}_i$. Each cluster of documents, s is characterized by a Dirichlet distribution with two parameters: a $K \times 1$ vector that describes the average attention across topics for that cluster and a concentration parameter that describes how close the documents' topic compositions are to the cluster average. Let m_{sk} be the probability of generating words for press releases in cluster s from the k^{th} topic and collect this across all K topics in $\boldsymbol{m}_s =$

$(m_{s1}, m_{s2}, \ldots, m_{sK})$. Let α represent a parameter that controls how much draws from the cluster-specific Dirichlet distribution resemble $\boldsymbol{m}_s$: as α increases, the draws will resemble $\boldsymbol{m}_s$ more. Given these parameters and each press release's cluster assignment, we suppose that each $\boldsymbol{\pi}_i$ is drawn from the corresponding cluster-specific Dirichlet distribution,

$$P(\boldsymbol{\pi}_i \,|\, \boldsymbol{m}_s, \alpha) \quad = \quad \text{Dirichlet}(\alpha \boldsymbol{m}_s).$$

We extend the hierarchy by assuming that each $\boldsymbol{m}_s$ is also drawn from a Dirichlet distribution with average attention to topics $\boldsymbol{m}$ and concentration parameter α_1,

$$P(\boldsymbol{m}_s \,|\, \alpha, \boldsymbol{m}) \quad = \quad \text{Dirichlet}(\alpha_1 \boldsymbol{m}).$$

To complete our model specification, we place a symmetric (uniform) Dirichlet priors on both $\boldsymbol{m}$ and $\boldsymbol{\theta}_k$. Let $\mathbf{1}_K$ represent a vector of ones of length K and α_2 and η represent concentration parameters. Then $\boldsymbol{m}$ and $\boldsymbol{\theta}_k$ are assumed to be draws from,

$$\begin{aligned} P(\boldsymbol{m} \,|\, \alpha_2, \mathbf{1}_K) &= \text{Dirichlet}(\alpha_2 \mathbf{1}_K) \\ P(\boldsymbol{\theta}_k \,|\, \eta, \mathbf{1}_W) &= \text{Dirichlet}(\eta \mathbf{1}_W). \end{aligned}$$

4 Model Selection in Statistical Models of Texts

The previous section introduced a novel model for measuring the contents of political texts. The model is straightforward to apply to other substantive problems and is easily extended to non-textual data, like roll call votes. But this model, like many other machine learning methods, requires assumptions about *model complexity*. For our model, we must set the number of topics in the model and the number of clusters. This is analogous to other models that require assumptions about the number of clusters of documents in hierarchical models for political texts (Quinn et al., 2010; Grimmer, 2010), number of clusters in clustering models (Fraley and Raftery, 2002), or the number of dimensions in factor analysis problems (Poole and Rosenthal, 1997; Clinton, Jackman and Rivers, 2004).

In spite of its central role in fitting models for political texts and their substantive interpretation, there is little guidance for determining model complexity when measuring latent features in social science applications. The most regularly-used approaches rely on fit statistics for model selection

that are only loosely connected to the substantive application at hand, such as post-hoc fit statistics (Fraley and Raftery, 2002) or evaluations based on out-of-sample predictions (Blei, Ng and Jordan, 2003). Other models use statistical tools for Bayesian methods, like the Dirichlet process prior, to simultaneously estimate models and perform model selection (Spirling and Quinn, 2011; Gill and Casella, 2009). And still other approaches eschew statistical guidance, instead relying upon a manual search by varying the number of topics across several model fits and evaluating the performance of the model using a substantive criteria (Quinn et al., 2010).

Both statistical and human-guided methods have their virtues—and limitations. Statistical methods for assessing model complexity are necessary because of the incredible number of parameters in statistical models for text. Even the most engaged and diligent researchers are unable to examine the hundreds of thousands of parameters produced by such statistical models. Fit statistics computed after fitting a model simplify this comparison, reducing assessments of fit from hundreds of thousands of parameters to only a few. And Bayesian methods based on nonparametric priors make the comparison even simpler, automatically choosing model complexity.

Statistical guidance from fit statistics and nonparametric priors therefore greatly simplifies model selection and evaluations of fit, but statistical guidance alone can cause researchers to select models that are substantively sub-optimal. The problem with only using statistical guidance to make model selection is most acute with post-hoc fit statistics, like BIC, AIC, or DIC (see Claeskens and Lid (2010) for a review). Each of the methods measures how well the model fits the data, while penalizing for additional model complexity. Each of the post-hoc methods of model fit measures performance based on *internal* criteria: they implicitly assume that the model and data are sufficient to evaluate the models.

To paraphrase Box (1979), statistical models of text are always *wrong*, but some are useful. And therefore, to make a final model determination when using machine learning methods for social scientific tasks, careful human evaluation is necessary. Language is sufficiently rich that simple statistical models are unable to accurately represent the data generating process for text data. From how statistical text models represent text to the assumed data generating process, statistical models for political texts extensively simplify language. To be clear, the simplicity is useful—it is effective at highlighting substantively interesting content in texts. But the implication

of this simplification is that we should view both the models and data used to fit those models with some skepticism. As a result, post-model fit statistics alone are insufficient to determine which model we should use. Rather, the key test is to determine whether the fitted models are useful for our task: the generation of substantively interesting and valid measures of political and social phenomena.

Testing models for how they will be used motivates the use of other statistics in machine learning tasks, but these measures alone are also inappropriate for social scientific purposes. Often, machine learning scholars want statistical models that *predict* well the contents of new documents (Manning et al., 2008). But social scientists are rarely interested in prediction, instead they focus on obtaining *substantive* summaries of texts (Mimno and Blei, 2011).[2] And unfortunately, there is often little connection between how well a model predicts new data and how well it identifies substantively interesting content. Chang et al. (2009) show that a model's performance on predictive tasks is often *negatively* correlated with a model's substantive fit.

We also avoid post-model fit statistics and model prediction for statistical and computational reasons. Post-hoc fit statistics do not vary the prior and implicitly limit the partitions that are considered. As Wallach et al. (2010) show, prior assumptions inform *both* the number of clusters in a data set and *how observations are allocated across clusters.* Even if scholars use post-hoc fit statistics to select a model complexity, they fail to vary how documents are allocated across the number of clusters. Further, post-hoc fit statistics and measures of prediction require fitting many models. For large collections of texts, this can require an impractical amount of computing time.

For social scientists to select models based on their substantive content, therefore, human guidance is needed in addition to the statistical information. But the same problems that motivate the use of statistical models of text make applying this guidance difficult: humans have limited time and cognition when evaluating the content of models. This makes current recommendations on how to use human guidance in model selection difficult to implement (Quinn et al., 2010). The current state of the art encourages users of software to develop impressions of how well the model is able to identify conceptually valid and distinct clusters. Then, the impressions developed across models are used to select a model complexity that provides the most conceptually valid and distinct

[2]This is not to say that social scientists are not interested in prediction or using texts to make predictions.

clusters.

Certainly this advice for model selection addresses some of the limitations of statistical methods for model selection. But there are two limitations with this impressionistic search that we address here with carefully-designed evaluation experiments. First, asking humans to form impressions about the quality of topics creates the opportunity to inject error into the model selection process. Even in one model it can be difficult for a researcher to assess every topic. Asking a user to repeat this task across many models places a significant cognitive burden on the researcher. This is likely to induce satisficing as researchers focus on only a few topics or fail to identify inconsistencies in their model (Krosnick, 1999). Second, statistical models for text do more than create a set of topics. For example, our model also represents each text as a mixture of topics and generates a partition of texts. The other modeling features need substantive evaluation too.

We propose a methodology for model selection that combines the strengths of statistical and human-based approaches to model selection. For statistical guidance, we extend the model developed in Section 3 to include three nonparametric priors: the Dirichlet process prior, the Pitman-Yor process prior, and the uniform process prior. Applying recent analytical results from Bayesian nonparametrics, we show that each of the priors makes different assumptions about the number of clusters in a data set and how documents are divided across those clusters, ensuring that we consider a diverse set of models. Using the models ensures that we have data-driven model choices that vary crucial assumptions, without fitting the large number of models required for post-hoc fit statistics. To select a final model from the fitted semiparametric models, we introduce a battery of experiments that measure model performance across three different features of our model: the grouping of documents into clusters, the identification of topics, and the representation of press releases as mixtures of topics. Our experiments carefully control how we elicit, and then use, human input. We then provide a set of graphical tools that allow users to vary the weight attached to the human evaluations when making a model selection. The result is rich and substance-based evaluations for guiding model selection.

To focus intuition, we develop our methodology for the clustering of documents, while assuming that the number of topics is known. Our methodology can be extended to the case where the topics are also unknown. To perform this extension, the nonparametric prior considered here would be

extended to hierarchical nonparametric priors, as in Teh et al. (2006). Our experiments to elicit human input are also straightforward to extend to new models and even fully parametric models—a point we address in the conclusion.

4.1 Nonparametric Priors for Bayesian Model Selection

The first step in our model selection methodology is to seek statistical guidance from a group of Bayesian semiparametric models. Each model uses a different nonparametric prior to vary the number of estimated clusters of press releases. The result is a set of candidate estimated models that will comprise the models we evaluate with our human experiments.

While each prior makes distinctive and consequential assumptions that lead to different model fits, the priors share common properties. Each prior is defined over *distributions* rather than parameters: in other words, a draw from each prior is a distribution rather than a parameter. Each model is set up so that the distribution drawn from the prior is then used to draw parameters for each observation that then generate the data. Crucially, the distribution drawn from each prior is discrete with probability 1, which implies that observations will share parameters. Two observations sharing the same parameter in the semiparametric models is equivalent to the observations being assigned to the same cluster in our fully parametric model.

The three priors have two basic parameters. Each prior has a *base* measure, G_0: essentially the *expected* distribution drawn from each prior (Orbanz and Teh, 2010). For our model, we will assume that our base measure is a hierarchical Dirichlet prior. Suppose that $\boldsymbol{m}$ is a $K \times 1$ vector of proportions. Then our base measure is defined as $G_0 \equiv \text{Dirichlet}(\alpha_1 \boldsymbol{m})$ and $\boldsymbol{m}$ is drawn from a Dirichlet distribution $P(\boldsymbol{m} \,|\, \alpha_0, \mathbf{1}_K) = \text{Dirichlet}(\alpha_0 \mathbf{1}_K)$.

The three priors also have a parameter or set of parameters that determine how concentrated draws from the prior are around the base distribution. For the purposes of stating our general class of models, we will represent the parameters across models with $\boldsymbol{\zeta}$.

If we generally call the three nonparametric priors $BP(G_0, \boldsymbol{\zeta})$ we can characterize our class of semiparametric cluster-topic models as,

$$
\begin{aligned}
P(G) &= BP(G_0, \boldsymbol{\zeta}) \\
P(\boldsymbol{m}_{s(i)}) &= G \\
P(\boldsymbol{\pi}_i \,|\, \boldsymbol{m}_{s(i)}, \alpha) &= \text{Dirichlet}(\alpha \boldsymbol{m}_{s(i)}) \\
P(\boldsymbol{\tau}_{ij} \,|\, \boldsymbol{\pi}_i) &= \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
P(w_{ij} \,|\, \tau_{ijk} = 1, \boldsymbol{\theta}_k) &= \text{Multinomial}(1, \boldsymbol{\theta}_k) \\
P(\boldsymbol{\theta}_k \,|\, \eta, \mathbf{1}_W) &= \text{Dirichlet}(\eta \mathbf{1}_W)
\end{aligned}
\tag{4.1}
$$

where $s(i)$ is a function that provides each observation's parameter. For example, if observation i is assigned to the 10^{th} parameter, then $s(i) = 10$. Two observations, i and j are assigned to the same cluster if $\boldsymbol{m}_{s(i)} = \boldsymbol{m}_{s(j)}$

Because each prior estimates the number of clusters included in a model, they are often used to determine a *natural* number of clusters to include in a model. But as we emphasize now using previous results on nonparametric priors (Wallach et al., 2010), each model makes specific and stringent assumptions that affects the number of estimated clusters (Welling, 2006). Both the Dirichlet process prior and the Pitman-Yor process prior create a few large clusters and many small clusters, while the uniform process prior creates a more even distribution of documents across clusters. And therefore the number of clusters identified is highly model-dependent.

To compare the nonparametric models, we analyze the predictive probability of a new observation being assigned to an existing cluster or a new cluster. We also give asymptotic descriptions for the expected number of clusters and cluster size. While the results that we present are standard in other fields or introduced in other work, we repeat them here to emphasize that even the use of nonparametric priors to create semiparametric models does not remove the need to make modeling choices about the number of clusters in our model. Rather, the use of nonparametric priors allows for *data driven* choices about the number of clusters.

4.1.1 Dirichlet Process Prior

The Dirichlet process prior is the most commonly used nonparametric prior, with applications in political science, statistics, medicine, and computer science (Gill and Casella, 2009; Spirling and Quinn, 2011; Escobar and West, 1995; Huelsenbeck et al., 2006; Blei and Jordan, 2006) and was originally developed in Ferguson (1973). A single concentration parameter ζ determines how close

draws from the prior are to the base distribution. This parameter also strongly influences the partitions of the data the Dirichlet process prior produces.

To see this, suppose that we have partitioned M documents according to their assigned Dirichlet parameter $\boldsymbol{m}_{s(i)}$ and define N_s as the number of observations assigned to the s^{th} cluster. Then, the probability that observation $M+1$ is assigned to the s^{th} cluster is $\frac{N_s}{M+\zeta}$ and with probability $\frac{\zeta}{M+\zeta}$ it is assigned to a new cluster (Sethuraman, 1994).

This makes clear that the properties of partitions generated by a Dirichlet process prior are strongly model-dependent. The process that generates new clusters results in a few large clusters. This property is often described as the "rich-get-richer" property of the model. Clusters with many observations—with large N_s values—are more likely to have new documents assigned to them. But clusters with only a few observations are less likely to see new documents. It also makes clear that ζ heavily influences the number of clusters identified by the model. Large values of ζ will induce many more clusters, whereas smaller values of ζ will cause the creation of fewer clusters (Sethuraman, 1994).

Given the strong dependence on ζ when determining each document's cluster assignment, it is not surprising that the value of ζ also strongly influences the expected number of clusters and the size of those clusters. As the number of documents, N, approaches infinity the expected number of unique clusters is approximately $\zeta \log(1+\frac{N}{\zeta})$.[3] The asymptotically expected number of clusters of size N_s also depends strongly on ζ. As N approaches infinity, the expected number of clusters of size N_s is $\frac{\zeta}{N_s}$ (Arratia, Barbour and Tavare, 2003; Wallach et al., 2010).

4.1.2 Pitman-Yor Process Prior

The Pitman-Yor process prior is a generalization of the Dirichlet process prior (Pitman and Yor, 1997). Like the Dirichlet process prior, the Pitman-Yor process has a concentration parameter ζ. But the Pitman-Yor process prior is more flexible and includes a *discount* parameter δ ($0 \leq \delta < 1$). Together, ζ and δ influence the cluster assignments.

To see how ζ and δ affect cluster assignments, suppose again that we have partitioned M documents into S unique clusters, with N_s documents assigned to the s^{th} cluster. Then the $M+1^{\text{st}}$

[3] $\zeta \log(1+\frac{N}{\zeta})$ also characterizes the variance in the number of clusters (Teh, 2010)

document is assigned to cluster s with probability $\frac{N_s-\delta}{M+\zeta}$ and to a new cluster with $\frac{\zeta+K\delta}{M+\zeta}$ (Pitman, 2002).

Like the Dirichlet process prior, the Pitman-Yor process will create several large clusters and many small clusters—though the discount parameter allows greater flexibility in the shape of partitions generated using this prior. If $\delta = 0$ then the Pitman-Yor process reduces the Dirichlet process prior. As the discount parameter δ is increased, however, new documents are less likely to be allocated to large clusters and more likely to be allocated to a new cluster.

The asymptotic behavior of cluster number and size from the Pitman-Yor process are strongly dependent on the concentration and discount parameters. As the number of documents approaches infinity, the expected number of unique clusters is $\frac{\Gamma(1+\zeta)}{\delta\Gamma(\delta+\zeta)}N^{\delta}$. And the expected number of clusters of size N_s is given by $\frac{\Gamma(1+\zeta)\prod_{z=1}^{N_s-1}(z-\zeta)}{\Gamma(\delta+\zeta)N_s!}N^{\delta}$ (Pitman, 2002).

4.1.3 Uniform Process Prior

Both the Dirichlet process and Pitman-Yor process prior allow the estimation of the number of the clusters to be included in the model, but the inferred number of clusters is strongly dependent on modeling assumptions. While there are differences across the models, both models create partitions with a few large clusters and many smaller clusters. This may be a reasonable modeling decision for political texts. For example, it may be useful to identify a few issues that many politicians engage regularly, with other issues receiving more scattered attention. But social scientists do not have strong prior beliefs over this structure of political debate. It is just as plausible that dividing documents more evenly across clusters is more useful for social scientific inferences from political texts. And more likely still is that in some instances the uneven distributions are useful and in other instances more even distributions over the clusters are useful.

This suggests the utility in considering nonparametric priors that allow for a more even allocation of documents across the clusters. We consider one such prior here: the uniform process prior. The prior was first introduced in Qin et al. (2003) and Jensen and Liu (2008), and studied extensively in Wallach et al. (2010). Like the Dirichlet process prior, the uniform process prior has a single concentration parameter ζ. But the uniform process prior assigns documents to clusters using a very different, and much more even, process. Suppose that we have allocated M documents

across K clusters. Then the $M+1^{\text{st}}$ observation is assigned to an existing cluster with probability $\frac{1}{K+\zeta}$ and assigned to a new cluster with probability $\frac{\zeta}{K+\zeta}$. That is, documents are assigned to existing clusters with a uniform probability, all together avoiding the "rich-get-richer" property of the Dirichlet process Prior and the Pitman-Yor process prior.

The uniform allocation of documents results in very different asymptotic properties for uniform process priors. Wallach et al. (2010) show that, as the number of documents approach infinity, the uniform process is expected to create approximately $\sqrt{2\zeta N}$ unique clusters, with approximate expected size ζ.

One concern about the uniform process prior is that it is not an *exchangeable* prior. Consider the density $P(c)$ of a complete partitioning c of the data into clusters. $P(c)$ can be computed via the predictive probabilities for each document given the documents that have already been assigned to clusters. In exchangeable priors, the order in which the documents are assumed to arrive is irrelevant. $P(c)$ will be the same no matter the order of document processing. The Dirichlet process and Pitman-Yor process priors both lead to exchangeable partitions as discussed in Wallach et al. (2010). The uniform process does not share this property. This fact may be cause for some concern, since we generally do not think of a collection of documents as being ordered. Even though the uniform process does not guarantee robustness to order statistically by producing an exchangeable clustering, Wallach et al. (2010) show that the model is *experimentally* robust to different document orders.

4.1.4 Summarizing the Model Assumptions in Estimating Model Complexity

We collect the important properties of the priors in Table 1. This table makes clear that, while nonparametric priors allow data-driven estimates of model complexity, this inference is also model-dependent. Using different priors, then, provides different perspectives on how many clusters to include in the model. Our approach for choosing among the models using substantive criteria is described in the next section.

Following the hierarchical model given by Equation 4.1, in Appendix A we provide our sampling algorithms for estimating the models. Building on work in Wallach et al. (2010) and Wallach (2008), we use a series of Gibbs sampling steps and then slice sampling to learn the parameters of

Table 1: Summarizing the Nonparametric priors (Wallach et al., 2010)

	Dirichlet Process	Pitman Yor Process	Uniform Process
Probability Existing Cluster s	$\frac{N_s}{M+\zeta}$	$\frac{N_s-\delta}{M+\zeta}$	$\frac{1}{K+\zeta}$
Probability New Cluster	$\frac{\zeta}{M+\zeta}$	$\frac{\zeta+K\delta}{M+\zeta}$	$\frac{\zeta}{K+\zeta}$
Expected Number of Clusters	$\zeta \log(1+\frac{N}{\zeta})$	$\frac{\Gamma(1+\zeta)}{\delta\Gamma(\delta+\zeta)}N^{\delta}$	$\sqrt{2\zeta N}$
Expected No. of Clusters of Size N_s	$\frac{\zeta}{N_s}$	$\frac{\Gamma(1+\zeta)\prod_{z=1}^{N_s-1}(z-\zeta)}{\Gamma(\delta+\zeta)N_s!}N^{\delta}$	ζ

the hierarchical Dirichlet distributions. The models are initialized using topics obtained from 1000 iterations of LDA, and clusters and topics are resampled jointly for 500 burn in iterations following Wallach (2008) and then summarized using a standard approach in computer science.[4]

4.2 Selecting Across Models: Human Based Experiments

To select among the candidate semiparametric models, we propose a suite of experiments designed to measure the substantive performance of our statistical models. Statistical tools for assessing model fit tend to measure how well a set of parameters approximate the data (Claeskens and Lid, 2010). Using this same basic idea, we assess how well the parameters in our model capture the underlying *substantive* meaning of our text. To do this, we design experiments to assess three features of our model: the quality of clusters, the validity of topics, and how well topics summarize a document. We then combine the information in these experiments to perform model selection with a new graphical tool that varies the weights.

Our method for evaluating the substantive fit of our models builds on two recent advances in the fitting and assessment of Bayesian models. Recent papers in political science, such as Gill and Walker (2005) and Freeman and Gill (2011), recognize that subject experts have valuable information about model parameters. Incorporating these beliefs into an analysis provides valuable information into the analysis. Like these papers, we use carefully-designed instruments to elicit and then use information from subject experts. But for our model assessments it is more useful to include this information after fitting our model than before with priors. This is because unsupervised models are able to discover substantive facts that are unknown before the study (Grimmer and

[4]To summarize the draws from our posterior, we use a standard approach for topic models that balances characterizing the posterior and the possibility of label switching. For more details, see McCallum (2002).

King, 2011). In this instance, using subject experts to elicit priors would hinder the ability of unsupervised models to discover new insights into our data.

While the discoveries from unsupervised models may be unanticipated, they are often intuitive *post-hoc*. We build upon recent work to perform substance-based posterior checks to measure model performance (Gelman et al., 1996; Chang et al., 2009; Mimno and Blei, 2011), by measuring how well our models fit the substance of the press releases. But, the ability to identify the intuitive properties of these discoveries, however, often depends on subject-expertise. Therefore, we perform our experiments using subject experts in American politics.

We now describe our experiments and how we combine the output to make a final model selection.

4.2.1 Cluster Quality

At the top of our hierarchy our model partitions documents based upon their attention to topics. To assess the substantive quality of the clusterings from each model, we employ a measure of cluster quality introduced in Grimmer and King (2011). We sample pairs of press releases where there is some disagreement across our models.[5] We then ask a set of subject experts who are unaware of how the models classified the press releases to evaluate if the press releases are about the same *topics*. We ask the coders to evaluate the pairs of the documents on a three point scale: unrelated (1), loosely related (2), closely related (3).

Using these evaluations, we compute the measure of cluster quality for our candidate models. Clustering models are designed to group together *similar* press releases and to separate distinct texts. With this intuition in mind, for each model z ($z = 1, \ldots, Z$), we compute the average evaluation of pairs assigned to the same category, $\text{mean(within cluster)}_z$, and the average evaluation of pairs assigned to different cluster, $\text{mean(between cluster)}_z$. Our measure of cluster quality for model z is then,

$$\text{CQ}_z \quad = \quad \text{mean(within cluster)}_z - \text{mean(between cluster)}_z$$

[5] If the models agree about how to group a pair of press, we exclude that pair from our evaluations. Pairs where all methods agree cannot provide information about the relative quality of each of the methods

4.2.2 Topic Quality

Our model also measures topics of discussions across concepts. These are intended as low level summaries of the political discussions across texts. We employ two measures of topic quality to measure how well our model parameters–the estimated topics–capture the substantive topics. We use a design from Chang et al. (2009) to measure the coherence of the topics identified by our model and we introduce a new design to measure how well our topics replicate the keywords readers of the text would identify.

Our first design replicates Chang et. al's model of word intrusion. First, we select the five words that are most probable for a given topic. We then select, at random, another word that is among the five most probable words for another topic and randomly insert it into the five word sequence. We then provide the sequence of words to a subject expert and ask them to identify the intruder word. We score the response a "1" if the expert identifies the intruders and a "0" otherwise. Our measure of word intrusion topic quality for model z, TQ(WI)_z, is,

$$\text{TQ(WI)}_z \quad = \quad 2 \times \text{Proportion Identified Intruders.}$$

where we multiply by 2 to ensure this measure has the same upper bound as our cluster quality measure and the proportion of identified intruders is computed across topics and raters for a single model.

A second way to asses topic quality is to ask if the words that our model identifies as keyword summaries of the text are actually good summaries of the documents. To assess this, we first identify a set of key words to summarize a document. To compute this, we use the model output to compute the probability of word token w in document i under model z with, $\Pr(y_{ij} = w)_z = \sum_{k=1}^{K} p(\tau_{ijk} = 1)_z p(y_{ij} = w \,|\, \tau_{ijk} = 1)_z$. We then ask human coders to read the press release and to identify 5 key words for the topics in the press releases. We score a probable word from the model a 1 if it overlaps with the user identified word and a zero otherwise. We then compute our measure of keyword topic quality for model z, TQ(KW)_z, as,

$$\text{TQ(KW)}_z \quad = \quad 2 \times \text{Proportion Keywords Identified.}$$

This evaluation deserves two caveats. First, this is also a measure of how well the topics summarize the documents (which we describe in the next section). Second, it may be that a model could score poorly on this evaluation but still be a useful model for text. Human evaluators may not be able to replicate topics before hand, but evaluate models highly after the fact. Therefore, we will want to carefully consider how much weight to attach to this evaluation in our final measure of model quality.

4.2.3 Topics as Summaries of Documents

Our final evaluation measures our model's representation of each document. Our topic model represents each press release as a mixture of topics. If the mixture of topics is an accurate (and useful) summary of the press release, then information about the topics in a press release should allow coders to identify the main idea in a press release. The final evaluation evaluates how well the models do this.

To construct this experiment, we first sample a set of documents. We then create two prompts for two different evaluators. In one prompt, we create lists of words from highly probable topics and ask the evaluator to guess the likely subject of the press release. In our second prompt, we provide an evaluator with the press release and ask her to briefly describe the subject of the press release. We then ask a third coder to evaluate the similarity of the responses on our three point scale. Averaging across press releases creates summary quality for model z, SQ_z,

$$\text{SQ}_z \quad = \quad \text{Average Rating of Summaries}_z - 1,$$

where we subtract 1 to maintain the same possible maximum across our evaluations.

4.2.4 Final Model Selection

Each of our evaluations considers different features of our model. We average the scores from the evaluations to create a final index that we use to select a single model from our candidate models. To vary the influence of each of our evaluations we compute a weighted average. Varying the weights

Table 2: Models Included in Evaluation and Experimental Scores

Model	ζ (Concentration)	CQ	TQ(WI)	TQ(KW)	SQ
Dir. Proc.	1	0.16	1.12	0.23	1.07
Dir. Proc.	5	0.07	1.10	0.23	1.27
Dir. Proc.	10	0.25	1.23	0.23	1.33
Dir. Proc	Learned	0.31	-	0.25	-
Uniform Proc	1	0.71	1.25	0.24	1.31
Uniform Proc	0.5	0.54	-	0.25	-

attached to each evaluation allows the user to vary the importance of the model features for the final selection stage. Formally, call the weights attached to each measure of quality $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3, \omega_4)$, and suppose that they sum to 1. Then our measure of model quality is, Model Quality$_z(\boldsymbol{\omega})$,

$$\text{Model Quality}_z(\boldsymbol{\omega}) \quad = \quad \omega_1 \text{CQ}_z + \omega_2 \text{TQ(WI)}_z + \omega_3 \text{TQ(KW)}_z + \omega_4 \text{SQ}_z$$

A particular set of values of $\boldsymbol{\omega}$ allows researchers to encode the relative importance of the features of the model for the substantive problem at hand. And by varying $\boldsymbol{\omega}$ across its possible values, researchers are able to identify the weights associated with selecting a model.

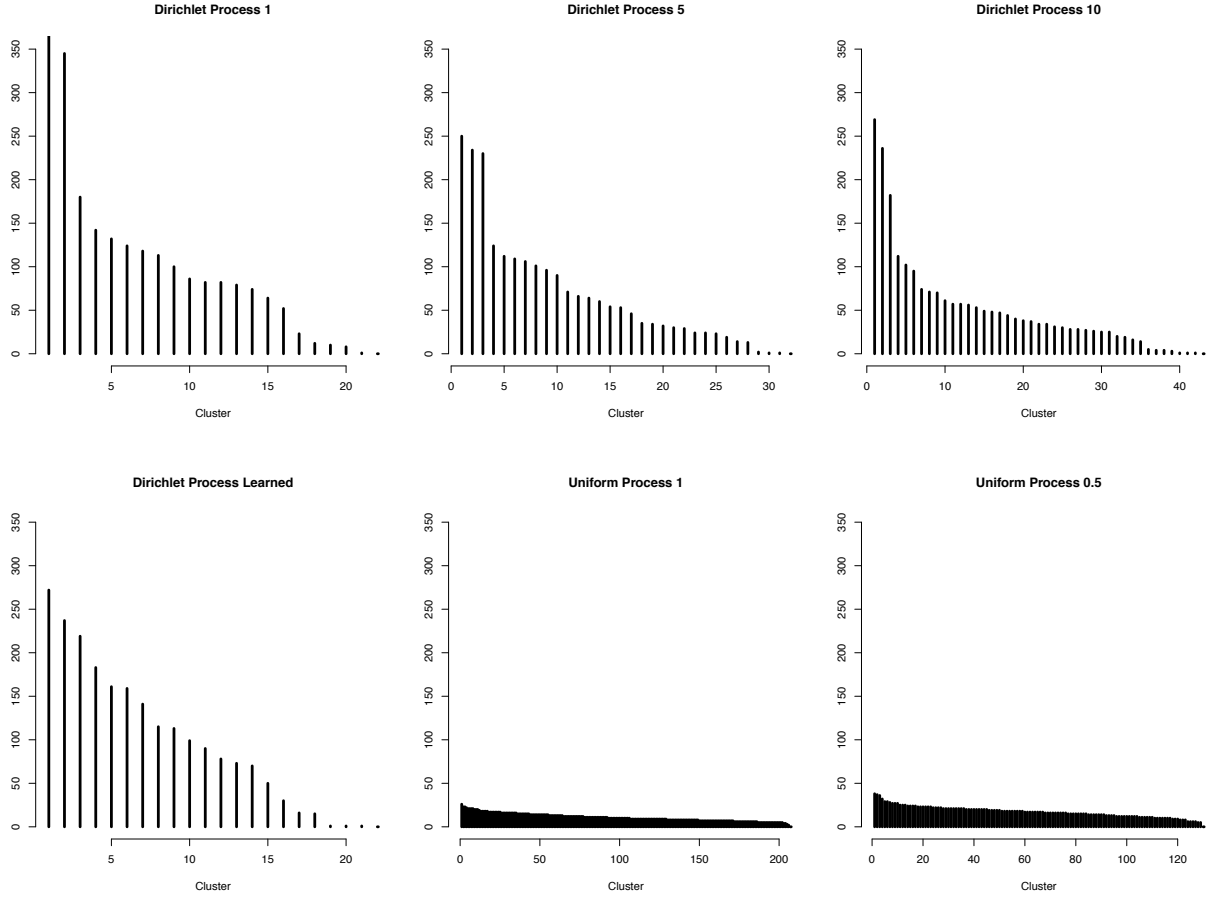
4.3 Selecting a Model for House Press Releases

To perform model selection we carried out our experiments on a subset of 2,225 press releases from our larger corpus and fixed the number of topics in each at 100. On this subset of press releases and with this fixed number of topics, we carried out our preliminary model comparisons using the Dirichlet process prior and the uniform process prior.[6] For the Dirichlet process prior we fit models with fixed concentration parameters and models that learn a value of the concentration parameter. For the uniform process prior we set the concentration parameter here, but in future work we will learn the concentration parameter from the data. Table 2 describes the models and provides summaries of their model fit scores.

The evaluations in Table 2 shows that the largest differences in our substantive evaluations is

[6] We currently are in the process of evaluating our models for the Pitman-Yor process prior as well. The comparison here is useful, because it compares the two most distinctive of our nonparametric priors.

Figure 1: Clustering Comparison: Model Selection Experiments



This figure shows that each prior and concentration parameter values within priors produce different numbers of estimated clusters and documents assigned to clusters.

exhibited in our measures of *cluster quality*. These differences represent the substantial differences in the clusters formed using our different nonparametric priors. To see these differences, Figure 1 shows the number of clusters estimated under each model and how documents are allocated across those clusters. Each histogram exhibits the expected number of estimated clusters from the model (horizontal axis) and the number of documents per cluster (vertical axis) for the 6 models in our preliminary evaluation.

As Section 4.1 would lead us to expect, each of the nonparametric priors results in a very different number of clusters and a different process for allocating documents across clusters. Each of the Dirichlet process priors—both models with the concentration parameters fixed and learned—

exhibit a power-law distribution of documents across clusters, with about 20-45 clusters estimated in each model. The models that use the uniform process prior, however, estimate a very different number of clusters and very different number of documents per cluster. True to its name, the uniform process prior evenly allocates documents across a large number of clusters.

Figure 1 shows that the nonparametric priors cause the models to estimate very different numbers of clusters and allocate across those documents quite differently. Figure 2 shows that this causes the models to create distinctive partitions of documents. To compare the documents, we use *variation of information*—a distance metric on clusters with a number of desirable properties (Meila, 2003; Grimmer and King, 2011). The left-hand plot of Figure 2 visualizes the distance matrix between clusters and the right-hand plot provides a two-dimensional scaling of the clusters, based on the distances from the variation of information metric.

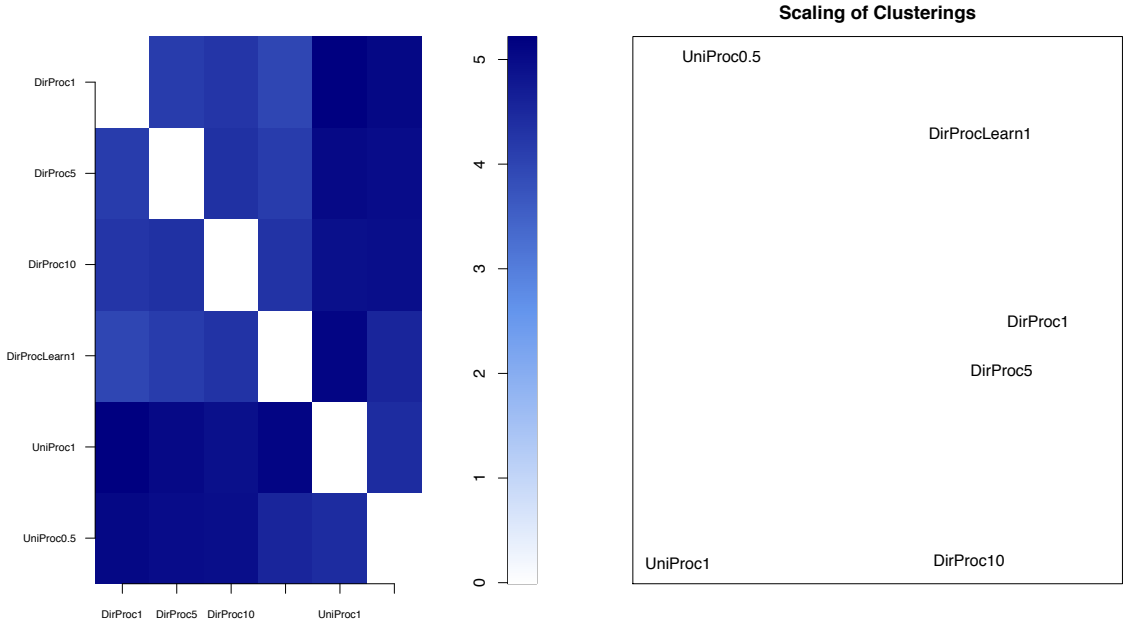
Both plots in Figure 2 show that the priors produce strikingly different clusterings of the documents. Dirichlet process priors produce clusterings that are similar–and therefore group together in the right-hand plot of Figure 2. The uniform process prior provides a very different partition than the Dirichlet process priors: the distance across priors is much greater than the distance within priors. This demonstrates the importance of varying the consequential (and subtle) assumptions that underlie clustering methods. And the differences are not just due to the uniform process priors creating many more clusters—the prior is also grouping together different documents than the Dirichlet process prior. Only about 30% of the pairs of documents that the uniform process prior clusters together are also clustered together by the Dirichlet process prior models. In short: the models have real consequences for determining which documents should be grouped together.

These preliminary experiments make clear the differences in the clustering of documents across our models. Our experiments allow us to select from these models using substantive approaches. To conduct the experiments, we employed 11 subject area experts.[7] The results of our evaluations are contained in Table 2.

Table 2 provides substantial evidence that the uniform process prior, with the concentration parameter set to 1, provides the best substantive fit for the press releases. The first column shows

[7]Our subject experts are either current graduate students who study American politics, or a set of highly trained undergraduate coders. Our coders exhibited high agreement, indicative of their shared ability to evaluate documents similarly.

Figure 2: Distances Between Clusters



This figure demonstrates that different models produce very different partitions of the press releases. The left-hand plot visualizes the distance matrix between clusterings, demonstrating that the distance across models is greater than the differences within clusters. The right-hand plot is a scaling of this distance matrix, demonstrating that the clusterings from the uniform process prior models and Dirichlet process prior models provide two very different partitions of the press releases.

that the clustering from the uniform process prior is rated much higher than the other partitions. Indeed, our evaluators rated the clusterings from uniform Process prior over twice as high as the clusterings from the Dirichlet process prior. And our scores show that this is not just a preference for models with a large number of clusters. First, built into our cluster quality metric is an implicit penalization for having too many clusters. If a pair of documents is placed in different clusters by our model, but our coders rate the pair as being highly similar, the partition is penalized. Second, the number of estimated clusters in our model only correlate loosely with the coders' evaluations. The highest rated clustering of the Dirichlet process prior models, the model with the concentration parameter learned, has only 22 clusters. This is the smallest number of clusters for the Dirichlet process models.

The uniform process prior performs well on evaluations of the topics and the summary of the documents with topics, though there is substantially less variation on these evaluations. On the word intrusion measure (Column 2, Table 2), the uniform process prior scored the highest, with the coders correctly identifying the intruder word 62.5% of the time. But this reflects the overall high quality of the topics in our models. Our coders correctly identified the intruder word 58.8% of the time, significantly more than than the 16.7% of the expected by chance. Likewise, our coders were able to generate overlapping keywords with our models about 12.5% of the time (Column 3, Table 2). This low number reflects the substantial difficulty in our keyword experiment. Finally, the models performed similarly on the summary quality test.

Finally, both the uniform process prior and the Dirichlet process prior with concentration parameter set to 10 had essentially equivalent scores, of 1.31 and 1.33, respectively. Note that these are scores that are quite high to the highest possible scores of 2 (a perfect score on the evaluations). This implies that our representations of texts as mixtures of topics still retains important *substantive* information about the texts.

To make a final model selection, we compute the scores for all models varying the weights placed on the evaluations. Then, for each weight we determine the method that scores the highest. Note, that this procedure is equivalent to evaluating a function over all the points in a three-dimensional (four component) simplex. This implies that we can use off the shelf tools to enumerate the weights and to compute the final model scores for each of our methods.

An immediate consequence of our weighting procedure for model selection is that a model can only be selected at this stage if there is at least one experiment where it scores the highest. This eliminates from consideration two versions of the Dirichlet process prior—the models with the concentration parameter set to 1 and 5. Figure 3 visualizes the choice between the two remaining methods. Each ternaryplot in Figure 3 varies the weights for three of the four weights. As we move to the corners of each plot, there is more weight placed on the evaluation placed at that corner. At each point in the plot—a weight attached to our experiments—we select either the uniform process prior (in blue) or the Dirichlet process prior with concentration parameter 10 (in red).

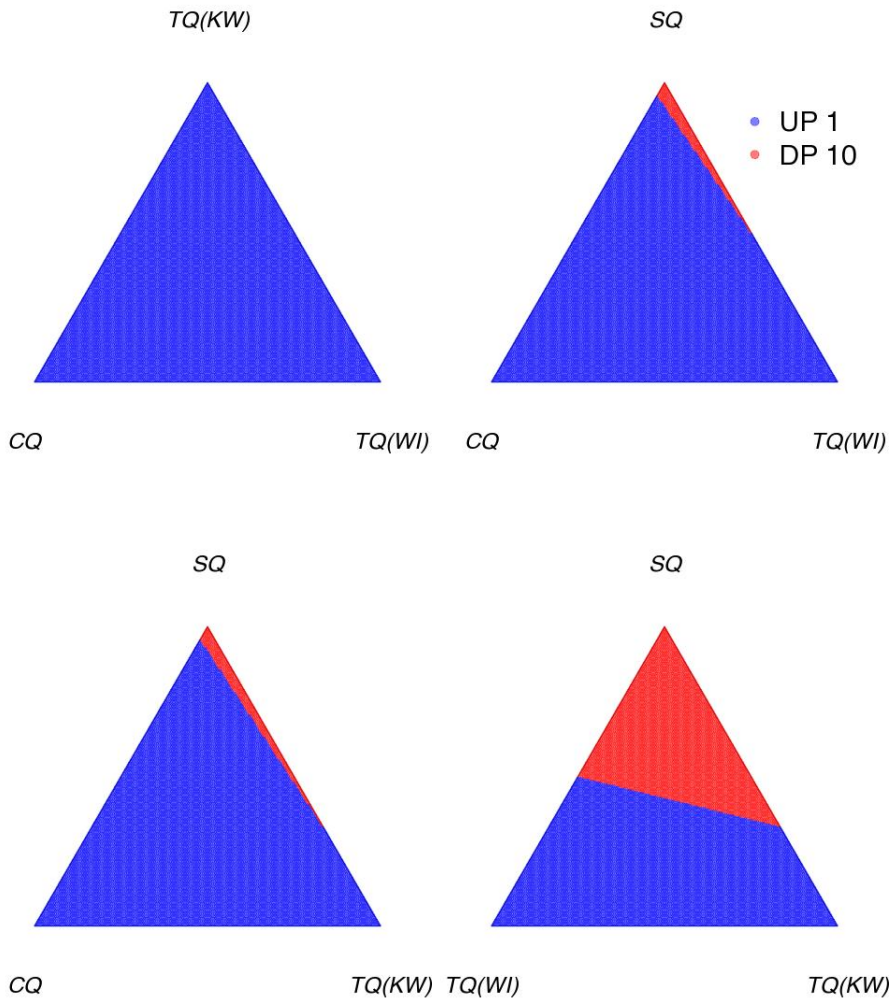
Each plot in Figure 3 shows that for almost all weight combinations, the uniform process prior is evaluated as having the highest model quality. Consider the far-left ternary plot, which shows that the uniform process prior scores highest on cluster quality, word intrusion, and keyword quality. And therefore this model is selected if we attach zero weight to summary quality. The middle two plots show that even if we attach substantial weight to summary quality, the output from the uniform process prior is selected as the superior model. There are only substantial areas where we would selected the Dirichlet process prior if we attach no weight to the cluster quality measure and substantial weight to summary quality, which is demonstrated in the far-right plot of Figure 3.

Given the superior performance across the vast majority of possible weight combinations, we select the uniform process prior as our final model to perform substantive evaluations. The substantial differences across clusterings, and the importance of the clusterings for our substantive analysis, suggests that we should place substantial weight on the cluster quality measure when selecting our final model. Further, the differences in topic-quality across models are negligible—something expected based on the similarity of the models.

5 Granular Measures of Political Discussion in House Press Releases

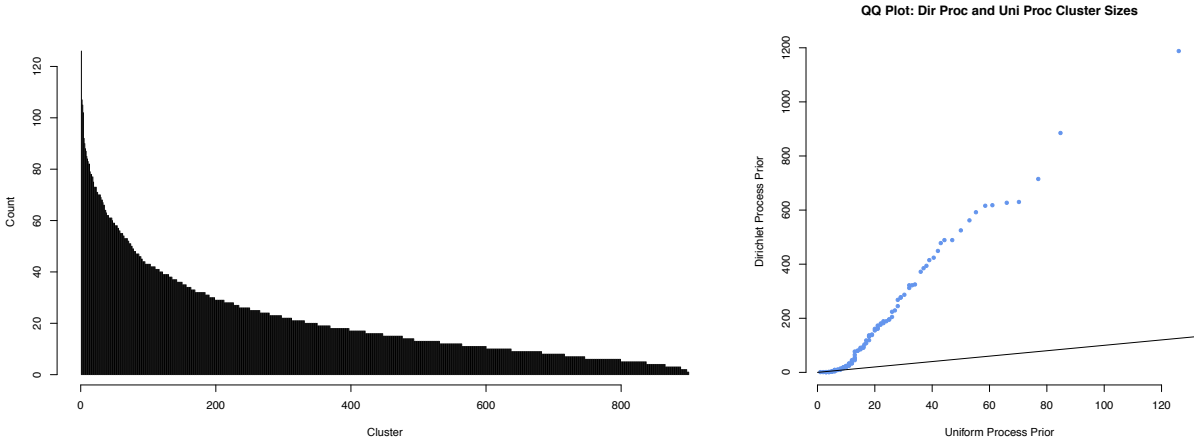
This section shows that our statistical model provides a substantively interesting grouping of the press releases. To show this, we fit the uniform process prior model to entire collection of press releases, setting the concentration parameter at 1 and the number of topics to 500. The distribution of documents across clusters is presented in Figure 4.

Figure 3: Final Model Selection



This figure shows that for almost all values of the weights attached to evaluations, our model selection procedure selects the uniform process prior.

Figure 4: Cluster Sizes for uniform Process Prior on Complete Data Set



This figure shows that the uniform process prior evenly partitions documents across a large number of clusters. We show below that the differences across clusters are meaningful—reflecting differences in the strategic use of language across senators when discussing a topic.

Figure 4 shows that the uniform process prior identifies a *very* large number of clusters: 900. And true to the assumptions of the uniform process prior, the clusters are much more evenly sized than clusters from a Dirichlet process prior on the same collection of press releases. The left-hand plot in Figure 4 sorts the clusters according to their size. A heuristic check of this histogram shows that many clusters have very similar number of press releases allocated to them. A closer look reveals this to the be the case: the average cluster contains about 21 press releases and the interquantile range of clusters containing 9 to 28 press releases. This allocation of clusters is *substantially* different than the partition from a Dirichlet process prior applied to the same data. The right-hand plot in Figure 4 compares the distribution of cluster sizes from the uniform process prior (horizontal axis) to the distribution of cluster sizes from the Dirichlet process prior (vertical axis) using a QQ-plot. This figure shows that the Dirichlet process prior creates a clustering with *much* longer tails. For example, the largest Dirichlet process prior cluster contains *ten times* as many press releases as the largest press release from the uniform process prior.

5.1 What Representatives Talk About

To summarize the clusters, we computed words that our model identified as highly likely to occur in each cluster. Specifically, for each cluster we computed the ten words that were most likely to

occur, given the mix of topics in the cluster and the probability of a word occurring in each topic. Previous work characterizes the content of topics using similar representations (Quinn et al., 2010) and our experiments confirm that these are useful summaries of the documents belonging to each cluster. Table 3 contains these distinguishing words.

Table 3: Cluster Size and Keywords for the Largest 50 Clusters

No. Docs	Keywords
126	academy,states,united,students,high,military,academies,naval,nominated,marine
107	spending,tax,economic,businesses,debt,budget,americans,job,business,obama
105	haiti,states,united,relief,earthquake,haitian,disaster,citizens,international,embassy
102	security,states,united,terrorist,administration,intelligence,plan,hearing,homeland,terrorists
92	haiti,tax,earthquake,relief,states,united,charitable,haitian,americans,contributions
90	tax,economic,unemployment,spending,job,businesses,recovery,americans,percent,democrats
88	spending,tax,debt,economic,percent,budget,businesses,job,plan,trillion
87	academy,military,states,high,united,students,academies,naval,honor,nominated
85	states,united,security,terrorist,terrorists,trial,administration,intelligence,guantanamo,obama
84	care,tax,insurance,costs,coverage,billion,cost,percent,medicare,vote
83	care,americans,insurance,democrats,tax,law,costs,percent,spending,businesses
82	haiti,states,united,earthquake,haitian,relief,disaster,emergency,citizens,rescue
82	care,insurance,coverage,costs,medicare,percent,law,billion,seniors,cost
79	economic,tax,budget,recovery,spending,businesses,fiscal,job,financial,americans
78	spending,budget,fiscal,debt,earmarks,vote,deficit,process,republicans,passed
78	spending,care,debt,tax,economic,americans,businesses,budget,percent,job
77	oil,spill,drilling,bp,plan,gas,offshore,safety,law,markey
77	states,united,iran,nuclear,administration,obama,international,sanctions,security,nations
75	states,united,haiti,children,haitian,guard,earthquake,relief,red,citizens
73	grant,liheap,billion,percent,states,grants,budget,programs,increase,food
73	recovery,rail,economic,project,transit,projects,grant,highspeed,reinvestment,investment
73	grant,project,economic,airport,grants,facility,safety,aviation,improve,job
73	veterans,states,military,united,afghanistan,troops,women,honor,iraq,war
71	tax,spending,debt,economic,percent,businesses,americans,democrats,budget,job
71	tax,businesses,business,economic,percent,job,billion,hire,recovery,employees
70	debt,spending,budget,trillion,fiscal,deficit,democrats,obama,vote,increase
70	care,states,spending,plan,united,democrats,americans,obama,republicans,budget
70	amendment,court,law,vote,supreme,elections,corporations,political,passed,election
70	spending,debt,budget,amendment,fiscal,vote,trillion,deficit,billion,percent
69	oil,hearing,spill,bp,markey,drilling,safety,gas,disaster,deepwater
68	recovery,training,grant,economic,grants,students,programs,job,businesses,business
68	tax,debt,spending,economic,businesses,democrats,americans,job,taxes,vote
67	care,democrats,americans,vote,tax,obama,insurance,republicans,costs,pelosi
66	care,spending,democrats,americans,republicans,obama,vote,budget,agenda,tax
66	spending,economic,tax,job,businesses,americans,percent,unemployment,democrats,obama
64	financial,plan,fannie,democrats,freddie,businesses,wall,spending,mae,billion
64	recovery,grant,broadband,economic,project,grants,projects,access,reinvestment,awarded
63	oil,spill,bp,drilling,hearing,gas,offshore,disaster,law,markey
62	care,vote,obama,hearing,amendment,passed,democrats,process,pelosi,negotiations
62	tax,businesses,economic,benefits,unemployment,business,job,americans,oil,housing
62	students,academy,high,schools,art,director,arts,military,student,academies
61	haiti,states,united,earthquake,relief,charitable,americans,tax,military,contributions
61	military,dont,troops,afghanistan,repeal,vote,study,forces,americans,women
61	financial,wall,economic,billion,spending,fannie,banks,freddie,administration,taxpayers
61	care,grant,recovery,grants,medical,programs,students,training,disease,projects
61	spending,tax,economic,plan,businesses,americans,job,obama,democrats,budget
60	spending,pay,tax,budget,vote,fiscal,economic,billion,raise,deficit
59	care,democrats,plan,spending,americans,vote,obama,law,insurance,passed
59	project,study,plan,lakes,flood,carp,corps,insurance,environmental,army
59	care,democrats,plan,obama,tax,costs,vote,administration,americans,cost

A quick check of Table 3 shows that our clusters group together substantively similar press releases. For example, our largest cluster uses words that solicit applications for the military academy or announce Congressional recommendations for the academies. Words that have a high likelihood of occurring in this cluster are `academy, states, united, students, miliarty`. A manual check of the press releases in this category confirm that they describe how representatives allocate their nominations to the service academy. For example, James Sensenbrenner, (R-WI) issued a press release where he announced, "that Claire Palmer of Elm Grove has been named his principal nominee to the United States Air Force Academy in Colorado Springs, CO" (Sensenbrenner, 2010*b*). Other categories reflect the major news events during the year. Our model groups together press releases about the mortgage crisis and the management of Fannie Mae and Freddie Mac, a cluster characterized by words like, `financial,plan,fannie,democrats,freddie,businesses,wall`. The model also identifies a cluster of press releases designed to claim credit for money allocated to the district. One cluster of press releases announces funding for airports—with characterizing words `grant,project,airport,grants`. Press releases grouped into this category included announcements that a "$1.5 million grant" was awarded to "continue construction of a new terminal building at the Santa Barbara Airport" (Capps, 2010) or that the "Gary/Chicago International Airport" received a "$5 million grant" for a "runway extension project" (Visclosky, 2010).

Both the keywords in Table 3 and our manual check of the clusters confirm that they group together substantively interesting speeches. But the number of clusters our model identifies, 900, is *many* more than that used in parametric models of Congressional speeches (Quinn et al., 2010; Grimmer, 2010). And analyzing the keywords in Table 3 might suggests that there are many redundant clusters in our data set, with topics seemingly being repeated. In the next section, we perform an in depth case study of three issue areas—the earthquake in Haiti, healthcare reform, and the oil spill—to show that the distinctions across our clusters are not redundant. Rather, they are substantively interesting divisions that capture important differences in language or focus that would be missed by other models with fewer clusters. The result is a *granular* and nuanced partition of Congressional press releases.

5.2 Substantive Differences Across Similar Clusters

To perform our comparisons of press releases allocated to seemingly similar clusters, we chose three of the most salient news stories during our study. We identified the largest clusters associated with each issue area: the earthquake in Haiti, the healthcare reform debate, and the BP oil spill. To demonstrate the conceptual validity of our groupings, we first created three aggregated clusters—one for each broader news story. Then following an evaluation suggested in Quinn et al. (2010), we show that the overtime variation in each category corresponds to major, salient events. Consider the top plot in Figure 5, which shows the overtime variation in the number of press releases about the Haitian earthquake and subsequent relief efforts. We see that after the earthquake on January 12th, the number of press releases about the quake jumps as members of Congress express their sympathy. The number of press releases rises again as Congress considers legislation to allow donations from Haiti in 2010 to count towards 2009 taxes.

The temporal dynamics for our other two categories also correspond to real events. The most press releases about health care occurs before the House passes the Senate bill and the corresponding reconciliation side car. And press releases about off-shore drilling and the BP oil spill increase as the oil spill continued and as Congress holds hearings about the spill.

This shows that the press releases are correctly classified into these broader categories. In the next section, we show that there are substantive differences across clusters that appear redundant.

5.2.1 Haiti Relief

We begin our case studies with a comparison of clusters that discuss the Haitian earthquake and subsequent relief effort–a total of five clusters and 415 press releases. To compare the clusters, we first identify the words that distinguish each cluster about the Haitian earthquake from the other Haitian earthquake clusters and place them in Table 4. By comparing only clusters about Haiti, we are able to identify the subtle language differences that distinguish each group of press releases.[8] We augment this comparison with a manual analysis of the press releases assigned to each category–allowing us to assess the distinguishing features of each cluster. We include three other pieces of information in each row of Table 4 and the other tables comparing across clusters.

[8]To make this comparison, we use the mutual information between word and cluster labels (Manning et al., 2008)

The first column contains the number of press releases assigned to each cluster, the second column describes the proportion of press release from Democrats, and the third column describes the average *extremity* of the representatives who issued the press releases. To compute this, we use DW-Nominate scores to measure the extent to which a representative is more extreme (positive) or less extreme (negative) than the average member of their party. We will use this below to assess who is issuing press releases about substantive topics.

Table 4: Distinctive Words Among Haiti Clusters

Row	No.Doc	Prop.D	Ext	Dist.Words
1	75	0.65	0.06	secur,homeland,children,unit,adopt,work,secretari,famili,humanitarian,orphan
2	105	0.67	0.03	888,4747,407,depart,center,org,red,cross,http,famili
3	61	0.77	0.02	charit,tax,deduct,2009,claim,legisl,said,return,allow,pass
4	82	0.59	-0.01	emerg,www,hospit,temporari,medic,tp,aid,respons,friend,gov
5	92	0.79	-0.02	tax,2009,deduct,charit,legisl,claim,return,contribut,year,allow

This reveals that five clusters discuss distinctive features of the earthquake disaster. Consider the first row of Table 4. This cluster contains press releases from members of Congress about expediting children adopt from Haiti and appeals to the Department of Homeland security for refugees from the Haitian crisis. For example, Pat Tieberi (R-OH) called on administration officials, "to implement a coordinated plan to ensure all orphans in Haiti be evacuated" (Tieberi, 2010), Pete Hoekstra (R-MI) announced that he "recently introduced a bill that would safely expedite the process for adoptions of Haitian orphans following the devastating earthquake" (Hoekstra, 2010), and Edolphus Towns (D-NY) announced his support for a plan to extend "temporary protected status to [Haitian] immigrants" (Towns, 2010).

In contrast, the second row of Table 4 describes press releases from Congressional officials that provide information about how to donate money to the relief effort, including the `red cross` number, "1-888-407-4747" (Edwards, 2010; Slaughter, 2010). The third row in Table 4 describes groups of press releases about proposed legislation to ensure that tax donations made in 2010 would be eligible for deductions on the 2009 taxes. And the fourth row describes how the government has been involved with the provision of emergency medicine.

The proliferation of clusters about Haiti, therefore, captures nuanced differences in how repre-

sentatives address the same core topic. This suggests that our model provides an ideal setting for the careful analysis of how representatives respond to major overseas tragedies.

5.2.2 Health Care Debate

The dominant political issue during the time period of our study was the passage of health care reform. In this section, we show that the largest clusters in our model identify substantial differences in how the parties discussed the legislation and demonstrates that the most liberal Democrats were likely to defend the legislation, while conservative Republicans were most likely to criticize the legislation. To show this, we grouped together the 9 largest clusters about health care reform and then identified the words that distinguish the health care press releases from each other.

Table 5: Distinctive Words Among Health Care Clusters

Row	No.Doc	Prop.D	Ext.	Dist.Words
1	84	0.39	-0.02	incom,pai,000,insur,tax,year,employ,make,million,compani
2	83	0.08	0.03	voic,statement,agenda,spend,r,idea,solut,govern,administr,instead
3	82	0.23	0.04	negoti,transpar,open,januari,vote,public,span,broadcast,speaker,pelosi
4	70	0.16	0.03	job,republican,listen,agenda,spend,address,elect,prioriti,fiscal,work
5	67	0.05	0.01	support,statement,need,takeov,debt,democrat,govern,r,economi,want
6	66	0.93	0.07	d,coverag,afford,provid,insur,benefit,medicar,premium,help,assist
7	62	0.13	0.04	repeal,replac,job,mandat,feder,law,kill,taxpay,cut,busi
8	59	0.07	0.02	medicaid,committe,dai,public,know,process,administr,medicar,democrat,r
9	59	0.07	0.08	januari,door,senat,democrat,negoti,deal,c,span,obama,close

To see the subtle differences captured in our model, consider the second row of Table 5. This captures groups of press releases that express Republicans' recognition of the special election of Scott Brown (R-MA), arguing that "[t]he people of Massachusetts—an historically liberal state—spoke with one voice to reject [Obama's] legislative agenda" (Lucas, 2010). James Sensenbrenner (R-WI) assures constituents that "I, along with my fellow Republicans, have heard you" (Sensenbrenner, 2010*a*). A different message is sent with the cluster in the eight row, asking represenetatives to repeal and replace the health care legislation, while the third row chastises Democrats (and some Republicans) for not conducting the health care negotiations in public.

Contrast this with the sixth row of Table 5. Here Democrats—particularly liberal Democrats—

defined the administrations' health care reform, arguing that it expands coverage to seniors. For example one press release announced that, "Congressman Maurice Hinchey (D-NY) today met with seniors at Kendal at Ithaca retirement community to discuss how Medicare coverage will improve as a result the health care reform measure he voted for and President Obama signed into law last week" (Hinchey, 2010). While members of both parties disputed the effect of the legislation on health care reform (row 1, Table 5).

This analysis shows that our clusters about health care reform result in meaningful divisions of our press releases, based on the positions advocated in those documents. Again, this provides a nuanced and granular perspective on the positions discussed in the press releases.

5.2.3 BP Oil Spill

This section briefly compares three clusters about press releases about the BP oil spill. This reveals clear divisions in how representatives discussed the same basic topic of the flow of oil out of a collapsed oil rig in the Gulf of Mexico.

Table 6: Distinctive Words Among BP Oil Spill Clusters

Row	No.Docs	Prop.D	Avg.Ext	Dist.Words
1	77	0.74	0.05	safeti,act,pass,prevent,requir,industri,said,develop,blowout,drill
2	63	0.46	0.06	r,job,moratorium,obama,product,t,legisl,administr,congress,foreign
3	69	0.71	0.04	subcommitte,question,dispers,ask,claim,tuesdai,guard,brief,commerc

The first row in Table 6 contains press releases about proposed legislation in Congress. For example, Eliot Engel (D-NY) issued a press release touting his vote in favor of "better methods and technologies for oil spill cleanup" (Engel, 2010). The second row of Table 6 contains a cluster of press releases that discuss the Obama administration's moratorium on deep water drilling and its consequences on the job market. The third cluster of press releases about the oil spill deal with BP's liabilities and how the money for BP should be used to compensate victims on the Gulf Coast. Some representatives described the mandate from the Obama administration for a "$20 billion into an escrow fund to pay claims filed against the company in the wake of the Gulf oil spill" as "Chicago-style political shakedown" (Price, 2010), while others "called on the oil company to

... direct funds to an escrow account dedicated to cleaning up the Deepwater Horizon spill and compensating its victims" (Loebsack, 2010).

5.3 Artificial Polarization in Political Discussion

The previous section shows that our model is able to retrieve subtle distinctions in political texts. In this section, we use the clusterings from our model to characterize a systematic bias in who discusses major policy issues. Specifically, we show that press releases about policy topics tend to come from more extreme members of Congress, while more moderate members of Congress avoid these discussions. This induces an *artificial polarization*, a polarization that exacerbates the well-documented polarization in roll call votes in Congress (Poole and Rosenthal, 1997).

For each cluster we determined the average level of polarization of the representatives who issued the press releases in that cluster.[9] We then used a bootstrap sampling method to calculate a p-value of obtaining an extremity score for a cluster of particular size. Specifically, we ask how likely an extremity score is if representatives randomly issued press releases in the cluster, without systematic ideological selection. We then compare the clusters of press releases with significant (p¡0.05) and positive extremity scores to the press releases with significant and negative extremity scores. For preliminary purposes here, we make this comparison by identifying words that distinguish extreme clusters of press releases–clusters with lots of press releases from extreme members–from moderate clusters of press releases–clusters with lots ofpress releases from moderate members. To identify the distinctive words, we use the smoothed-log odds algorithm introduced in Monroe, Colaresi and Quinn (2008). Table 7 contains the 10 words for both the extreme clusters (first row) and the moderate clusters (second row).

Table 7: Extreme Representatives Discuss Policy More Often

Bias	Key Words
Extreme	peopl,american,statement,care,health,obama,right,r,democrat,republican
Moderate	announc,grant,program,local,provid,counti,fund,improv,area,said

What this reveals is that clusters that are characterized by significant extremity also are charac-

[9] We smoothed this estimate using a hierarchical model to avoid attributing large extremity to clusters with small numbers

terized by words that used in policy discussions. This provides good evidence that representatives at the political extremes emphasize policy at a disproportionate rate. Table 7 shows that extreme clusters are characterized by their use of words like `health care`, references to President `Obama`, and `statements` about legislation under consideration in Washington. Moderate clusters–clusters filled with documents from moderate Republicans and Democrats—are characterized by their focus on pork barrel politics. Press releases from moderates are characterized by `announcing grants` for `local programs` that are used as `funds` to `improv` local infrastructure.

6 Conclusion

This paper provides a new model for political texts and a new methodology for model selection. We contribute a new model that simultaneously clusters documents and learns the topics that comprise each document. Using this model, we provide a methodology for model selection that combines human and statistical guidance. Statistically, we showed how to extend our model to a class of semiparametric models, each of which estimates the number of clusters. But the estimated number of clusters is necessarily model dependent. We therefore proposed a second stage of model selection where we carefully elicit the input of subject experts.

Applying this methodology to a new collection of House press releases, we show that our model is able to provide substantively interesting, and granular, estimates of how discussion occurs in House press releases. Using these estimates, we show that there is an artificial polarization in political debate, with extreme members much more likely to discuss policy issues, while more moderate members avoid policy and instead emphasize appropriations.

Our framework—while developed in the context of a particular model—can be extended to many other models and settings. In future work we plan to extend our approach to other text models and even to nontextual data, like estimates of ideal points from roll call votes.

A MCMC Algorithms

We characterize the posterior distribution using a collapsed Gibbs algorithm to assign words to topics and documents to clusters. Dirichlet–multinomial conjugacy allows us to integrate out $\boldsymbol{\pi}_i$, $\boldsymbol{\theta}_k$, $\boldsymbol{m}_s$, and $\boldsymbol{m}$. After integrating out these parameters, we are able to sample the topic assignments

word tokens and cluster assignments for documents. This also is where we determine the number of clusters. Based upon these draws, we then can recover $\boldsymbol{\pi}_i$, $\boldsymbol{\theta}_k$, $\boldsymbol{m}_s$, and $\boldsymbol{m}$. To obtain data-driven estimates of $\alpha, \alpha_0, \alpha_1$, η, and ζ we use a slice sampling algorithm (Neal, 2003). Full details on the algorithms are available in a supplemental appendix.

References

Arratia, R, A Barbour and S Tavare. 2003. *Logarithmic Combinatorial Structures: A Probabilistic Approach*. Monographs in Mathematics. European Mathematical Society.

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

Blei, David and Michael Jordan. 2006. "Variational Inference for Dirichlet Process Mixtures." *Journal of Bayesian Analysis* 1(1):121–144.

Box, George. 1979. Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics: Proceedings of a Workshop*.

Capps, Rep. Lois. 2010. "Rep. Capps Announces $1.5 Million in Funding for Santa Barbara Airport Terminal Construction.".

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang and David Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Neural Information Processing Systems*.

Claeskens, Gerda and Nils Lid. 2010. *Model Selection and Model Averaging*. Cambridge University Press.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(02):355–370.

Edwards, Rep. Donna. 2010. "Rep. Edwards Issues Statement on Tragic Earthquake in Haiti.".

Engel, Rep. Eliot. 2010. "Rep. Engel Votes for Safer Drilling, New Clean-Up Technologies, to Prevent Future Oil Spill Disasters.".

Escobar, Michael and Mike West. 1995. "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association* 90(430):577–588.

Fenno, Richard. 1978. *Home Style: House Members in their Districts*. Addison Wesley.

Ferguson, T. 1973. "A Bayesian Analysis of Some Nonparametric Problems." *Annals of Statistics* 1(2):209–230.

Fraley, Chris and Adrian Raftery. 2002. "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association* 97(458):611.

Freeman, John and Jeff Gill. 2011. "Dynamic Elicited Priors for Updating Covert Networks." Washington University, St Louis Mimeo.

Gelman, Andrew, John Carlin, Hal Stern and Donal Rubin. 1996. *Bayesian Data Analysis*. Chapman & Hall.

Gill, Jeff and George Casella. 2009. "Nonparametric Priors for Ordinal Bayesian Social Science Models: Specification and Estimation." *Journal of the American Statistical Association* 104(486):453–454.

Gill, Jeff and Lee Walker. 2005. "Elicited Priors for Bayesian Model Specification in Political Science Research." *Journal of Politics* 67(3):841–872.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1):1–35.

Grimmer, Justin. 2011. "Representational Style: What Legislators Say and Why It Matters." Stanford University Mimeo.

Grimmer, Justin and Gary King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization." *Proceedings of the National Academy of Sciences* 108(7):2643–2650.

Hinchey, Rep. Maurice. 2010. "Rep. Hinchey Meets with Ithaca Seniors to Outline How Medicare Improvements in Health Care Reform Will Benefit Them.".

Hoekstra, Rep. Pete. 2010. "Rep. Hoekstra Introduces Bill to Expedite Haitian Adoptions.".

Hopkins, Daniel and Gary King. 2010. "Extracting Systematic Social Science Meaning from Text." *American Journal of Political Science* 54(1):229–247.

Huelsenbeck, JP, S Jain, SWD Frost and SLK Pond. 2006. "A Dirichlet Process Model for Detecting Positive Selection in Protein Coding DNA Sequences." *Proceedings of the National Academy of Sciences* 103:6263–6268.

Jensen, Shane and Jun Liu. 2008. "Bayesian Clustering of Transcription Factor Binding Motifs." *Journal of the American Statistical Association* 103:188–200.

Krosnick, Jon. 1999. "Survey Research." *Annual Review of Psychology* 50(1):537–567.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(02):311–331.

Lipinski, Daniel. 2004. *Congressional Communication: Content and Consequences*. University of Michigan Press.

Loebsack, Rep. Dave. 2010. "Rep. Dave Loebsack Calls on BP to Set Up Independent Account for Payments to Victims, Cleanup in Gulf.".

Lucas, Rep. Frank. 2010. "Change We Don't Want.".

Manning, Christopher et al. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Mansbridge, Jane. 2003. "Rethinking Representation." *American Political Science Review* 97(4):515–528.

McCallum, Andrew. 2002. "MALLET: A Machine Learning Toolkit." UMASS Amherst Software.

Meila, Merina. 2003. "Comparing Clusterings by the Variation of Information." *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop* .

Mimno, David and David Blei. 2011. Bayesian Checking of Topic Models. In *Empirical Methods in Natural Language Processing*.

Monroe, Burt, Michael Colaresi and Kevin Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16(4):372.

Neal, R.M. 2003. "Slice Samling." *Annals of Statistics* 31(705–767).

Orbanz, P and YW Teh. 2010. Bayesian Nonparametric Models. In *Encyclopedia of Machine Learnings*. Springer.

Pitman, Jim. 2002. "Combinatorial Stochastic Processes." Tech. Rep. 621, Department of Statistics, University of California Berkeley.

Pitman, Jim and Marc Yor. 1997. "The Two-Parameter Poisson-Dirichlet Distribution Derived From a Stable Subordinator." *Annals of Probability* 25(2):855–900.

Poole, Keith and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press.

Porter, Martin. 1980. "An Algorithm for Suffix Stripping." *Program* 14(3):130–137.

Price, Rep. Tom. 2010. "Rep. Price Issues Statement on Chicago Style Political Shakedown.".

Qin, ZS, AL McCue, W Thompson, L Mayerhofer, CE Lawrence and JS Liu. 2003. "Identification of Co-Regulated Genes Through Bayesian Clustering of Predicted Regulatory Binding Sites." *Nature Biotechnology* 21:435–439.

Quinn, Kevin et al. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1).

Sensenbrenner, Rep. James. 2010*a*. "March Madness, Washington Style.".

Sensenbrenner, Rep. James. 2010*b*. "Rep. Sensenbrenner Announces Principal Nominee to US Air Force Academy.".

Sethuraman, J. 1994. "A Constructive Definition of Dirichlet Priors." *Statistica Sinica* 4:639–650.

Slaughter, Rep. Louise. 2010. "Statement by Congresswoman Slaughter on Haitian Earthquake.".

Spirling, Arthur and Kevin Quinn. 2011. "Identifying Intraparty Voting Blocs in the UK House of Commons." *Journal of the American Statistical Association* .

Teh, Yee Weh, Michael Jordan, Matthew Beal and David Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101(476):1566–1581.

Teh, YW. 2010. Dirichlet Processes. In *Encyclopedia of Machine Learning*. Springer.

Tieberi, Rep. Patrick. 2010. "Rep. Tiberi Works to Help Speed Up Adoptive Parents Link with Their Haitian Children.".

Towns, Rep. Edolphus. 2010. "Rep. Towns Issues Statement on Temporary Protected Status for Haitian Refugees in US.".

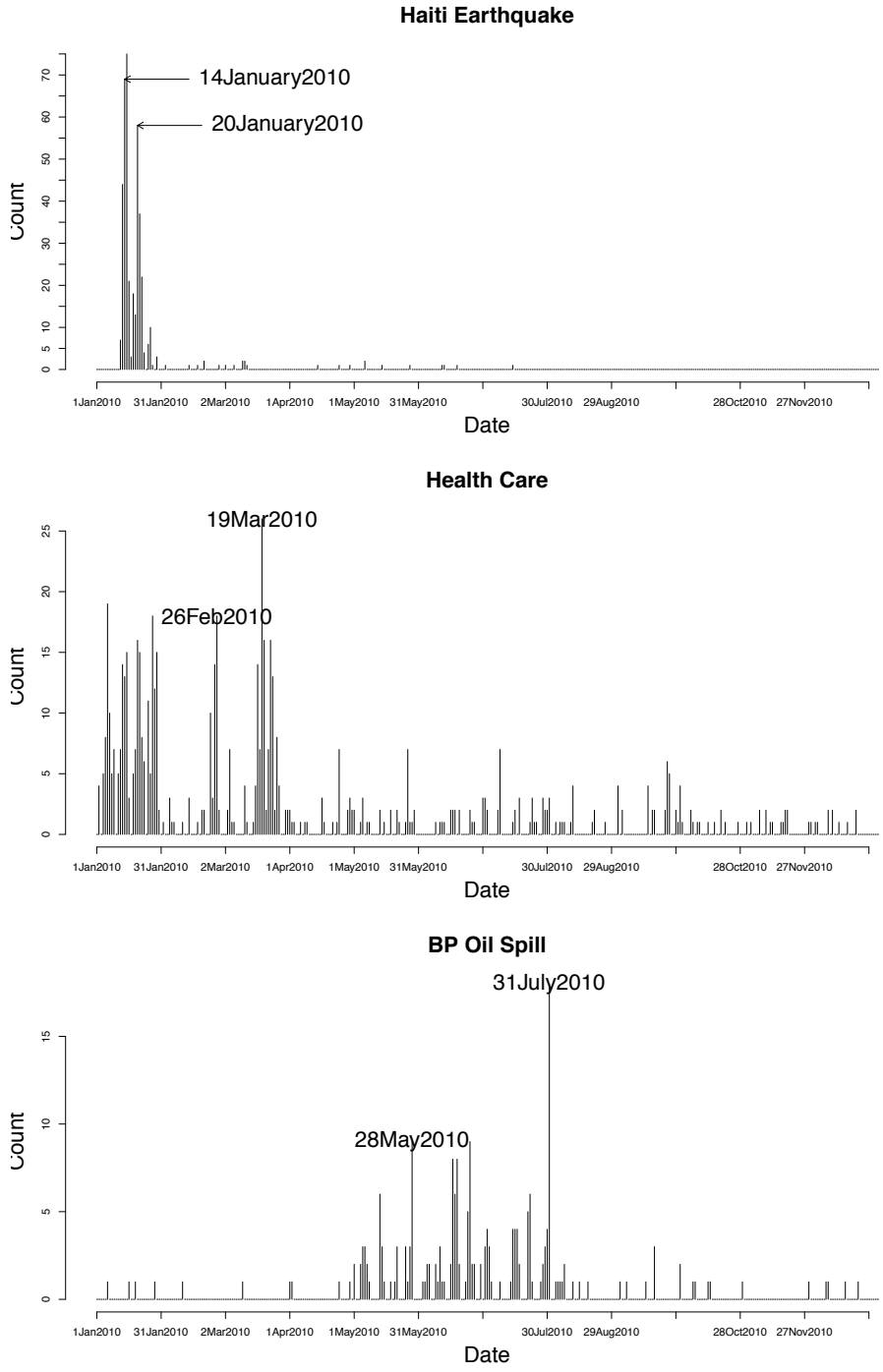
Visclosky, Rep. Pete. 2010. "Rep. Visclosky Announces Fifth Installment of FAA Funding for Gary/Chicago International Airport.".

Wallach, Hanna. 2008. Structured Topic Models for Language PhD thesis University of Cambridge.

Wallach, Hanna, Lee Dicker, Shane Jensen and Katherine Heller. 2010. An Alternative Prior for Nonparametric Bayesian Clustering. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 9.

Welling, M. 2006. Flexible Priors for Infinite Mixture Models. In *Workshop on Learning with Non-parametric Bayesian Methods*.

Figure 5: Overtime Variation in Cluster Attention



This figure provides one check of the conceptual validity of our clusters, first proposed in Quinn et al. (2010). The number of press releases in each set of clusters increase along with salient events.