# Structural topic models for open-ended survey responses[1]

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Gadarian, Bethany Albertson, David Rand[2]

This draft: May 30, 2013

[1]Our thanks to the Caltech SURF program, IQSS's Program on Text Analysis, and Dustin Tingley's Dean support for supporting Jetson's initial participation during the summer of 2012. Brandon Stewart gratefully acknowledges funding from a National Science Foundation Graduate Research Fellowship. Alex Storer helped get computers to do their job. We thank the following for helpful comments and suggestions: Neal Beck, Justin Grimmer, Jennifer Jerit, Luke Keele, Gary King, Mik Laver, Rose McDermott, Helen Milner, Rich Nielsen, Brendan O'Connor, Mike Tomz and participants in the Harvard Political Economy and Applied Statistics Workshops, UT Austin Government Department IR Seminar, Visions in Methodology 2013, and Stanford Methods Seminar.

[2]Working paper, send comments to: dtingley@gov.harvard.edu

**Abstract**

Despite broad use of surveys and survey experiments within political science, the vast majority of survey analysis deals with responses to options along a scale or from pre-established categories. Yet, in most areas of life individuals communicate either by writing or by speaking, a fact reflected in earlier debates about open and closed-ended survey questions. Collection and especially analysis of open-ended data is relatively rare in the discipline and when conducted is almost exclusively done through human coding. We present an alternative, semi-automated approach, the Structural Topic Model (STM) [Blinded For Review], that draws on recent developments in machine learning based analysis of textual data. A crucial contribution of the method is that it incorporates information about the document, such as the author's gender, political affiliation, and treatment assignment (if an experimental study). This paper focuses on how the STM is helpful for survey researchers and experimentalists. The STM makes analyzing open-ended responses easier, more revealing, and capable of being used to estimate treatment effects. We illustrate these innovations with analysis of text from surveys and experiments.
Word count: 8,500

# 1 Introduction

Despite broad use of surveys and survey experiments within political science, the vast majority of survey analysis deals with responses to options along a scale or from pre-established categories. Yet, in most areas of life individuals communicate either by writing or by speaking, a fact reflected in earlier debates about open and closed-ended survey questions. Collection and especially analysis of open-ended data is relatively rare in the discipline and when conducted is almost exclusively done through human coding. We present an alternative, semi-automated approach, the Structural Topic Model (STM) [Blinded For Review], that draws on recent developments in machine learning based analysis of textual data. A crucial contribution of the method is that it incorporates information about the document, such as the author's gender, political affiliation, and treatment assignment (if an experimental study). This paper focuses on how the STM is helpful for survey researchers and experimentalists. The STM makes analyzing open-ended responses easier, more revealing, and capable of being used to estimate treatment effects. We illustrate these innovations with several experiments and an analysis of open-ended data in the American National Election Study (ANES).

In practice, we believe that many survey researchers and experimentalists avoid open-ended response data because it is costly to analyze in a systematic way. There are also debates about the desirability of using open and closed-ended response formats. We provide relatively low-cost solutions that occupy a middle ground in these debates and innovate in two ways. First, we show how survey researchers and experimentalists can efficiently analyze open-ended data alongside a variety of common closed-ended data, like a subject's party preferences or assignment to an experimental condition. Second, we provide a suite of tools for survey researchers and experimentalists that enable preprocessing of textual data, model selection, and visualization. We also discuss best practices and tools for human intervention in what otherwise is a unsupervised learning model, such as how a researcher could implement pre-analysis plans, as well as a discussion of limitations to the unsupervised learning model[1] at the foundation of our research strategy.

[1] As opposed to supervised models which require a hand-coded training set, see Grimmer and Stewart

We proceed by first laying out the advantages and limitations of incorporating open-ended responses into research designs (Section 2). Next we present our estimation strategy and quantities of interest, as well as contrast our approach to existing methodologies (Section 3). Having set up our research strategy, we analyze open-ended data from a survey experiment on immigration preferences and a laboratory experiment on public goods provision (Section 4). We also analyze the "most important problem" data from the ANES. In each example we showcase both our methodology for including covariates as well as the software tools we make available to researchers. Finally we conclude with a discussion of future research possibilities (Section 5).[2]

# 2 Why open-ended responses?

There was a point at which research on survey methodology actively debated whether questions should be open or closed form (Geer, 1991; Lazarsfeld, 1944; Krosnick, 1999). That era is no more; the majority of survey analyses are composed predominately of closed-ended questions and open-ended questions rarely analyzed. This is despite the fact that prominent scholars writing on the topic identified advantages with each methodology (Lazarsfeld, 1944; Krosnick, 1999).

There are advantages and disadvantages to both closed and open-ended data. One view of open-ended responses is that they provide a direct view into a respondent's own thinking. For example, RePass (1971, p. 391) argues that open-ended questions query attitudes that "are on the respondent's mind at the time of the interview," attitudes that were presumably salient before the question and remain so afterwards. Similarly, Iyengar (1996, p. 64) notes that open-ended questions have the advantage of "nonreactivity." That is, unlike closed-ended questions, "open-ended questions do not cue respondents to think of particular causes or treatments."[3]

(2013) for details.

[2]Supplemental appendices include estimation details (Section 6.1), a comparison to alternative models (Section 6.2.1), a range of simulation studies (Section 6.2.5), and tools for applied users (Section 6.3).

[3]On this point, Kelley (1983, p. 10) notes that the opinions of the American electorate are so wide ranging that any closed list is bound to omit good opinions.

A major concern about open-ended questions is that open-ended questions chiefly require that subjects "articulate a response, not their underlying attitudes" (Geer, 1988, p. 365). Furthermore, non-responses to open-ended questions may stem from ineloquence rather than indifference; subjects may not respond to open-ended questions because they lack the necessary rhetorical device (Geer, 1988). A related concern is that open-ended questions may give respondents too little of a frame of reference in order to form a coherent response (Schuman, 1966).

Open-ended responses have traditionally been considered more difficult to analyze than their closed counterparts (Schuman and Presser, 1996), as human coding is almost always used. The use of human coders typically involves several steps. First, researchers need to define the dimensions on which open-ended data will be coded by humans and generate examples in order to guide the coders. This is typically guided by the researcher's own prior theoretical expectations and potentially reading of some examples. Next human coders are unleashed on the data and numerical estimates for each document compared across coders (Lombard *et al.*, 2006; Artstein and Poesio, 2008).

Our view is that while such pragmatic concerns are reasonable, they ought not be our ultimate consideration and instead what is crucial is whether open-ended questions give real insights (Geer, 1991, pg. 360). Rarely have survey researchers/experimentalists used automated text analysis procedures and when they have, covariate information, either in the form of randomized treatment conditions or pre-treatment covariates (e.g., gender or political ideology), is not used in the textual analysis (Simon and Xenos, 2004). Researchers still might have good reason to use human coders, but we believe adoption of our methods will only assist them in using them more effectively.

## 2.1 Our Contributions

The model below has a number of advantages over only using human coders. First, it allows the researcher to *discover* topics from the data, rather than assume them. These topics may or may not correspond to a researcher's theoretical expectations. When they do correspond, researchers can leverage the wide variety of quantities of interest that the STM generates. When they do not correspond, researchers may consider revising their

theoretical model for future work, or retain their model and turn to standard human coding procedures.

Second, it allows analysts to do this while studying how the prevalence and content of topics change with information that is particular to each respondent; for example, whether the respondent received the treatment or background demographic data. We argue our model can fruitfully be used at either an exploratory stage prior to using human coders or as part of making causal inferences about the effect of treatments/frames/covariates on topics. Thus our approach can serve a variety of purposes. The next sections demonstrate the usefulness of text analysis tools for analyzing open-ended responses.

# 3 Statistical Models of Text

The core innovation of the paper is to bridge survey and experimental techniques, which include randomization of frames or encouragements to adopt a particular emotional status or way of looking at political issues, with new techniques in text analysis. Our approach allows the analyst to incorporate covariates (e.g. attributes of the respondent, treatment condition), with a model of the topics that are inferred directly from the written text. Crucially for experimental applications, this enables us to calculate treatment effects, and uncertainty estimates, on open-ended textual data. We believe that we are the first to do so in a way that builds in the structural information about the experiment, though we share similar motivations to Simon and Xenos (2004) and Hopkins (2012). In this section we outline the notation and core ideas for statistical topic models, then we overview the structural topic model (STM) including quantities of interest and conclude by providing an overview of material available in the supplemental appendix.

## 3.1 A Heuristic Understanding of Statistical Topic Models

Statistical topic models allow for rich latent topics to be automatically inferred from text. Topic models are often referred to as "unsupervised" methods because they *infer* rather than *assume* the content of the topics under study, and have been used across a variety fields (Quinn *et al.*, 2010; Grimmer, 2010; Blei *et al.*, 2003; Wang and Blei, 2011). We emphasize that this is conceptually different from "supervised" methods where the analyst

defines the topics *ex ante*, usually by hand-coding a set of documents into pre-established categories (e.g., Laver *et al.*, 2003).

Within the class of unsupervised statistical topic models, topics are defined as distributions over a vocabulary of words which represent semantically interpretable "themes." Topic models come in two varieties: *single-membership* models and *mixed-membership* models. Previous work in political science has focused on single-membership models which have emphasized document meta-data (Quinn *et al.* (2010); Grimmer (2010), see also Grimmer and Stewart (2013) for a general review). In mixed-membership models, the most notable of which is Latent Dirichlet Allocation (Blei *et al.*, 2003; Blei, 2012), a document is represented as a mixture of topics, with each word belonging to exactly one topic; thus, each document can be represented as a vector of proportions that denote what fraction of the words belong to each topic. In single-membership models, each document is restricted to only one topic, so all words within it are generated from the same distribution. We focus on mixed-membership models, highlighting the comparison to single-membership alternatives in the appendix (Section 6.2.3).

In mixed-membership models, each document (indexed by $d$) is assumed to be generated as follows. First a distribution over topics ($\theta_d$) is drawn from a global prior distribution. Then for each word in the document (indexed by $n$), we draw a topic for that word from a multinomial distribution based on its distribution over topics ($z_{d,n} \sim \text{Mult}(\theta_d)$). Conditional on the topic selected, the observed word $w_{d,n}$ is drawn from a distribution over the vocabulary $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$ where $\beta_{k,v}$ is the probability of drawing the $v$-th word in the vocabulary for topic $k$. So for example, our article (the one you are reading), which is just one article among all journal articles ever written, might be represented as a mixture over three topics which we might describe as: survey analysis, text analysis and experiments. Each of these topics is actually a distribution over words with high frequency words associated with that topic (e.g. the experiments topic might have "experiment, treatment, control, effect" as high probability words).

In Latent Dirichlet Allocation (LDA), the model described above is completed by assuming a Dirichlet prior for the topic proportions such that: $\theta_d \sim \text{Dirichlet}(\alpha)$. Esti-

mation proceeds by variational expectation-maximization (EM) where the local variables $\theta_d, \vec{z_d}$ are estimated for each document in the E-step, followed by maximization of global parameters $\alpha, \beta_{1:K}$ which are the parameters describing global topic proportions ($\alpha$) and the $K$ distributions over words which define the topics ($\beta_k$).[4]

The expressive power of statistical topics models to discover topics comes at a price. The resulting posterior distributions have many local modes, meaning that different initializations can produce different solutions. This can arise even in simple mixture models in very low-dimensions (Sontag and Roy, 2009; Buot and Richards, 2006; Anandkumar *et al.*, 2012). In Section 3.4 we present a framework for model evaluation focused on semantic interpretability as well as robustness checks.

## 3.2 Structural Topic Model

The Structural Topic Model (STM) innovates on the models just described by allowing for the inclusion of covariates of interest into the prior distributions for document-topic proportions and topic-word distributions. The result is a model where each open-ended response is a mixture of topics. Rather than assume that topical prevalence and content are constant across all participants, the analyst can incorporate covariates over which we might expect to see variance.

We explain the core concept of the model here (complete details in appendix 6.1). As in LDA, each document arises as a mixture over $K$ topics. In the STM, topic proportions ($\theta$) can be correlated, and the prevalence of those topics can be influenced by some set of covariates $X$ through a standard regression model with covariates $\theta \sim \text{LogisticNormal}(X\gamma, \Sigma)$. For each word ($w$) in the response, a topic ($z$) is drawn from the response-specific distribution, and conditional on that topic a word is chosen from a multinomial distribution over words parametrized by $\beta$ which is formed by deviations from the baseline word frequencies ($m$) in log space ($\beta_k \propto \exp(m + \kappa_k)$). This distribution can include a second set of covariates $U$ (allowing, for example, Democrats to use the word "estate" more frequently than Republicans while discussing taxation). We discuss the difference between the two sets of covariates in more detail in Section 3.3.

[4]Variational EM uses a tractable factorized approximation to the posterior. See Grimmer (2011).

Thus there are three critical differences in the STM model as compared to the LDA model described above: (1) topics can be correlated, (2) each document has its own prior distribution over topics, defined by covariates $X$ rather than sharing a global mean, and (3) word use within a topic can vary by covariate $U$. These additional covariates provide a way of "structuring" the prior distributions in the topic model, injecting valuable political information into the inference procedure.

Like LDA, estimation of the STM uses variational EM. In the variational E-step we loop through each document to infer the proportion of words in the document attributable to each topic ($\theta$) and the assignment of each word ($z$). Then in the M-step, we infer the global parameters $\kappa, \gamma, \Sigma$ which control the priors on topical prevalence and content.[5]

The STM provides fast, transparent, replicable analyses that require few *a priori* assumptions about the texts under study. Yet it is a computer-*assisted* method, and the researcher is still a vital part of understanding the texts as we describe in the examples section. The analyst's interpretive efforts are guided by the model and the texts themselves. But as we show the STM can relieve the analyst of the burden of trying to develop a categorization scheme from scratch (Grimmer and King, 2011) and perform the often-mundane work of associating the documents with those categories.

## 3.3 Estimating Quantities of Interest

A central advantage to our framework for open-ended survey response analysis is the variety of interpretable quantities of interest beyond what is available from LDA. In all topic models, the analyst estimates for each document the proportion of words attributable to each topic, providing a measure of topic *prevalence*. The model also calculates the words most likely to be generated by each topic, which provides a measure of topical *content*. However in standard LDA the document collection is assumed to be unstructured; that is each document is assumed to arise from the same data generating process irrespective of additional information the analyst might possess. By contrast, our framework is designed to incorporate additional information about the document or its author into the estima-

[5]The STM prior is not conjugate to the likelihood and thus does not enjoy some of the theoretical guarantees associated with mean-field variational inference in the conjugate exponential family.

tion process. This allows us to measure systematic changes in topical prevalence and topical content over the conditions in our experiment, as measured by the $X$ covariates for prevalence, and the $U$ covariates for content. Thus we can easily obtain measures of how our treatment condition affects both how often a topic is discussed (prevalence) and the language used to discuss the topic (content).

The inference on the STM quantities of interest is best understood by reference to the familiar regression framework. For example, consider topical prevalence; if we observed the topics for each survey response, we could generate a regression where the topic is the outcome variable, and the treatment condition or other respondent controls (such as gender, income, party affiliation), along with any interactions are the explanatory variables. This regression would give us insight into whether our treatment condition caused respondents to spend a larger portion of their written response discussing a particular topic. In our framework for analysis, we conduct this same regression, while simultaneously estimating the topics. This framework builds on recent work in political science on single-membership models, specifically Quinn *et al.* (2010) and Grimmer (2010), which allow topical prevalence to vary over time and author respectively. Our model extends this framework by allowing topical prevalence to vary with *any* user specified covariate. We also extend the framework to topical content. Word use within a particular topic comes from a regression, in this case a multinomial logistic regression, where the treatment condition and other covariates can change the rate of use for individual words within a topic.

In addition to these corpus-level changes, we also get an estimate of the proportion of words in each survey response attributable to a particular topic. Thus, we can retrieve the same types of quantities that would arise from human coding without the need to construct a coding scheme in advance. These document-level parameters can be used to construct useful summaries such as most representative documents for each topic, most representative documents for each treatment condition, or variation in topic use across other covariates not in the model.

We can also use the model to summarize the semantic meaning of a topic. Generally

these summaries are the highest probability words within a topic, however this tends to prioritize words that have high frequency overall but may not be semantically interesting. Following the insights of Bischof and Airoldi (2012) who demonstrate the value of exclusivity in summary words for topics, we label topics using simplified Frequency-Exclusivity (FREX) scoring [Blinded For Review]. This summarizes words with the geometric mean of the probability of appearance under a topic and the exclusivity to that topic. We find that these words provide more semantically intuitive representations of topics.

In Figure 1 we list some of the quantities of interest with a simple interpretation. These quantities can be combined to create more complex aggregates, but we expect these summaries will suffice for most applications.

## 3.4 Model Specification and Selection

Researchers must make important model specification and selection decisions. We briefly discuss the choice of covariates and the number of topics. We discuss theoretical implications of model specification choices, quantitative metrics, and methods for semi-automated model evaluation and selection.[6]

### 3.4.1 Choices in Model Specification

In the STM framework, the researcher has the option to choose covariates to incorporate in the model. These covariates inform either the topic prevalence or the topical content latent variables with observed information about the respondent. The analyst will want to include a covariate in the topical prevalence portion of the model ($X$) when they believe that the observed covariate will affect *how much* the respondent is to discuss a particular topic. The analyst also has the option to include a covariate in the topical content portion of the model ($U$) when they believe that the observed covariate will affect the words which a respondent uses to discuss a particular topic. These two sets of covariates can overlap, suggesting that the topic proportion and the way the topic is discussed change with particular covariate values. The STM model includes shrinkage priors or regularization,

[6]We use standard text pre-processing conventions, such as stemming (Manning *et al.*, 2008). The appendix provides complete details along with software to help the user manage and pre-process their collections of texts (Section 6.3.1).

1. QOI: Topical Prevalence Covariate Effects
    - Level of Analysis: Corpus
    - Part of the Model: $\theta, \gamma, X$
    - Description: Degree of association between a document covariate $X$ and the average proportion of a document discussing each topic.
    - Example Finding: Subjects receiving the treatment on average devote twice as many words to Topic 2 as control subjects.
2. QOI: Topical Content Covariate Effects
    - Level of Analysis: Corpus
    - Part of the Model: $\kappa, U$
    - Description: Degree of association between a document covariate $U$ and the rate of word use within a particular topic.
    - Example Finding: Subjects receiving the treatment are twice as likely to use the word "worry" when writing on the immigration topic as control subjects.
3. QOI: Document-Topic Proportions
    - Level of Analysis: Document
    - Part of the Model: $\theta$
    - Description: Proportion of words in a given document about each topic.
    - Example Use: Can be used to identify the documents which devote the highest or lowest proportion of words to a particular topic. Those with the highest proportion of words are often called "exemplar" documents and can be used to validate that the topic has the meaning the analyst assigns to it.
4. QOI: Topic-Word Proportions
    - Level of Analysis: Corpus
    - Part of the Model: $\kappa, \beta$
    - Description: Probability of observing each word in the vocabulary under a given topic. Alternatively, the analyst can use the FREX scoring method described above.
    - Example Use: The top 10 most probable words under a given topic are often used as summary of the topic's content and help inform the user-generated label.

Figure 1: Quantities of Interest from STM

which draws the covariate effects towards zero. An analyst concerned about overfitting to the covariates can increase the degree of regularization.

The analyst must also choose the number of topics. There is no "right" answer to this choice. Varying the number of topics varies the level of granularity of the view into the data. Therefore, the choice will be dependent both on the nature of the documents under study and the goals of the analysis. While some corpora like academic journal articles might be analyzed with 50-100 topics (Blei, 2012) due to the wide variety in their content, survey responses to focused questions may only consider a few topics. The appropriateness of particular levels of aggregation will vary with the research question.

### 3.4.2 Model Selection Methods

It would be useful if all of these choices could be evaluated using a simple diagnostic. It is tempting to compute an approximation to the marginal likelihood and calculate a model selection statistic, but we echo previous studies in emphasizing that this maximizes model fit and not substantive interpretation (Chang *et al.*, 2009). Instead we advocate quantitative evaluations of properties of the topic-word distributions. Specifically, we argue that a semantically interpretable topic has two qualities: (1) it is *cohesive* in the sense that high probability words for the topic tend to co-occur within documents, and (2) it is *exclusive* in the sense that the top words for that topic are unlikely to appear under other topics.

These two qualities are closely related to Gerring (2001)'s "consistency" and "differentiation" criteria for concepts in empirical social science.[7] Semantic cohesion has previously been studied by Mimno *et al.* (2011) who develop a criterion based on co-occurence of top topic words and show that it corresponds with human evaluation by subject matter experts.[8] While semantic coherence is a useful criterion, it only addresses whether or not

[7]These qualities also appear in the evaluation of single-membership clustering algorithms (Jain, 2010). We speculate these qualities are implicitly central to many conceptual paradigms both in quantitative as well as qualitative political science.

[8]Newman *et al.* (2010) first proposed the idea of using point-wise mutual information to evaluate topic quality. Mimno *et al.* (2011) then proposed a closely related measure which they named *semantic coherence* demonstrating that it corresponded with expert judgements of NIH officials on a corpus of NIH

a topic is internally consistent, but does not, for example, penalize topics which are alike. From the standpoint of social science inference, we want to be sure that we are evaluating a concept that is both well defined and our measure captures all incidence of the concept in the survey responses.

For this we turn to the *exclusivity* of topic words, drawing on previous work on exclusivity and diversity in topic models (Eisenstein *et al.*, 2011; Zou and Adams, 2012; Bischof and Airoldi, 2012). If words with high probability under topic $i$ have low probabilities under other topics, than we say that topic $i$ is exclusive. A topic which is both cohesive and exclusive in all likelihood is likely to be semantically useful.

In order to select an appropriate model, we generate a set of candidate models (generated by differing initializations, tuning parameters, or processing of the texts) and then discard results which are below the 75% quantile (which can be set by the researcher) in either semantic coherence or exclusivity. We then either randomly select a model or manually examine the remainder and select the model most appropriate to our particular research question. We provide methods for calculating exclusivity and semantic cohesion with our estimation software.

While this simple measure is computationally efficient and interpretable, it cannot replace human judgement. The insight of the investigator is paramount here and we strongly suggest careful reading of example texts. In these cases, the STM can direct the reader to the most useful documents to evaluate by providing a list of exemplar texts for each topic. An intermediate step between automated diagnostics and judgement of the principal investigator is to use human evaluations on tasks for cluster quality. Chang *et al.* (2009) and Grimmer and King (2011) describe human evaluation protocols for testing topic quality which can easily be applied to our setting. Of course, researchers are free to not use these selection methods, or create other methods. Researchers might also incorporate pre-analysis plans, which specify sets of words they expect to appear together in topics of interest and select based upon those criteria.

grants as well as human judgements gathered through Amazon's Mechanical Turk.

## 3.5 Validating the Model: Simulations Tests and Examples

When introducing any new method, it is important to test the model in order to validate that it performs as expected. Specifically, we were driven to answer two critical questions about the performance of the structural topic model:

1. Does the model recover treatment effects correctly (i.e. low false-positives and low false-negatives)?

2. How does analysis compare to first estimating topics with LDA, then relating the topics to covariates?

In the supplemental appendix (Section 6) we address both of these questions in turn using a battery of tests that range from Monte Carlo experiments on purely simulated data through applied comparisons on the examples presented in Section 4. Here we briefly address each question providing an overview of our simulations and deferring the details to the appendix.

As shown in Figure 2 the model recovers the effect of interest when it exists, and does not induce a spurious effect when the effect is actually zero (false positives). A separate, but related, concern is the effects of multiple testing. While our simulation results demonstrate that STM does not systematically overestimate treatment effects, it does not address concerns of accurate $p$-values in the presence of multiple testing. In Section 6.3.2 of the appendix we discuss how false discovery rate methods and pre-experiment plan approaches can be incorporated into the topic model framework to address these concerns. In Section 6.2.6 we show results of a permutation test on one of our applied examples. In this test we randomly permute the treatment variable across documents and refit the model, showing that we do not find spurious treatment effects.

### 3.5.1 Comparison to LDA and other Alternate Models

Statistical methods for the measurement of political quantities from text have already seen widespread use in political science, and the number of available methods is growing at a rapid rate. How does analysis with the STM compare to existing unsupervised models?

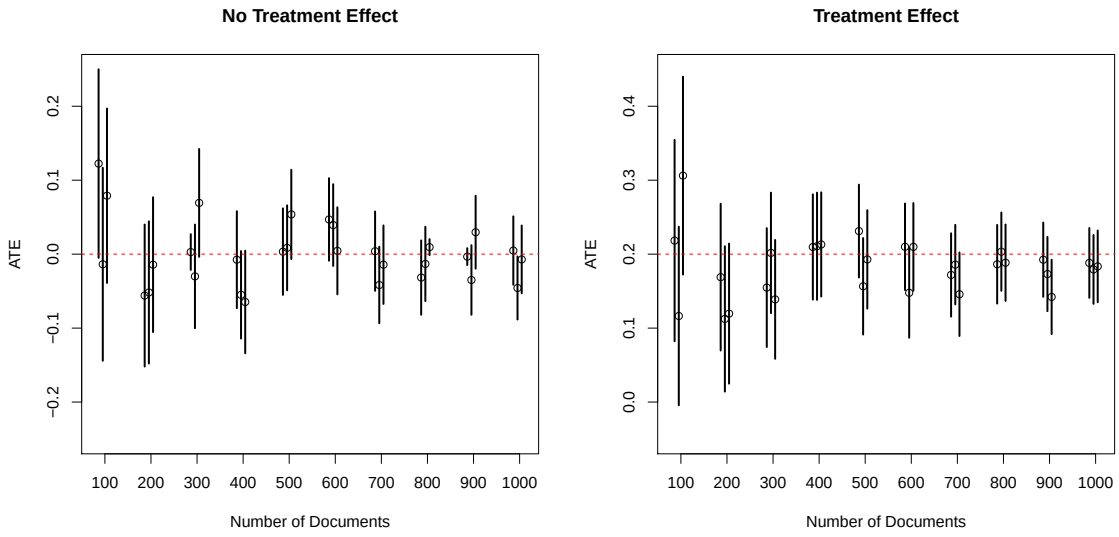



Figure 2: Estimated average treatment effect with 95% confidence intervals holding expected number of words per document fixed at 40, and the concentration parameter fixed at 1/3. The STM is able to recover the true ATE both in cases where there is no treatment effect (left) and cases with a sizable treatment effect (right). As expected, inferences improve as sample sizes increase.

In the appendix (Section 6.2.1), we contrast our approach with three prominent alternative models in the literature, focusing on the advantages of including covariates. While no single method will be superior for all applications, we argue that the STM framework is particularly applicable to the analysis of open-ended survey responses. Specifically we contrast the benefits of the STM with 'vanilla' Latent Dirichlet Allocation (Blei, 2012), factor analysis (Simon and Xenos, 2004), and single-membership models (Quinn *et al.*, 2010; Grimmer, 2010). Both LDA and factor analysis provide the mixed-membership structure, which allows responses to discuss multiple topics, but cannot incorporate the rich covariate information we often have available, while the single-membership models developed in political science can incorporate a narrow set of covariate types and are limited to a single topic per document which may be too restrictive for our application. Compared to other unsupervised techniques, we believe the STM provides the most versatility for survey researchers and experimentalists.

In Section 6.2.1 we provide an extensive comparison to LDA which shows that the STM provides more accurate estimation of quantities of interest when compared to using LDA with covariates in a two-stage process. We show Monte Carlo simulations which are consistent with the theoretical expectation that LDA will tend to attenuate continuous covariate relationships on topical prevalence. Figure 3 shows one such simulation for the case a continuous covariate that operates differently under the treatment and control condition. LDA is unable to capture the dynamics of the effect in many of the simulated datasets. Section 6.2.1 also overviews some diagnostics for LDA models which indicate when the inclusion of additional structure as in the STM is useful for inference. Finally, we provide an analysis of actual documents from the immigration experiment discussed below using LDA and characterize the differences between those solutions and the ones attained by the STM.

A comparison to supervised methods would be beyond even the scope of our appendix. Supervised methods provide a complement to unsupervised methods and can be used when an analyst is interested in a specific, known quantity in the text. Thus supervised methods can be seen as occupying a place on the spectrum between close-ended questions

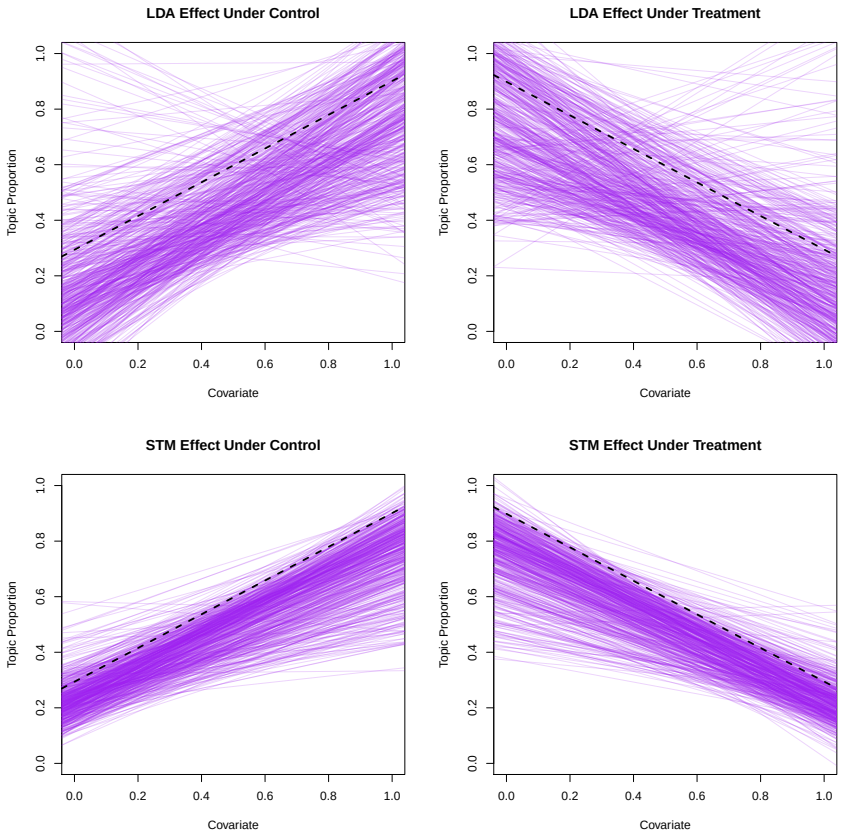



Figure 3: STM vs. LDA recovery of treatment effects. Each line represents the estimated effect from a separate simulation, with the bold line indicating the true data generating process. While the two stage LDA process often captures the approximate effect, it exhibits considerably higher variance.

which provide an *a priori*, analyst-specified assessment of the quantities of interest, and the unsupervised analysis of open-ended response which provide a data driven assessment of the quantities of interest with *post hoc* analyst assessment. We provide an implicit comparison to supervised approaches by comparing our unsupervised methods to human coders in Sections 4.3 and 6.2.8.

### 3.5.2 Additional Material

The appendix provides a number of additional details which we split into three major sections:

1. Model Estimation (6.1) gives details on the variational Expectation-Maximization based approach to optimization of the model parameters.

2. Model Validation tests (6.2) includes simulations mentioned above as well myriad other validations.

3. Getting Started (6.3) overviews two additional software tools that we provide (`txtorg`, a tool for preprocessing and handling large bodies of text, and a topic visualization tool for helping users browse their documents and assess model results) and discussions of common questions that might arise and how they connect to the topic modeling framework, including multiple testing (Section 6.3.2), mediation analysis (Section 6.3.4), and pre-analysis plans (Section 6.3.1).

# 4 Data Analysis

The purpose of this section is to illustrate the application of the method to actual data. Our goal is to show how to estimate the relationships between covariates and topics with corresponding uncertainty estimates, how to interpret model parameters, and how to automatically identify passages that are the best representations of certain topics. To illustrate these concepts we rely on several recent studies that recorded open-ended text as well as recently released data from the ANES.

## 4.1 Public views of immigration

Gadarian and Albertson (2013) examine how negatively valanced emotions influence political behavior and attitudes. In one of their surveys, they focus on immigration preferences by using an experimental design that in the treatment encourages some subjects to become worried about immigration and in control to simply think about immigration. To categorize these open-ended responses they turned to human coders who were instructed to code each response along the dimensions of enthusiasm, concern, fear, and anger, each along a three point scale.

### 4.1.1 Topic Analysis

To estimate the STM we use an indicator variable for the treatment condition, 7-point party identification self-report, and an interaction between party identification and treatment condition as covariates. The interaction term lets us examine whether individuals who are Republican respond to the treatment condition differently from those who are Democrats. In this particular application, the influence of these parameters was estimated on topic proportions or the prevalence of topics within responses. To address multi-modality, we estimated our model 200 times, with 200 different starting values and applied the model selection procedure described in Section 3.4. This left us with nine models, from which we randomly selected one. However, a close examination of these nine models indicate that all have very similar results in terms of the topics discovered and differences in topic proportions across treatment conditions.

We estimated three topics in total in our analysis. The two topics most associated with the treatment and control groups, respectively, are presented in Figure 4. The first topic is the "crime" and "welfare" or "fear" topic and the second topic is a much more neutral topic, stressing citizenship, and the difficulties faced by immigrants. To get an intuitive sense of the topics Figures 5 and 6 plot representative responses[9] for topics 1 and 2.

[9]The predicted probability of that response being in the given topic is high relative to other responses within the corpus.

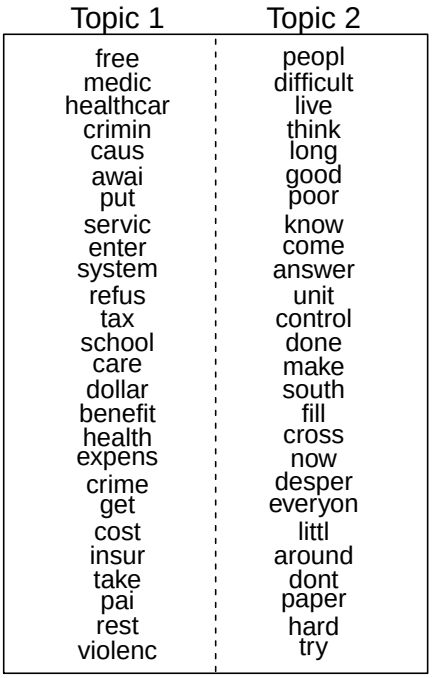



Figure 4: Vocabulary Associated with Topic 1 and 2

" problems caused by the influx of illegal immigrants who are crowding our schools and hospitals, lowering the level of ducation and the quality of care in hospitals."

Figure 5: A Representative Response from Topic 1

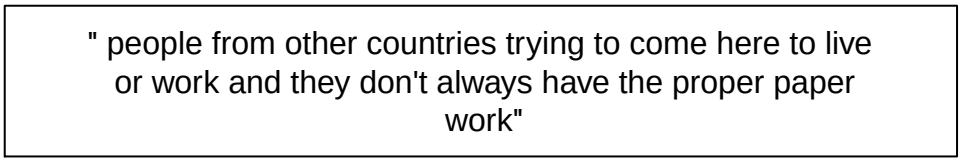



Figure 6: A Representative Response from Topic 2

### 4.1.2 Covariate Analysis

Next we move to differences across the treatment groups. On average, the difference between the proportion of a treated response that discusses Topic 1 and the proportion of an untreated response that discusses about Topic 1 is .27 (.21, .32).[10] This shows that the study's encouragement to express worries about immigration was effective. In addition, on average over both treatment and control, Republicans talk about fear and anger toward immigrants much more than Democrats do: by our estimates, the difference between the proportion of a Republican response that talked about Topic 1 and proportion of a Democrat response that talked about Topic 1 was .30 (.23, .37).

The ability to estimate moderating effects on the treatment/control differences is a key contribution of our technique. The interaction between party ID and treatment also heavily influences topics. The difference between the proportion of a treated Republican response that talked about Topic 1 and the proportion of an untreated Democrat response that talked about Topic 1 is very large: .50 (.42,.59). What does this mean? An untreated Democrat will talk about Topic 1 29% of the time and Topic 2 33% of the time. A treated Republican will talk about Topic 1 80% of the time and Topic 2 8% of the time.[11]

Topic proportions by these two covariates are displayed graphically in Figure 7. The first plot in Figure 7 shows a treatment effect of response proportions in Topics 1 and 2, comparing treated to untreated. The second plots a Loess-smoothed line of the proportion of each response in Topic 1 on party identification, where 7 is strong Republican and 0 is a

[10]Estimates within the parentheses represent a 95% confidence interval.

[11]Words which don't fall into Topic 1 or Topic 2 fall into Topic 3, a topic which includes general immigration words, such as "immigration", "illegal", and "america".

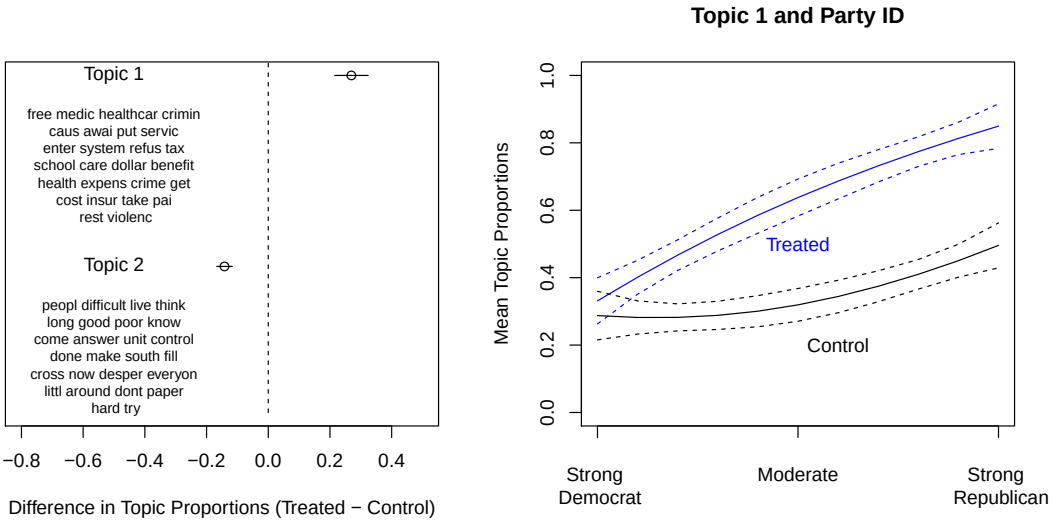



Figure 7: Party ID, Treatment, and the Predicted Proportion in Topic 1

strong Democrat. In general, these accord with our expectations about how the treatment and party identification should be associated with the responses.

### 4.1.3 Aggregate Comparison with Human Coders

The traditional way text has been analyzed in survey or experimental settings is to have human coders code each response based on a set of coding instructions. Fortunately in this example Gadarian and Albertson (2013) did just this, using two research assistants. How do our results compare with those of the coders? The comparison between the hand-coding and the results from our algorithm is described in detail in the 6.2.8. In summary, the results from STM and the hand-coding are similar and both methods find a treatment effect. In addition, there is significant correlation between the hand coding of individual responses and the predicted topic proportions from our unsupervised learning model.

Of course, since our model is unsupervised, the topics discovered by our model do not perfectly match the topics the coders were instructed to use. The coders categorized the vast majority of responses into fear and anger, but because the topic model by design tries to distinguish between documents, its definition of topics do not align directly with fear and anger, and some documents with a low proportion of Topic 1 from our analysis are also

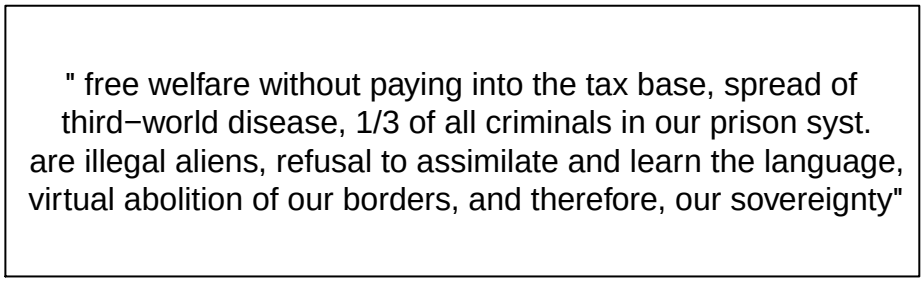



Figure 8: Fearful Response with High Topic 1

hand-coded with fear and anger. We would expect that the documents with low predicted proportion of Topic 1, but hand-coded as fear and anger would have fewer characteristics associated with Topic 1, for example, they might talk about crime and social security less relative to other reasons for being fearful or angry and immigrants. Figure 8 presents a document that has a high predicted proportion of Topic 1 that the coders both agree include fear or anger while Figure 9 presents a response with low relation to Topic 1 but still coded to be of high fear.

It is clear from these responses that both are angry at or fearful of illegal immigrants, but the reasoning behind their emotion is different. In addition, these two may have a very different view of legal immigration in general. One advantage of the topic model is even if the overwhelming majority people are either fearful of or angry at illegal immigrants, it will refine the topics in order to distinguish between documents, so even if a category pre-determined by the researcher applies to almost all responses, the topic model will find a finer distinction between them.

## 4.2 Intuition versus Reflection in Public Goods Games

Rand *et al.* (2012) study how intuitive versus reflective reasoning influences decision making in public goods games using a number of experimental conditions. In the "free-write" experimental contrast, subjects were primed to consider a time when they have acted out of intuition in a situation where their action worked out well or a time when they reflected and carefully reasoned in a situation where their action worked out well. After

" as an arizona resident who lives 18 miles from the mexican−us border, and who has also spoken to some of these illegals while hiking in the huachuca mtns., i know these people, mostly, come here out of sheer desperation. sure, some are the same lazy, fat, undereducated jerks that lurk around our own mid−level businesses. but most simply are people who want what we all do: a comfortable life with as little thinking and suffering as possible, while reproducing at will. they have told me, babies in arms,that if they remain at home, they have no future but an early death. that they, maybe, should reduce their birth rate and/or not have children at all, if they cannot support them, simply will never occur to citizens of a catholic country, living a day's walk from a rich country that can be easily milked for what they consider a fortune in life support. there is no answer to this, so long as 95% of mexico's wealth is controlled by 5% of its people, and the only riches the others have lie in their children."

Figure 9: Fearful Response with Low Topic 1

this encouragement, everyone played a single, incentivized, one-shot public goods game. In the "time" experimental contrast subjects were either forced to make a decision quickly or encouraged to take their time, after which all players participated in the same public goods game. Rand *et al.* (2012) find that subjects contribute more under the treatments where subjects are primed for intuition or are under time pressure, concluding that cooperation is intuitive. After both the free-write and time experiments, subjects were asked to write about the strategy they used while playing the public goods game. We analyze the players' descriptions of their strategies and their relationship to game contributions.

### 4.2.1 Decision explanations across treatment conditions

We contrast the topics present in the strategy descriptions across the different treatment conditions. The topic model reflects how the experimental conditions influence strategy descriptions. In the "free-write" experimental contrast, respondents primed to think intuitively talk about their strategy very differently than those who received the reflection priming. Listed in Figure 10, Topic 1 is associated with the intuitive priming, and Topic 2 is associated with the reflection priming. Topic 1 has words that reflect intuition, for example "mood", "good", "hope", "felt", and "believ". Topic 2, on the other hand, includes words such as "decis", "interest", "highest" and "self". The estimated topical difference between the two treatments are shown in Figure 11.

In the "time" experimental contrast, people who are given less time to think about their decision in the public goods game use more feeling and trusting words to describe their decisions, as shown below in Topic 1 of Figure 12. Words in this topic reflect concern over morality and feeling, with reference to words like "believe", "feel", and "god". In contrast, people who are given more time to think about their decision of whether or not to contribute use a more calculating vocabulary to describe their decision, with words like "receiv", "game", and "thought". This topic is shown below in Topic 2 of Table 12. The treatment effect for both of these topics is shown in Figure 13.

The topics in the intuition priming experiment and the time pressure experiment show some similarities. Topic 1 in both cases uses words associated with feeling and trusting. Topic 2 in both cases is more about thinking, maximizing payoff, and making


Figure 10: Topics from Intuition vs. Reflection Priming

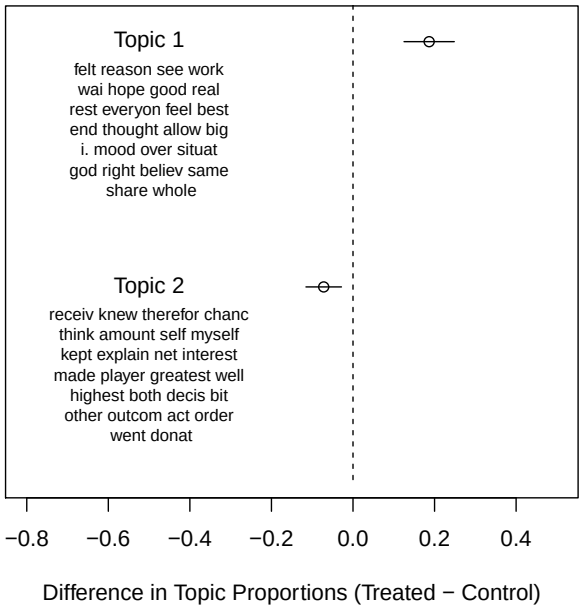



Figure 11: Intuition Treatment Effect on Topics

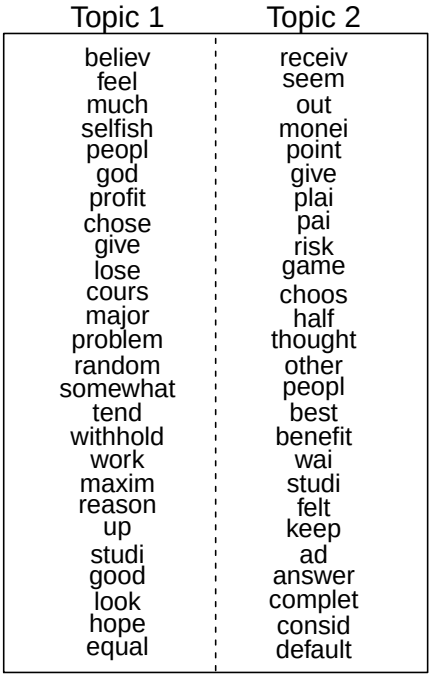



Figure 12: Topics from Time Pressure Experiment

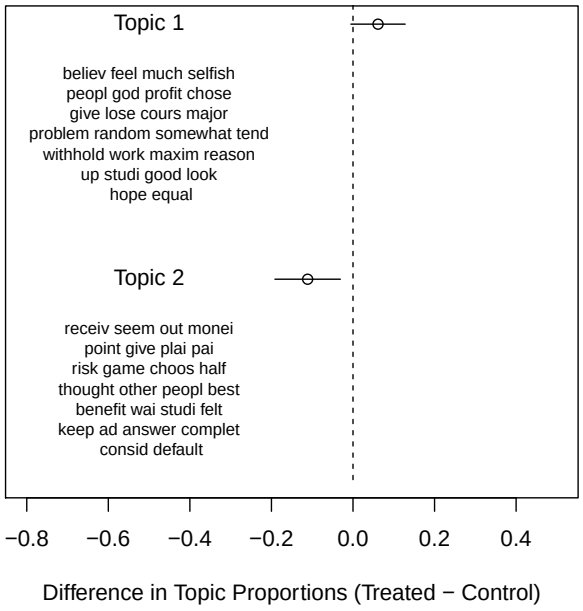



Figure 13: Time Pressure Treatment Effect on Topics

choices. These results show a nice coherence in the experimental design and shows how topic models can directly connect to the theoretical model, where the intuitive-cooperation theory expects that these exact distinctions to be important.

### 4.2.2 Respondents Who Talk About Their Intuition Cooperate More

We also examine the relationship between references to a topic and contributions in the game. As Rand *et al.* (2012) find, forcing people to think quickly increases contributions, and priming people to think intuitively also increases contributions. We expect therefore that people who talk about intuition, or whose responses are more in line with Topic 1, will also contribute more, but that people who talk about strategy and maximizing their profits, more in line with Topic 2, will contribute less.

In both cases, respondents with a higher predicted proportion of Topic 1 in their response are more likely to contribute. Figure 14 and Figure 15 plot a Loess smoothed line of contributions plotted on the predicted topic proportions for each document. For each of the experiments, responses with a higher predicted proportion of Topic 1 have overall higher contribution. However, as the predicted proportion of Topic 2 increases, the overall level of contributions falls. Therefore, people who talk more about intuition, trust, and their feelings, are more likely to contribute. People who talk more about strategy and maximizing profits are less likely to contribute.

### 4.2.3 Using Covariates for the Vocabulary: Gender

The topic model is not only able to assess the influence of covariates on the topic proportions, it is also able to use covariates to show how different types of respondents use different vocabulary to talk about the same topic. A researcher might be interested in how women talk about their strategy when primed with intuition, compared to how men talk about their strategy when primed with intuition. Figure 16 shows a word cloud of the intuition topic where the word size represents the frequency with which a word is used. We find that men talk about their intuition in terms of their own feelings, while women describe their intuition in terms of their morality. On the left side of the plot are words that men use more frequently within the intuition topic, including “trust”, “good”, and

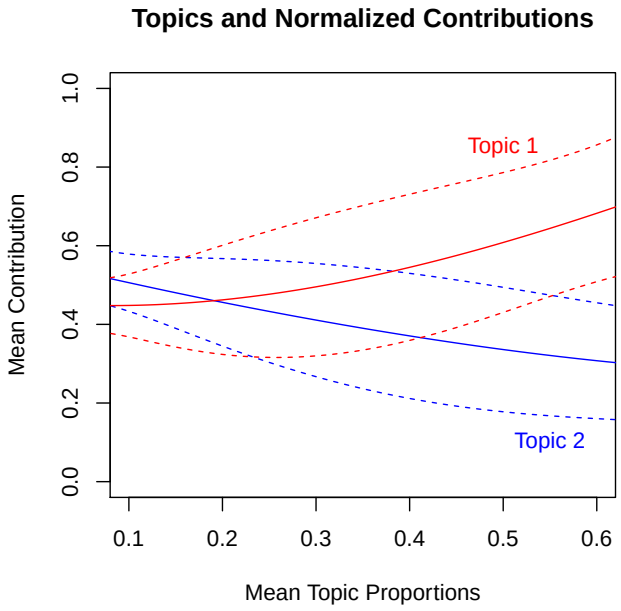



Figure 14: Time Pressure vs. Delay Topics and Contributions

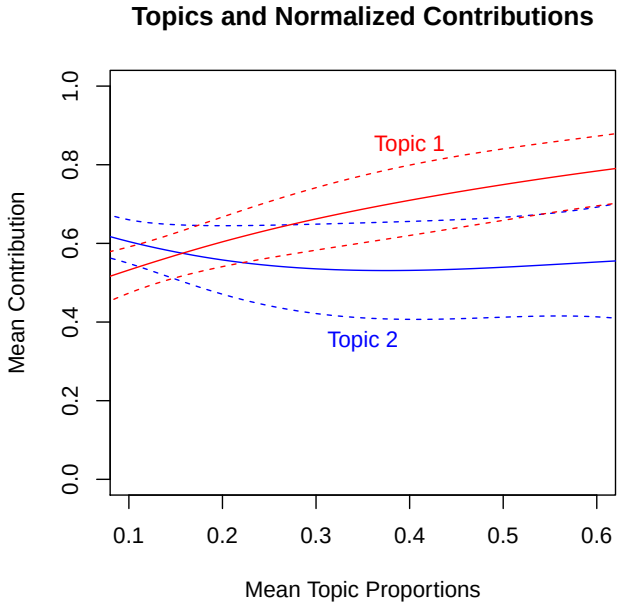



Figure 15: Intuition Topics and Contributions

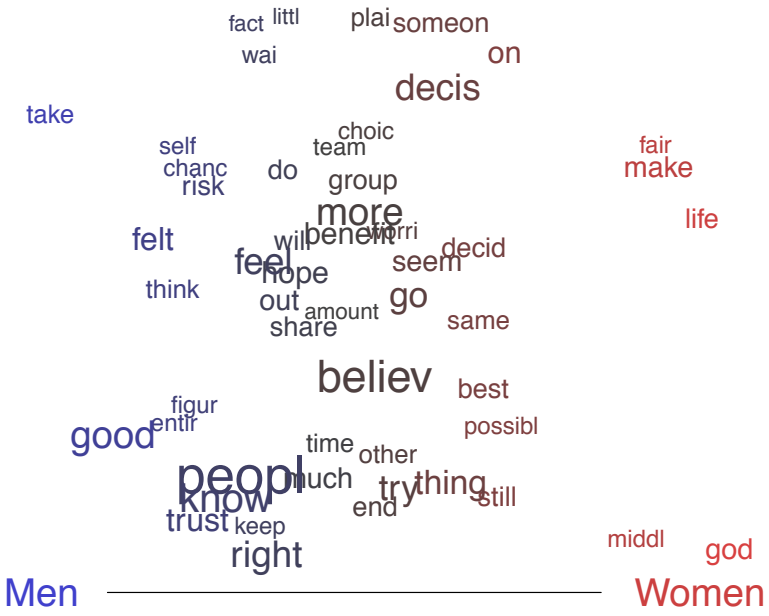



Figure 16: Intuitive Topic Allowing for Different Vocabularies Based on Gender

"risk", words associated with describing their own intuition. On the right side of the plot are words that women use more frequently within the intuition topic, including "god", "life", and "middle", words associated with morality or fairness (see also Figure 17).

By allowing topics to vary in vocabulary by gender, we expect that we will also better be able to measure treatment effects. If women and men simply explain their intuition differently, we shouldn't treat these two slightly different vocabularies as completely different topics. By allowing flexibility in the topics, we expect that we can more precisely measure the prevalence of topics within documents. Future research will explore the relationship between better estimated vocabularies and more precise treatment effects on topic proportions.

## 4.3 ANES

Our model has more general applications to open-ended survey questions even when no treatment condition is included in the study. In this section, we apply the Structural Topic Model to analyze open-ended responses from the American National Election Sur-

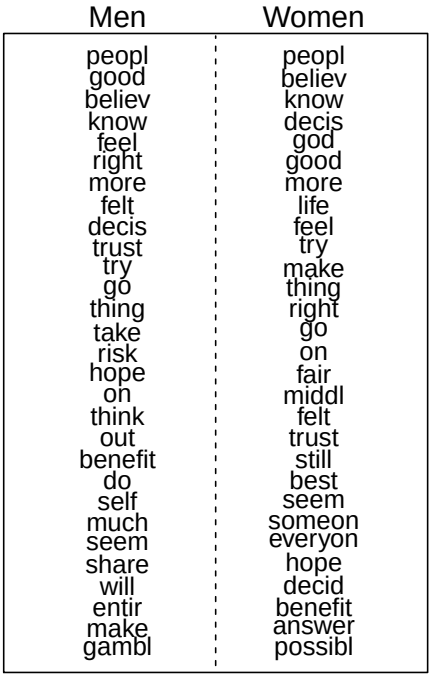



Figure 17: Comparison of Women and Men's Vocabulary After Intuition Treatment

vey (ANES). A sample of 2,323 respondents were interviewed after the 2008 presidential election. Each were asked to identify the most important and second most important political problem facing the United States, as well as the most important and second most important personal issue in the election. The original data was recently recoded by the ANES into a set of stable categories using human coding. We show that the STM is consistent with the human coding of the open-ended responses, while also uncovering new categories that are specific to the 2008 election.

We analyze open-ended responses that identify the most important political problem for each individual and use Party ID, education, age, and an interaction between Party ID and education as the covariates. Figure 18 displays the top topics from a 60 topic model and the frequency of these topics within our data.[12] The topics correspond closely to general topics we would expect: high frequency topics include "econ", "war", "don't know", "unemployment, recession", "job", and "terror".

The ANES hired human coders to code each of the open-ended responses into one of 69 categories.[13] Table 1 compares the aggregate categorization of the top categories between the STM and the hand-coding. The aggregate numbers of responses coded into each category are very similar across the STM and the ANES hand-coded data, even though the topic categories are not perfectly aligned between the STM and the pre-determined human categories. The major difference in the aggregate numbers is between the "Budget" and "Unemployment" categories. This difference is because the ANES had a catch-all "Economics" category, separate from "The Economy", "Budget", and "Unemployment", and this catch-all "Economics" category contains responses that the STM included in the "Deficit, debt", "Budget", or "Unemployment and job" categories.

Not only are the aggregate numbers are similar, many of the individual responses coded by the STM are similarly coded by the hand-coding scheme. For example, we

[12]Words used to label the topics are the most likely words within that topic. The number of words printed is determined by an algorithm that calculates the gap between the probability of the last word printed and the probability of the next most likely word.

[13]Some responses were placed into multiple categories if the coders determined that the response included multiple topics.

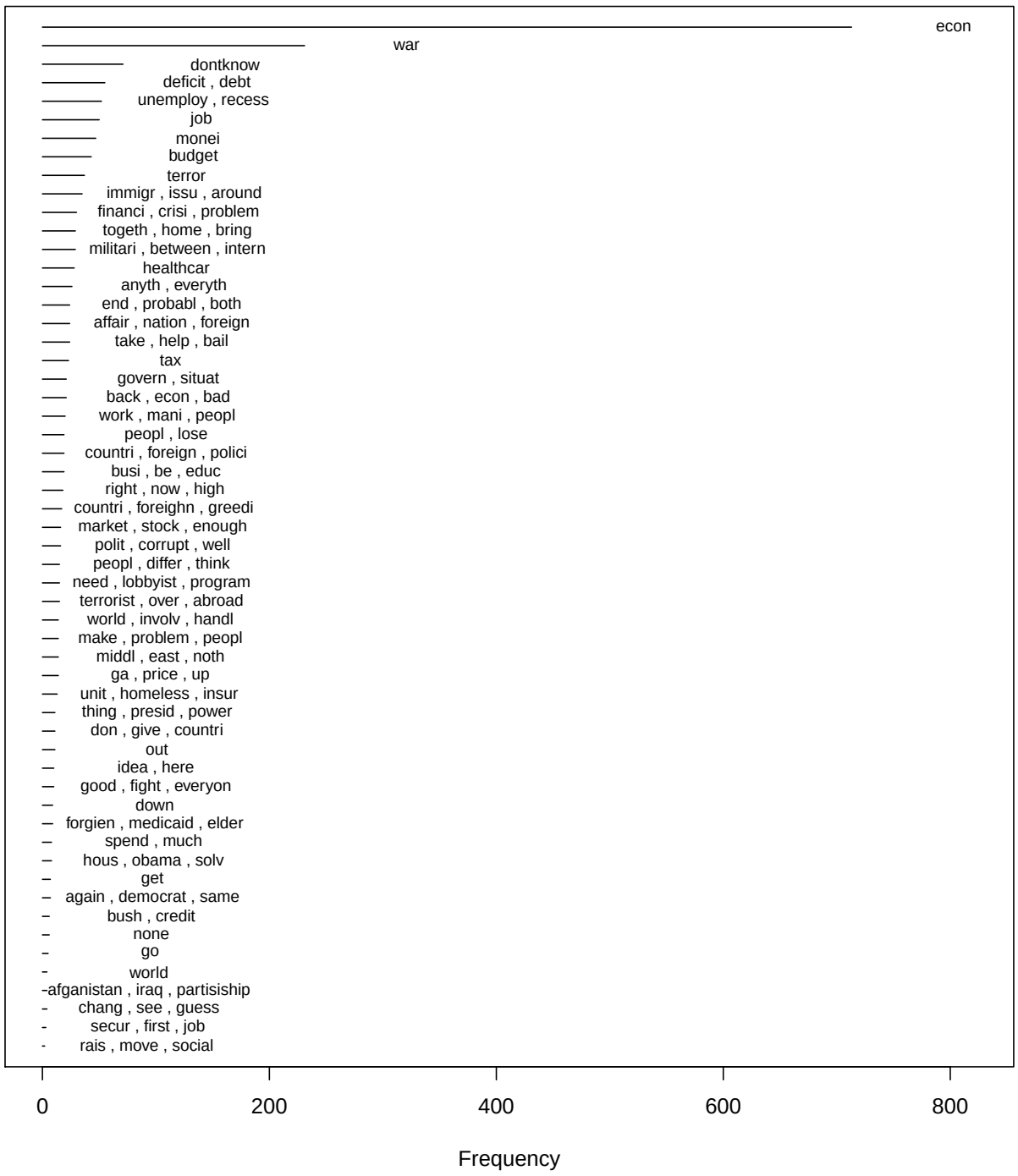



Figure 18: STM Topics from ANES Most Important Problem

| STM Category | STM Count | Hand-Coding Category | Hand-Coding Count |
|---|---|---|---|
| Economy | 777 | The Economy | 788 |
| War | 277 | War, or Iraq War | 281 |
| Don't Know | 82 | Don't Know | 90 |
| Deficit, Debt and Budget | 125 | Budget | 86 |
| Unemployment and Job | 151 | Unemployment | 119 |
| Terror | 56 | Terrorism | 69 |

Table 1: Comparison of to Hand Coding

compare responses that the STM estimated to have more than 5% of the topic "econ" with responses that were hand-coded to be at least partially in the topic "The Economy". Of the responses that were coded by the STM into "econ", only one of them, "the economic recession" was not also hand-coded into "The Economy". Of the responses that were hand-coded into "The Economy", 10% of them were not coded by the STM into "econ". These responses usually included the word "finance" or "financial crisis" in lieu of "economy" and therefore were classified by the STM into the "financial crisis" category.

While most of the responses in the ANES were placed into one category, the human coding did allow an individual response to fall into multiple categories. Overall, about 19% of responses were hand-coded into multiple categories. Whereas Table 1 shows that the STM is consistent with the hand-coding in putting responses into categories, a comparison of multiple categories provides yet another dimension along which we can compare hand-coding to the STM because the STM also allows responses to be a mixture over topics. We would expect, for example, that responses that the ANES coded into a single category would also be heavily centered over one topic in the results from the STM.

For each response we use the STM to calculate the number of topics that contribute to at least 5% of the individual's response. Using this method of comparison, we find high correlation between the number of topics the STM assigns to a response and the hand-coding assigned to a response. Of the responses the hand-coding coded into only one-category, 85% were also coded by the STM into only one category. Of the responses that the STM coded into one category, 90% were also hand-coded into one category. Overall, 92% of the responses were coded either in the same number of categories between the STM and hand-coding or only had one category difference. To consider a specific example, of

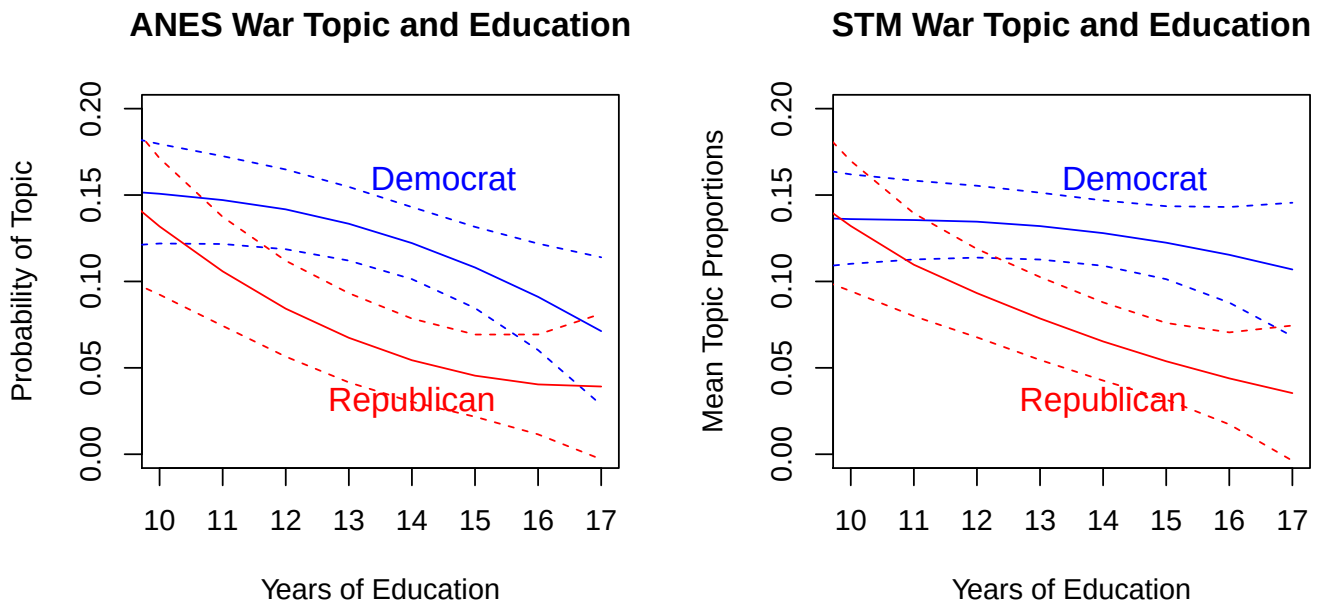



Figure 19: Comparison of Covariate Relationships

the responses that were hand-coded into the "Economy" and "Unemployment" categories, the two most prevalent topics from the STM were also "econ" and "unemployment".

The STM recovers very similar covariate relationships to those discovered by the ANES hand-coders. Figure 19 shows the relationship between Party ID, education, and the "war" topic in each case. The relationship between the covariates and the "war" topic look very similar between the two different models; respondents with higher education write less about war, and Democrats write more about war than Republicans. One benefit of the STM is that it produces a continuous measure of topics within each document, whereas the hand-coded data produces only a categorization. The continuous measure is much easier to work with when looking at covariate relationships, which is one advantage of using the topic model in place of, or in addition to, human-coding.

An additional advantage of using the unsupervised approach to categorize open-ended responses in the ANES data is that categories that are specific to the time period in which the survey was administered. Since this survey was administered directly after the 2008

election, many respondents indicated that their most important problem was "Obama", the "democrats", or "Bush", none of which are categories within the ANES hand-coding scheme.[14] Categorizations can change quickly with time period and researchers might be interested in these changes. Thus with respect to categorization, the STM allows for flexibility and adaptability difficult with human-coding.

Of course, there are disadvantages to using the unsupervised approach to the categorization of these open-ended responses. In particular, pre-determined categories that have a low-incidence rate within the open-ended responses are unlikely to show up within the STM. For example, human coders assigned one response to the category "China". Since very few responses mention China, the STM does not discover a topic related to China, and therefore this categorization would be lost using the unsupervised approach. Due to low-frequency responses like the one response to China, there are also some topics recovered by the STM that are not particularly meaningful, with most frequent words like "get" or "go", which represent a hodge-podge of low-frequency responses. These topics would become better defined with more data.

However, the costs of human-coding thousands of responses may balance out downsides associated with the unsupervised approach. Indeed, as the number of responses increases, the STM becomes more accurate and human coding becomes more unwieldy. At minimum, the ANES could use this unsupervised method in conjunction with human-coding. This approach would save significant time and money, would allow for the discovery of topics specific to the survey time period and would point human coders to ambiguous responses where human classification is more important.

# 5 Conclusion

Spurred by recent efforts in digitalization of text, researchers are rapidly reinventing the evidence base of the social sciences by introducing new measures of political and social behavior constructed form large text archives (Grimmer and Stewart, 2013; Hopkins and King, 2010; King *et al.*, 2009; Lazer *et al.*, 2009). At the same time political science has

[14]The ANES has a general category for "Politicians", but does not include categories for specific politicians.

shown increasing interest in survey experiments, merging the discipline's long-standing use of surveys with the inferential strengths of the experimental framework (Druckman *et al.*, 2006).[15] Yet interestingly these two trends have been remarkably distinct and analyses of textual data from surveys and experiments are rare. We show how the Structural Topic Model recently introduced by [Blinded For Review]provides a unified approach that enables scholars to take advantage of recent advances in automated text analysis. Crucially, it enables the analysis to investigate a range of interesting quantities of interest while incorporating information like a respondent's experimental condition or other covariates like partisan affiliation or gender. This in turn enables the estimation of standard quantities of interest with associated measures of uncertainty. All of these methods will be available in easy to use open-source software.

The proposed methodology is not without limitations. Like any unsupervised method there are model selection issues arising both from the multi-modality of the objective function as well as alternative specifications. We address these concerns in a principled and explicit manner introducing new model selection criteria to help guide analysts towards high quality solutions in an automated manner, as well as visualization tools that facilitate manual investigation of the models under consideration. In the supplemental appendix we engage with a range of potential criticisms such as concerns about false positives arising from the use of treatment assignment for both measurement and effect estimation. As with any research design, best practices will depend on the inferential goals and criteria within a research field. STM can be used for exploration of a corpus about which little is known *ex ante*, or along with rigorous pre-analysis plans that define clear prior predictions of expected topics. When the researcher has a body of text with metadata where the documents take on a mixture of documents, the STM can be useful for exploration, description and prediction. Opportunities for future research are immense, including new substantive applications, incorporation of respondent ranking in text (e.g. denote one problem the 'first' most important problem and another the 'second') and technical extensions like techniques for non-parametric selection of the number of topics (Paisley

[15] According to Hainmueller *et al.* (2012) the "top 3" general interest political science journals published 72 papers with survey experiments between 2006 and 2010.

*et al.*, 2011).

# 6 Appendix

This appendix outlines a number of important details about the STM. We have organized it into three sections (1) Model Estimation, (2) Model Validation, and (3) Getting Started. The Model Estimation section (6.1) presents the details of model estimation, which describes an Expectation-Maximization based approach to optimization of the model parameters.

With this formal statement in mind, the Model Validation section overviews theoretical and simulation evidence for the properties of the STM. Section 6.2.1 discusses the STM compared to alternative tools for text analysis, highlighting both similarities as well as important points of departure. Next in Section 6.2.5 we turn to a set of Monte Carlo based evaluations of the STM. We also include a discussion of inference in one of the paper's empirical examples using a two-step LDA based analysis that does not formally incorporate covariates in estimation.

The final section includes information directed at the applied user. Section 6.3.2 discusses the multiple testing problem and the use of the STM with pre-analysis plans. Section 6.3.4 gives an example of using the STM alongside mediation analysis. Section 6.3.1 demonstrates two-software tools we have developed to enable efficient work-flow for researchers using our methods.

## 6.1 Model Estimation

Before providing an overview of estimation we give the full data generating process:

1. Draw $\vec{\eta}_d | X_d\gamma, \Sigma \sim \mathcal{N}(\mu = X_d\gamma, \Sigma)$
2. For $n \in 1, \ldots, N_d$:
    - Draw topic assignment $Z_{d,n}|\vec{\eta}_d$ from Multi$(\theta = \frac{\eta_d}{\sum_\eta})$)
    - Draw word $W_{d,n}|z_{d,n}, \vec{\kappa}$ from Mult$(\beta \propto \exp(m + \kappa_k z_{d,n} + \kappa_c U_d + \kappa_I U_d z_{d.n})$

There are two substantive differences to the notation from the main paper. We leave the topic proportions in their unnormalized form and we expand out the notation for

prior distribution over words. Note also that $\gamma$ and $\kappa$ are maximized with a ridge regression penalty which is equivalent to a zero-mean Normal prior distribution with variance proportional to the observation variance.

Model estimation proceeds via mean-field variational EM where the unnormalized document proportions $\eta$ and the individual word assignments $z$ are inferred during the variational E-step, and the global parameters are inferred in the M-step. Estimation is complicated by the non-conjugacy of the logistic normal and the multinomial. To deal with non-conjugacy, in the variational E-step we use a second-order Taylor series approximation to the optimal variational update as in Wang and Blei (2013). Note that this means inference in our model does not enjoy the theoretical guarantees of mean-field variational inference in the conjugate exponential family (see for example Grimmer (2011) for an introduction to conjugate variational approximations for political scientists). Nonetheless, the approximation described here has been shown to work well in practice in numerous applications as well as in evaluations by Wang and Blei (2013).

Full details are available in a companion technical manuscript and code will be available on Github upon publication, if not before. However, for completeness we overview the estimation below, omitting derivations.

### 6.1.1 General Inference Strategy

In mean-field variational bayes, we form a deterministic factorized approximation to the posterior which minimizes the KL-divergence between the true posterior and the approximation (Bishop, 2007; Grimmer, 2011). This turns the Bayesian inference problem into an optimization problem. Thus we avoid the complications of convergence monitoring in MCMC and also typically enjoy substantially faster run times.

In our case, we approximate the posterior distribution over the document-level latent variables $p(\eta, \vec{z})$ with the factorized posterior $q(\eta)q(\vec{z})$ which given the conditional independence assumptions of the model further factorizes as $q(\eta)\prod_n q(z_n)$. We then seek to minimize the KL-divergence to the true posterior, $\mathrm{KL}(q||p)$. In conjugate exponential family cases, standard derivations (Bishop, 2007) show that the optimal updates for the

variational posterior are:

$$q^*(\eta) \propto \exp(E_{q(z)}[\log\, p(z, \eta)]) \tag{1}$$

$$q^*(z) \propto \exp(E_{q(\eta)}[\log\, p(z, \eta)]) \tag{2}$$

Since the logistic-normal is not conjugate to the multinomial these distributions will not be an easily normalized form. Instead we use an approximation (described in more detail below) such that $q(\eta)$ will be multivariate normal distribution with mean $\lambda$ and variance matrix $\nu$, and $q(z)$ will be discrete with parameter $\phi$.

The global parameters $(\gamma, \Sigma, \kappa)$ are set to their MAP estimates during the M-step. Thus the inference algorithm is analogous to a standard EM algorithm, except that we take the expectations with respect to the variational distributions $q(\eta)$ and $q(z)$.

Thus in each stage of the EM algorithm we go through the following steps (note we have placed substantive interpretations of the parameters in parentheses):

- For each document
  - Update $q(\eta)$ (document-topic proportions)
  - Update $q(z)$ (assignments to topic for each word)
  - Repeat until the $L_2$ norm of the change in $q(\eta)$ is below the threshold
- Update $\gamma$ (coefficients for topic prevalence)
- Update $\Sigma$ (global covariance matrix controlling correlation between topics)
- Update each $\kappa$. (topical-content parameters describing deviations from baseline word rate)
- Repeat until convergence.

Convergence can be assessed by monitoring either an approximation to the bound on the marginal likelihood, or by change in the variational distribution $q(\eta)$. In practice we monitor change in the variational distribution at the document level and monitor convergence on the marginal likelihood at the global level as in Wang and Blei (2013).

### 6.1.2 E-Step

In our case, the necessary expectations are not tractable due to the non-conjugacy, and so we form a Gaussian approximation to the update step using a second-order Taylor expansion. The next section closely follows the derivations in Wang and Blei (2013). We start by letting $f(\eta)$ represent the expectation over the log of the joint distribution. Taking the Taylor expansion around the maximum of the function, denoted $f(\hat{\eta})$, yields:

$$q(\eta) \propto \exp\left(f(\eta)\right) \tag{3}$$

$$\approx \exp\left(f(\hat{\eta}) + \frac{1}{2}(\eta - \hat{\eta})^T \nabla^2 f(\hat{\eta})(\eta - \hat{\eta})\right) \tag{4}$$

$$= \mathcal{N}(\lambda = \hat{\eta}, \nu = -\nabla^2 f(\hat{\eta})^{-1}) \tag{5}$$

where $\nabla^2 f(\hat{\eta})$ is the hessian of $f(\eta)$ evaluated at the mode. Note that the gradient does not appear in the approximation since we know that the gradient is zero (by virtue of being centered at the maximum). This is the Laplace approximation to the optimal variational update.

We find the mode of $f(\eta)$ using the quasi-Newton method BFGS, with objective and gradient,

$$f(\eta) = \left(\eta - \log\left(\sum_k \exp(\eta_k)\right)\right)^T \bar{t_z} - \frac{1}{2}(\eta - \mu_d)^T \Sigma^{-1}(\eta - \mu_d) \tag{6}$$

$$\nabla f(\eta) = \bar{t_z} - \theta \sum_k [\bar{t_z}]_k - \Sigma^{-1}(\eta - \mu_d) \tag{7}$$

where we use $\theta$ to indicate the simplex mapped $\eta$, and $\bar{t_z}$ to indicate the sufficient statistics for $q(z)$.

Once $q(\eta)$ is updated, $q(z)$ is easily updated as a function of the mean of $q(\eta)$, $\lambda$, by

$$\phi_{n,k} \propto exp(\lambda_k)\beta_{k,w_n} \tag{8}$$

### 6.1.3 M-Step

In the M-step we update the global parameters which control topic prevalence and topical content. The parameters for topical prevalence, are simply penalized maximum likelihood

estimation for a multivariate normal. Thus the coefficients are updated as:

$$\hat{\gamma} = (p_\gamma I + X'X)^{-1}X'\lambda$$

where $p_\gamma$ is our user specified penalty parameter and $I$ is the identity matrix. Note that this is equivalent to solving the ridge regression problem. In different contexts alternative methods for solving this minimization problem may be preferable using the Cholesky or QR decomposition methods.

$\Sigma$ is estimated as in Blei and Lafferty (2007) with the exception that each document now has its own mean via $X_d\hat{\gamma}$ and each document has a full covariance matrix $\nu_d$. This yields the update,

$$\hat{\Sigma} = \frac{1}{D}\sum_d \nu_d + (\lambda_d - X_d\hat{\gamma})(\lambda_d - X_d\hat{\gamma})^T$$

In the estimation for topical content, we base our estimation strategy off of the work of Eisenstein *et al.* (2011), again doing MAP estimation, in this case using a Normal prior. The idea is that each topic is represented as a deviation from the baseline word frequency $m$ in log-space. Thus for topic $k$, the rate is $\exp(m + \kappa_k)$.

For estimation we have $\kappa$'s and the constant $m$. We set $m$ to the empirical log-probability of the words in the dataset. This leaves MAP estimation of $\kappa$ as the main inference problem. Addressing $\kappa_k$ we get the following objective,

$$\mathcal{L}_\kappa = \sum_{k=1}^{K}\sum_{j=1}^{J} E_q[\log\, p(\kappa_{k,j}|v_2)] + \sum_{d=1}^{D}\sum_{n=1}^{N} E_q[\log\, p(w_{n,d}|z_{n,d},\beta)]$$

This is equivalent to the log posterior of a multinomial regression on words under a ridge regression penalty. We optimize using BFGS with likelihood and gradient,

$$\mathcal{L}_{\kappa_k} = \langle c_k\rangle\kappa_k - \langle C_k\rangle\log\sum\, \exp(\vec{\kappa} + m_i) - \frac{1}{2}\kappa^2/v_2$$

$$\nabla\mathcal{L}_{\kappa_k} = \langle c_k\rangle - \sum_j \langle C_{jk}\rangle\beta_{jk} - \kappa/v_2$$

where $c_k$ is $V$-length vector of expected counts for each term in the vocabulary for the given topic. $C_k$ is the summation over that vector producing a scalar equal to the expected number of tokens assigned to topic $k$.

When estimating with covariates, for each covariate $j$ we consider the word rate deviation for any document of that covariate $\kappa_j$ as well as a topic-covariate interaction $\kappa_{j,k}$. Thus the form for the word probabilities is,

$$\beta_{k,j} \propto \exp(m + \kappa_k + \kappa_j + \kappa_{j,k}) \tag{9}$$

We then solve each block of the $\kappa$'s separately with objective functions analogous to the one above.

### 6.1.4 FREX and Semantic Coherence Scoring

In this section we provide technical details for a few of our model diagnostics. FREX words are a way of choosing representative words for analysis. It builds off the work of Bischof and Airoldi (2012) but with a substantially simpler model. Specifically FREX is the geometric average of frequency and exclusivity. We define the exclusivity term as:

$$\text{Exclusivity}(k, v) = \frac{\beta_{i=k,j=v}}{\sum_i \beta_{j=v}}$$

where $i$ indexes the topics and $j$ indexes the vocabulary words. Then our FREX score is:

$$\text{FREX}(k, v) = \left( \frac{\omega}{\text{ECDF}(\text{Exclusivity}(k, v))} + \frac{1 - \omega}{\text{ECDF}(\beta_{k,v})} \right)^{-1}$$

where ECDF is the empirical CDF function, and $\omega$ is the weight given to exclusivity, here .5 by default.

Semantic coherence comes from Mimno *et al.* (2011) and is based on co-occurrence statistics for the top $n$ words in a topic. Thus where $D()$ is a function of a word index $v_i$ which outputs the number of documents containing its arguments, we get

$$\sum_{n=2}^{N} \sum_{m=1}^{n-1} \log \left( \frac{D(v_n, v_m) + 1}{D(v_m)} \right)$$

Intuitively this is a sum over all word pairs in the top topic words, returning the log of the co-occurrence frequency divided by the baseline frequency. The one is included to avoid taking the log of zero in the event that a pair of words never co-occurs (which is possible with very short documents). This measure is closely related to pointwise mutual information.

## 6.2 Model Validation

This section of the appendix overviews a series of validation exercises including comparisons to existing unsupervised models. Understanding that no single test can validate the model, we build confidence in our approach by marshalling evidence from tests using simulated data, permutation tests and additional tests using the real documents in our presented experiments. When using real documents, we use the immigration dataset from Gadarian and Albertson (2013) throughout for continuity.

In Section 6.2.1 we compare the STM to prominent unsupervised models within the literature, highlighting the contrasts to existing approaches. We discuss the merits of including covariates and jointly estimating their effects on the topics rather than using a two-step approach. In Section 6.2.5 we use simulations to demonstrate that the STM is able to recover parameters of interest. In Section 6.2.6 we move from simulated data to a permutation test on actual documents in which we randomly permute the treatment indicator to demonstrate that the STM does not find spurious treatment effects. In Section 6.2.7 we provide a comparison with a two-stage approach using standard LDA, highlighting the differences on the immigration data. Finally, in Section 6.2.8 we compare the STM results on the immigration data to the analysis of the same data with human coders.

### 6.2.1 Comparison to Alternative Models

Computer-assisted methods for the measurement of political quantities of text have already seen widespread use in other areas of political science (Laver *et al.*, 2003; Slapin and Proksch, 2008; Grimmer, 2010; Quinn *et al.*, 2010).[16] In this section we contrast our approach with three alternative unsupervised text analysis models in the literature, focusing on the advantages of including covariates.

[16]In restricting our focus to the political science literature we naturally omit the large literature in computer science, statistics and other social sciences. See Blei (2012) for a review of Latent Dirichlet Allocation and related models.

### 6.2.2 LDA and Factor Analysis

The standard Latent Dirichlet Allocation model can provide rich, interpretive semantic depictions of text. However, central to LDA is the assumption of *exchangeability*, the idea that the ordering of the documents in the corpus is irrelevant. Not only do we believe that survey respondents naturally have systematic, and easily-measurable, differences (e.g. gender, income etc.) but in experimental settings the design of the study suggests non-exchangeability between treatment and a control condition. We could always simply ignore this and run the standard LDA model, but even in a best case scenario where comparable topics are estimated, the resulting estimate of the covariate effects will be inefficient. We describe the process of running LDA (without covariate information) followed by analysis of covariate relationships with the results resulting topics as a "two-step" approach.

Indeed, Hopkins (2012) provides a recent example of the best use of the two-step approach. He uses the correlated topic model, a close relative of LDA, to study open-ended survey responses from 30,000 respondents on healthcare. One of his primary objectives is to show trends over time and reaction to elite framing in press releases. Models such as the STM would allow for an analysis that directly incorporates the effects of time on topic proportions and also makes use of the rich demographic meta-data available on the individual responses.

In Simon and Xenos (2004), the authors propose the analysis of open-ended survey response using latent semantic analysis in a three stage process: data preparation, exploratory factor analyses and hypothesis testing. Specifically they advocate the use of latent semantic analysis (LSA) because it is "rigorous, systematic, and, most of all, efficient." We agree with the goals of their work. Indeed our model exists in a lineage that extends from latent semantic analysis; LDA was developed as a probabilistic formulation of latent semantic analysis which is appropriate to discrete data such as word counts, the STM is a related alternative to LDA appropriate to the inclusion of rich covariate information.[17] While LSA was amongst the best available techniques for the dimension

[17]Buntine and Jakulin (2005) shows that multinomial principal components analysis and LDA are

reduction of text at the time of the article, the STM represents advances in statistical rigor and efficiency of the intervening decade. Furthermore, because The STM produces an approximate posterior distribution over topic classifications, it is more amendable to the kind of hypothesis testing advocated by Simon and Xenos (2004). By propagating our uncertainty about a document's coding, we can lower the risks of measurement error.[18]

We also note that the clear data generating process of the STM, allows us to avoid the numerous practical issues that are discussed in length by Simon and Xenos (2004) such as, weighting schemes of the word-frequency matrix and selection of rotation. The topics themselves are also more interpretable than factor loadings, specifically in the STM or LDA, the topic loading for a document is the percentage of words attributable to a latent topic; by contrast, eigenvalues and document scores have little *prima facie* meaning to the user. Thus, we argue that the STM offers all the advantages of the framework laid out by Simon and Xenos (2004) with the addition of numerous quantities of interest.

### 6.2.3 Single-Membership Models with Covariates

The literature on single-membership models in political science offers another appealing approach. Work such as the Dynamic Topic Model (Quinn *et al.*, 2010) and the Expressed Agenda Model (Grimmer, 2010), take into account systematic variation amongst documents based on *specific* characteristics such as time and authorship. These models allow the analyst to specify political knowledge about the documents at the outset. However, these existing single-membership models have focused on modelling particular types of structure. The STM generalizes these models by allowing any structure which can be captured in a general class of linear models. This allows the STM to model time, authorship, or time and authorship interactions all within the same framework. The STM also provides variational posteriors over the parameters which provide estimates of uncertainty

closely related models. In addition to the numerous theoretical benefits, the use of the LDA framework side steps the numerous practical difficulties of the standard principal components analysis advocated by Simon and Xenos (2004). In particular the user is not faced with un-interpretable quantities such as negative factor loadings or concerns about rotations.

[18]See Grimmer (2010) for arguments on this in the context of text data, and Blackwell *et al.* (2011); Treier and Jackman (2008) on measurement error in other contexts.

(similar to the Expressed Agenda Model but not the Dynamic Topic Model).

These models also assume that responses contain words originating from only a single topic. This can be advantageous as it often makes the corresponding optimization problem easier, with fewer local modes. However, we argue that in most open-ended responses, survey respondents are likely to give multi-faceted answers which could touch on a variety of semantic themes. At a minimum it is desirable to relax this restrictive assumption and see if the results change substantively. Furthermore, each of these models assume that topics are talked about in the same way, whereas as the STM allows (for example) men and women to talk about the same topics differently, as we describe in our empirical examples and simulations.

Single-membership models in political science do establish and excellent tradition of validation of the topics based on external political phenomena and careful reading (Grimmer and Stewart, 2013). These validations can be performed in our setting by using model output to examine documents which contain a large percentage of words devoted to a particular topic. In the STM model, the analyst can also use the covariate relationships in both topic prevalence and topical content to verify that the results are sensible and in line with theoretical expectations. The best validations will necessarily be theory-driven and application specific, but we direct interested readers to Krippendorff (2004) for some excellent guidelines and Quinn *et al.* (2010) for some excellent applied examples.

Thus, previous work highlights a useful insight; including information about the structure of documents can improve the learning of topics. These changes can be motivated in the language of computer science via the no-free lunch theorem (Wolpert and Macready, 1997), or statistically via literature on partial pooling (Gelman and Hill, 2007). The source of the improvement is the identification of units which are "similar" and allowing the model to borrow strength between those units, without making the stronger exchangeability assumptions.

We can summarize the relative merits of the STM framework with Table 2 inspired by the table of the common assumptions and relative costs of text categorization in Quinn

*et al.* (2010).

| | Hand Coding | Factor Analysis | LDA | Single Member | STM |
|---|---|---|---|---|---|
| Categories Known | Yes | No | No | No | No |
| Mixed-Membership | No | Yes | Yes | No | Yes |
| Covariates on Prevalence | No | No | No | Limited | Yes |
| Covariates on Content | No | No | No | No | Yes |
| Interpretable QOIs | Yes | No | Yes | Yes | Yes |
| Uncertainty Estimates | No | Limited | Yes | Yes | Yes |

Table 2: Relative Merits of STM and competing approaches for analysis of open-ended survey response.

#### 6.2.4 Estimation vs. Two-Stage Process

We can contrast the STM's joint estimation of covariates and topics with a two-stage process in which some method (factor analysis, LDA, human classification) is first run and then the estimated topic proportions are entered into a regression of interest. The intuitive appeal of a two-stage estimation with LDA is the ease of use and the separation of hypothesis testing from measurement. Indeed the analyst might reasonably be concerned that joint estimation of the topics and covariates will induce spurious correlations between the topics and the variables of interest (although in the next section we demonstrate using Monte Carlo simulations that this is not the case).

The hidden cost of two-stage estimation is the inability to adequately account for our uncertainty. Specifically, we are unable to propagate the uncertainty in our classifications to our regressions with covariates, since the naive approach will condition on the observed values of the topics developed in the first stage. Furthermore, the model implicitly assumes that, in expectation, the topics are discussed in the same way and with the same prevalence across all respondents. In Section 6.2.7 we explore this approach and highlight problems with it.

The concern about spurious correlations is reasonable, and any time we use unsupervised methods for measurement we must be careful to validate our findings. Spurious correlations can arise even in a two-stage process if, for example, the analyst iterates betweens specifications and observing covariate relationships. In the STM model, we address these concerns with regularizing priors which draw the influence of covariates to zero

unless the data strongly suggests otherwise. In practice, this means that the analyst can specify the degree to which she wants the covariates to enter into the model. Standard diagnostics such as posterior predictive checks or cross-validation can also be used. Posterior predictive checks in particular have been shown to be able to diagnose violations of assumptions in LDA, providing one method of checking whether additional structure needs to be incorporated in the model (Mimno and Blei, 2011).

### 6.2.5 Monte Carlo Analysis

In this section we use simulated data to test the model. Unfortunately the process of writing natural language text bears little resemblance to the bag-of-words data generating process assumed by almost all topic models. In some cases this is to the model's disadvantage, because the interdependencies amongst words are more deeply complex than the model's assumptions. Yet, in many cases it is an advantage; there are features to natural language text, that make it amenable to analysis of this sort: people write with finite vocabularies, use words consistently and tend to write about relatively focused topics. Simulation parameters must be carefully chosen to reflect patterns we can reasonably expect to find in text. Accordingly we provide a series of tests in this and the following sections which move from purely simulated documents to full analyses on real texts.

The advantage of purely simulated data is that we can compare recovery to the known truth. The evaluation of parameter recovery is complicated by issues of identification in the model. At the most basic level topics are invariant to label switching (Topic 1 and Topic 2 could be switched since the number is irrelevant), and there can be myriad causes of local modes in the posterior. Indeed under some very general conditions the LDA inference problem is probably NP-hard (Sontag and Roy, 2009). In order to compare comparable units between model fits, we use the Hungarian algorithm to approximate the best possible match (Papadimitriou and Steiglitz, 1998; Hornik, 2005). The Hungarian algorithm solves the optimal assignment problem (which is typically NP-hard) in polynomial time, a speed advantage which becomes important when the number of topics is above 4.

**Basic Simulations** For our most basic simulations, documents are created according to the generative model for LDA with different means for each topic depending on the treatment assignment. Thus the generative process for the corpus parameters is:

$$\begin{aligned}\beta_k &\sim \text{Dirichlet}(.05)\\ \alpha_{t=0} &= (0.3, 0.4, 0.3)\\ \alpha_{t=1} &= (0.3 - ATE, 0.4, 0.3 + ATE)\end{aligned}$$

where each of the $k$ topics for $k \in 1, 2, 3$, is drawn from a symmetric 500-dimensional Dirichlet with concentration parameter of .05. The mean of the document topic proportions is based upon the treatment assignment, where the first half of the documents are assigned control, and the second half are assigned treatment. Then for each document

$$\begin{aligned}N_d &\sim \text{Poisson}(\zeta)\\ \theta_d &\sim \text{Dirichlet}(\alpha_d * G_0)\\ \vec{w}_d &\sim \text{Multinomial}(N_d, \theta_d\beta)\end{aligned}$$

We consider a host of different parameters but for space present here only the results across the following dimensions: Number of Documents (100 to 1000 by 100 increments), $\zeta$ the expected number of words (40), the size of the ATE (0, .2), and $G_0$ the concentration parameter for the Dirichlet (1/3).

We estimate the STM model on each dataset using default parameters for the prior. We then calculate the modal estimates for the topic proportions ($\theta$) and perform a standard OLS regression to get estimates of the treatment effect and confidence intervals.

The results are shown in Figure 20. In the case of no treatment effect as well as a sizable treatment effect, the STM is able to recover the effects of interest. As expected, uncertainty shrinks with sample size.

**Interaction Simulations** Here we consider a slightly more complex simulation. Specifically we consider a continuous variable which under control has a negative effect on a topic, and under treatment has a positive effect on a topic. In each case we simulate with 100 documents, a vocabulary size of 500 and 50 words per document in expectation. The

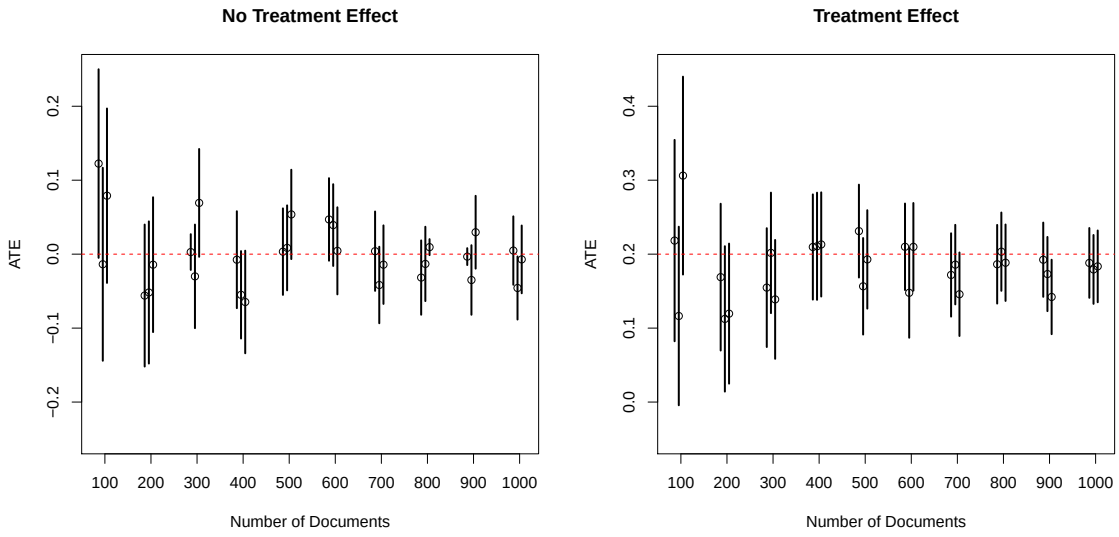



Figure 20: Estimated average treatment effect with 95% confidence intervals holding expected number of words per document fixed at 40, and the concentration parameter fixed at 1/3. The STM is able to recover the true ATE both in cases where there is no treatment effect (left) and cases with a sizable treatment effect (right). As expected, inferences improve as sample sizes increase.

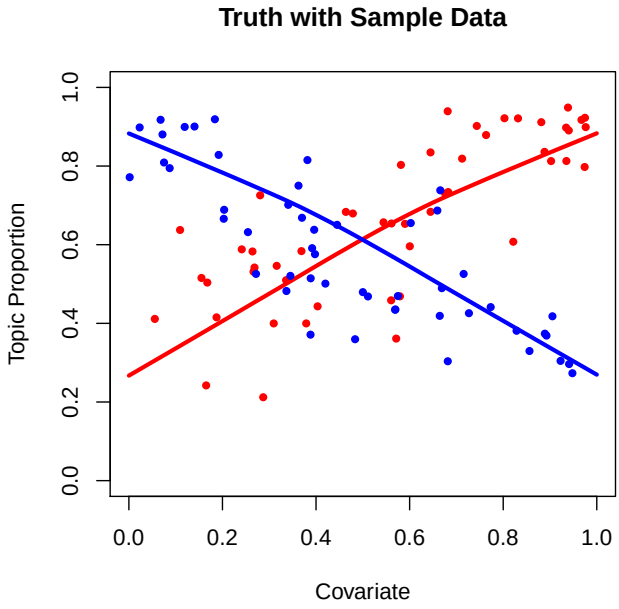



Figure 21: Simulated covariate effect, where treatment assignment causes a continuous variable along the $X$ axis to have either a positive or negative effect on a topic. Lines shown are the true relationships for each simulation while the dots indicate the true latent parameters for a single dataset.

true relationships are plotted in Figure 21 along with a sample of what the true latent data might look like for a particular simulation.

Using the STM and LDA we fit a model to each of 250 simulated datasets. For each dataset we plot a single line to indicate the treatment effect and the control effect (separated here for clarity). Lines are plotted semi-transparently for visibility with the true relationship super-imposed as a dotted black line. The fitted lines are shown in Figure 22.

Since it can be difficult to see the distribution, we show a histogram of the recovered slopes with a thicker black line demarcating the true value. These are plotted in Figure 23.

These simulations indicate that while LDA does sometimes correctly find the relationship it often is unable to capture the dynamic appropriately, in some cases reversing

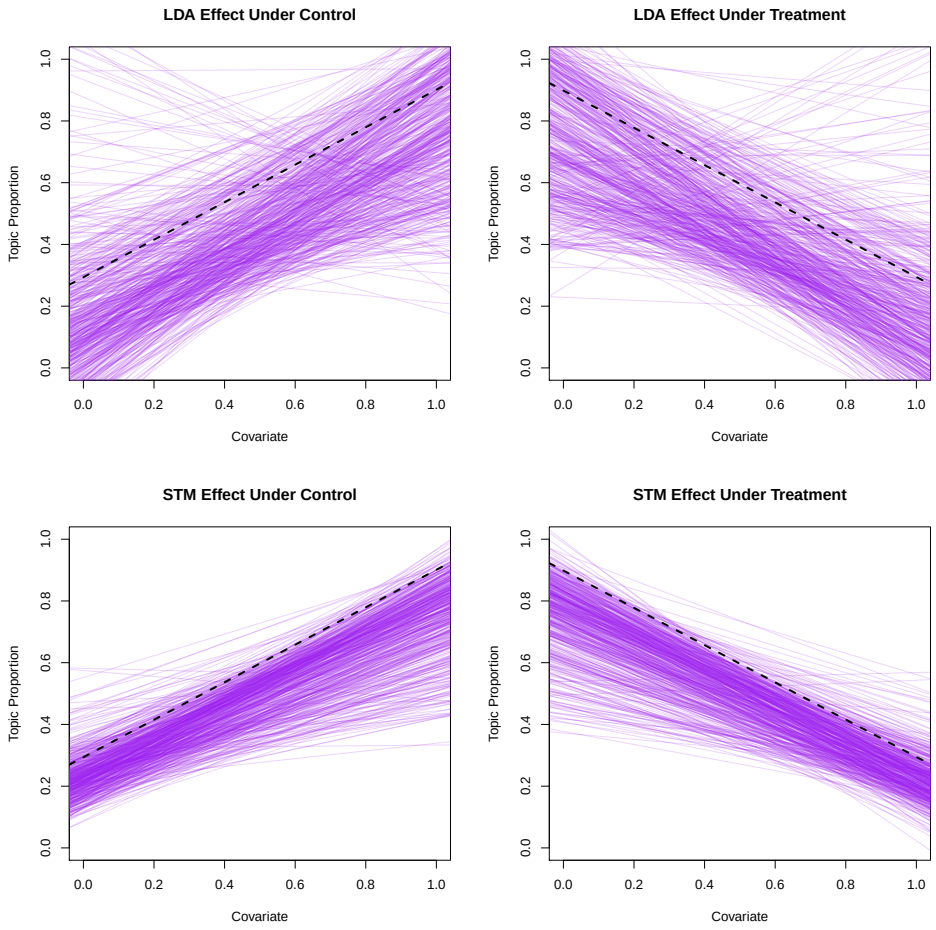



Figure 22: Covariate relationships recovered for treatment and control cases by LDA and STM.

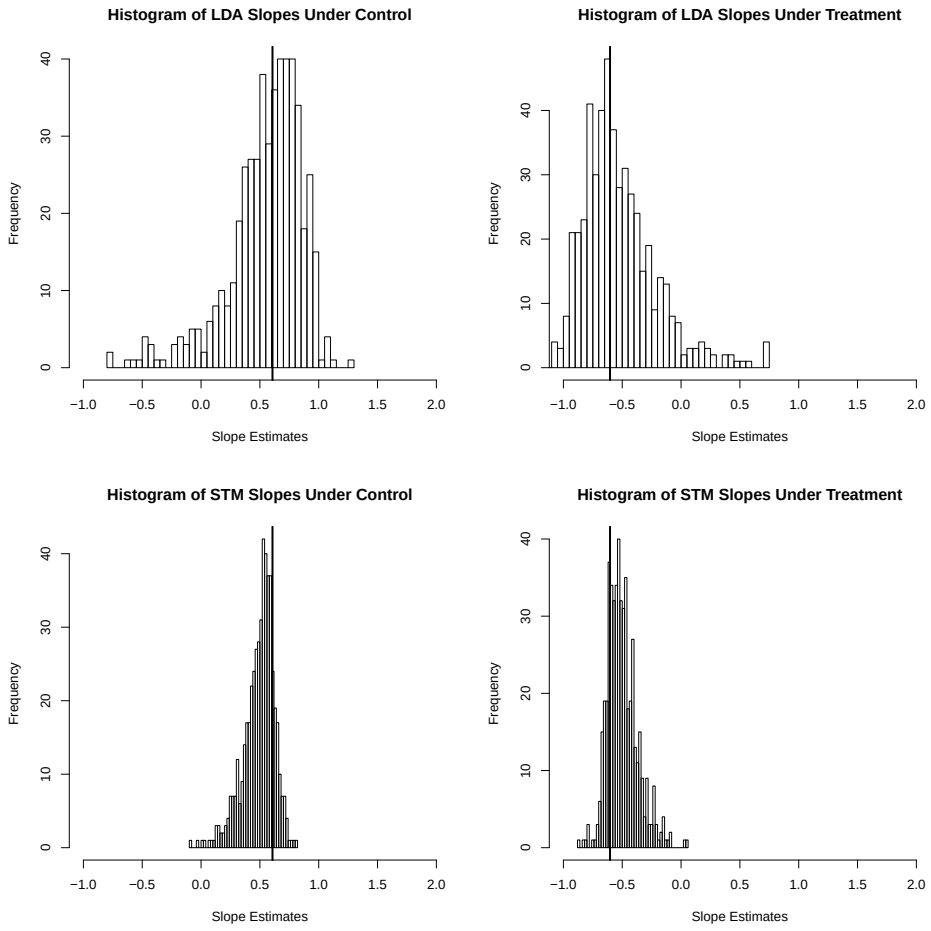



Figure 23: Slopes of continuous covariate effects recovered by STM and LDA

the sign of the effect. The STM by contrast tightly adheres to the true effect.

The reason the majority of the simulations appear below the line in both cases is an artifact of the data generating process. Document proportions are simulated in a continuous space but in practice the expected topic proportion is effectively discretized by the number of words. As a result, the true proportions are going to be slightly underestimated at both ends. The element of interest, both for the simulation as well as in practical examples is the slope of the line which constitutes the effect of the covariate under the two regimes.

### 6.2.6 Permutation Analysis

In this section we use one of our examples of real text data to show that when we randomly permute treatment assignment between text documents, we do not recover a treatment effect. This is further evidence that our model does not induce an effect when one does not in fact exist. In addition, under the true treatment assignment we see the most extreme effect, verifying that our results are not an artifact of the model, but instead reflect a consistent pattern within the data.

To do this, we use the Gadarian and Albertson (2013) data to estimate the effect of treatment on anxiety toward immigrants.[19] We estimate our model 100 times on this data, each time with a random permutation of treatment and control assignment to the documents. We then calculate the largest effect of treatment on any topic. If the results relating treatment to topics were an artifact of the model, then we would find a significant treatment effect regardless of how we assigned treatment to documents. If the results were a reflection of a true relationship between treatment and topics within the data, we would only find a treatment effect in the case where the assignment of treatment and control align with the true data.

Figure 24 shows the results of the permutation test. Most of the models have effect sizes clustered around zero, but the estimation that included the true assignment of treatment and control, indicated by the dotted line, is far to the right of zero. This indicates that the estimation itself is not producing the effect, but rather that the relationship

[19]To simplify the permutation test, we do not include Party ID as a covariate.

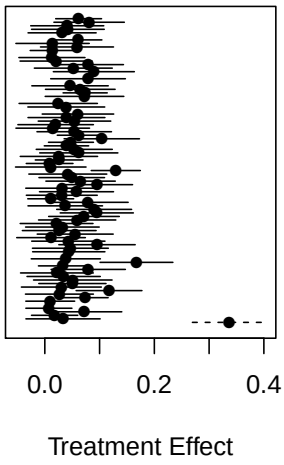



Figure 24: Maximum Treatment Effect Across Permutations

between treatment and topics arises within the data.

Note that more than 5% of the intervals don't cover zero, this arises, among other factors, from searching for the largest effect size. Such a search would be inappropriate in the standard use of permutation tests but provides a more rigorous test of our claim. It also is suggestive evidence against a concern with multiple testing in this framework. We return to this issues later, but essentially were multiple testing over topics putting us at risk of finding consistently large spurious effects, we would also see large spurious effects in our permutation test. As we don't, we take that as evidence for a cautious optimism that multiple testing concerns are not a huge problem here.

### 6.2.7 Comparison with LDA

An important question we asked ourselves in developing our approach was 'is there any need to go beyond a two-step process using Latent Dirichlet Allocation (LDA) first and then estimating covariate relationships after the fact?' We addressed this with simulated data above, but it is useful to add to those findings an assessment on real data. Specifically we are interested in characterizing whether LDA is able to recover a simple binary

treatment effect in one of our real datasets.

In order to demonstrate the differences in solutions attained by the two methods, we re-analyze the immigration data from Gadarian and Albertson (2013). As with STM, we first remove all stopwords and stem the corpus before conducting our analysis. We take an additional step of removing any word which only appears once in the corpus and any documents which subsequently contain no words. The unique words cause instability in the LDA analyses whereas STM is relatively unaffected due to the use of prior information.[20] We estimate LDA using collapsed Gibbs sampling (CGS), with fixed hyperparameters both at .05, running for 1000 iterations discarding 950 for burnin.[21] We estimate STM using default priors and using *only* the binary treatment indicator. As in the permutation test this simplifies the interpretation and crucially it biases *against* STM because it means we are using considerably less information than is available.

We start by running each algorithm 250 times with different starting values. We evaluate each run on our two selection criteria: *semantic coherence*, which measures the frequency with which high probability topic words tend to co-occur in documents, and *exclusivity* which measures the share of top topic words which are distinct to a given topic. In each case we use a comparison set size of 20 which was chosen in advance of viewing the results. We emphasize that our model does not directly optimize either criterion and

[20]The stabilizing influence is most likely a result of the structured priors in STM. Inference for topic assignment of words which only appear once comes from other words in the same document as the single-appearance words. Thus imagine the word "unlikely" appears only once in the corpus. In LDA its topic assignment will be based on the proportion of words assigned to each topic within the same document. In STM the assignment will be based both on the other words within the same document, but also other words in documents with similar covariate profiles. This tends to lead to inference which is more stable across initializations.

[21]Collapsed Gibbs sampling (CGS) for LDA is far in excess of what is needed for convergence. Many analysts simply use the final pass rather than averaging over iterations, but we use the last 50 in order to decrease sampling variability. To remove sampling variability as an issue we had originally used a variational approximation of LDA, but the algorithm produced pathological results for the K=3 case we consider here when estimating the Dirichlet shape parameter on topic prevalence. We do however emphasize that CGS is generally considered to be more accurate than standard variational approximation for LDA.

thus it provides a fair basis for comparison between the two models. Figure 25 shows the space of all LDA and STM solutions with larger numbers being better along each dimension.

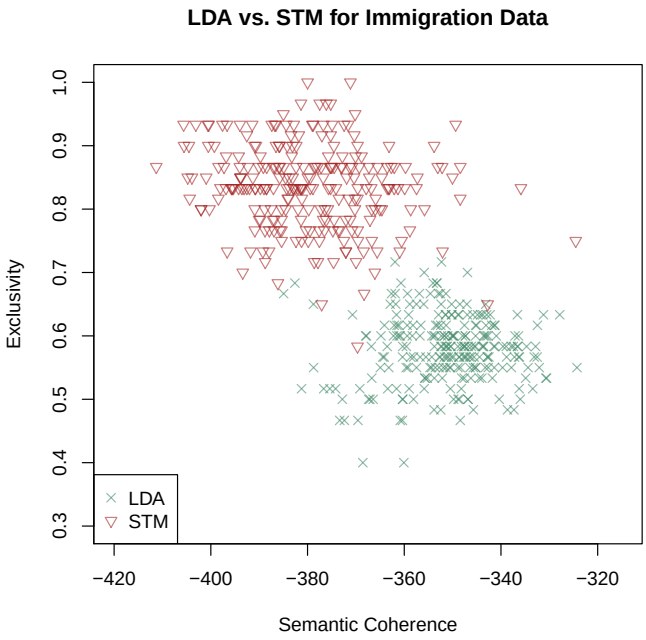



Figure 25: Semantic Coherence and Exclusivity for 250 runs of LDA and STM using different starting values.

We note that in general, STM provides more exclusive solutions where LDA favors semantic coherence. We note however that STM does provide a few solutions with the maximum level of semantic coherence attained by either model but with higher exclusivity than LDA. The plot also highlights the implicit tradeoff between coherence and exclusivity; that is, it is trivially easy to have either extremely high coherence (by having all topics have the same top words) or high exclusivity (by picking completely disjoint sets which do not co-occur in actual documents) and thus it is useful to examine both criteria in concert. We hope to explore whether these general trends hold across different document sets in future work.

In order to examine more closely examine results that are favorable to each model,

we randomly choose a high performing solution using the same steps discussed in the paper. First we discard any solution below the 75% quantile along each dimension. Then we randomly select a solution from the remaining models for analysis. While in actual data analysis it would be best to look at several options and choose the most useful representation of the data, for the purposes of fairly comparing the two algorithms we made one random selection and did not search over other alternatives.

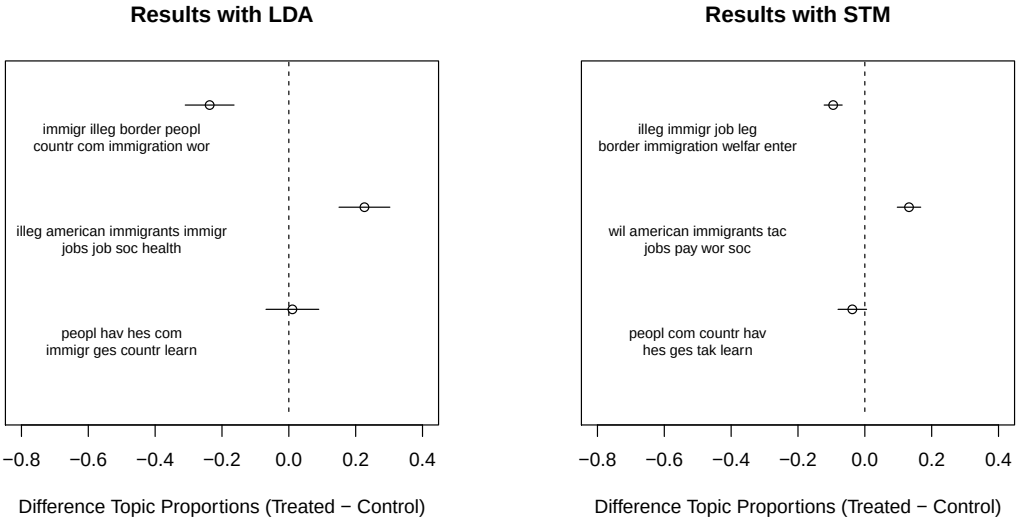



Figure 26: Treatment effects from a randomly selected high performing LDA and STM run on the immigration data.

We plot the two results in Figure 26 (aligning the results so comparable topics are on the same line). Words shown are the highest probability word stems (rather than the FREX words we use elsewhere in the paper, here we use highest probability words for comparability with standard LDA practice). Our model selection procedures are successful in that both solutions are reasonable and quite similar. The estimates of the treatment effects for LDA are higher but with much larger confidence intervals suggesting considerably more heterogeneity in the treatment and control groups.

This is consistent with the topics themselves which are more focused in the STM model. Looking over the different topics from the LDA estimation, we see that the words

are not very exclusive to topics: "immigr", "illeg", "people", "countr", and "com" all appear in multiple topics. In contrast, the STM has much more exclusive topics: there are no overlapping words across topics. Even though the topics between LDA and STM are quite similar, interpretation in the more exclusive STM topics is more straightforward. The topic reflects both that the respondent is talking about immigration and the reasons for the respondent's concern. By contrast in the LDA version of the topic, the user can only determine the general topic, which seems to be common across all estimated topics.

This analysis is intended to provide a single concrete example to show the contrast between LDA and STM with real data. We emphasize that these results should not be taken as universal and analysis was done to emphasize comparability between the two methods as opposed to best practice. Thus we highlight two points: (a) LDA will not generally produce larger treatment effect estimates with wider confidence intervals, and (b) a careful analysis would examine documents rather than just looking at highest probability words. Our analysis of the immigration data is contained in the main body of the paper, but this section's analysis helps to contrast the two approaches with actual documents, as a complement to the simulated data used above.

### 6.2.8 Comparison with Hand-Coded Results

Here we cover the specifics of our comparisons between our results and the results from the hand-coding. The first thing to note is our topics do not match up perfectly with the categories that the coders used. This should not be a surprise, since we are using an unsupervised method whereas the human coders received supervision via a set of instructions. In particular, our topics do not have many words that seem to evoke "concern" or "enthusiasm" for immigrants. However, this is somewhat consistent with the human coded results because the human coders coded very few documents as expressing these topics. RA 1 classified only 18% of those who took the survey as having either enthusiasm or concern for immigrants. RA 2 classified 23% as having either enthusiasm or concern. These low numbers are in line with our analysis in that we did not retrieve very many words related to concern or enthusiasm and hence our results are consistent with Gadarian and Albertson (2013)

| | No Concern | Concern |
|---|---|---|
| 0. Think | 0.76 | 0.23 |
| 1. Worried | 0.87 | 0.13 |

Table 3: Treatment and Concern for Immigrants, RA 1

| | No Concern | Concern |
|---|---|---|
| 0. Think | 0.65 | 0.27 |
| 1. Worried | 0.76 | 0.16 |

Table 4: Treatment and Concern for Immigrants, RA 2

Even though the topics and coder categories do not match perfectly, the vocabulary associated with Topic 1 is more closely in line with fear of immigrants and anger toward immigrants, and so we will compare Topic 1 to the fear and anger categories. Aggregating across the treatment conditions, RA1 classified 56% of respondents as having negative views of immigrants (either fear or anger), and RA2 classified 79% of respondents as having a negative view of immigrants. The data clearly has a much richer and present vocabulary that corresponds to fear and anger toward immigrants.

The treatment effect we uncovered using the Structural Topic Model corresponds to what Gadarian and Albertson (2013) find with human coders. To see this effect from the human coding, Table 6.2.8 and 6.2.8 shows the breakdown by treatment and control, where "Concern" refers to responses that were coded either into concern or enthusiasm. Tables 6.2.8 and 6.2.8 show the breakdown when looking at the negative views of immigrants that is captured by our Topic 1, where "Negative" refers to responses that were coded as either containing fear or containing anger toward immigrants. The treatment increases the likelihood of a response with fear or anger under the coding scheme, in line with what the topic model uncovers.

An additional way to compare our results with that of the human coders is to consider

| | Not Negative | Negative |
|---|---|---|
| 0. Think | 0.60 | 0.39 |
| 1. Worried | 0.26 | 0.74 |

Table 5: Treatment and Negative Views Toward Immigrants, RA 1

| | Not Negative | Negative |
|---|---|---|
| 0. Think | 0.32 | 0.60 |
| 1. Worried | 0.06 | 0.86 |

Table 6: Treatment and Negative Views Toward Immigrants, RA 2

the correlation between the hand-coding and the document-level probabilities of topics. Because the topics are different than the pre-determined hand-coded topics, we should not necessarily expect the predicted document proportions and the coders' classification to completely line up. This is particularly the case because the hand coders put the majority of posts into either fear or anger. However, we should see some relationship between the probability that a response is in the fear and anger topic estimated by the Structural Topic Model and the human coding of these same responses.

To examine document-level agreement, we take observations where the coders were in 100% agreement about the classification. We would expect documents where the coders agree to be the most clearly classifiable. Figure 27 presents a histogram of the documents by the predicted proportion of a document in that topic, coloring each document by the category given by the coders. The x-axis on the histogram is the predicted proportion of a document in Topic 1. Documents on the left side of the barplot have very low proportions of Topic 1, while on the right side of the histogram have very high proportions of Topic 1. The colors on the bar plot represent the coders' classification of the same documents. For example, all of the documents with between 0 and .1 proportion of Topic 1 are coded either as having enthusiasm or not given a category. On the other hand, almost all of the documents with over .8 of Topic 1 were either coded with fear or anger.

We do see a correlation between the topics assigned to documents by the topic model and the classification done by the coders. Documents with high proportions of Topic 1 are more likely to be coded with fear and anger, and are rarely coded with enthusiasm or not categorized. We might be able to, for example, use the topic probabilities generated by the topic model to predict whether the coders would categorize a post into either enthusiasm or no category with a fairly high success rate.

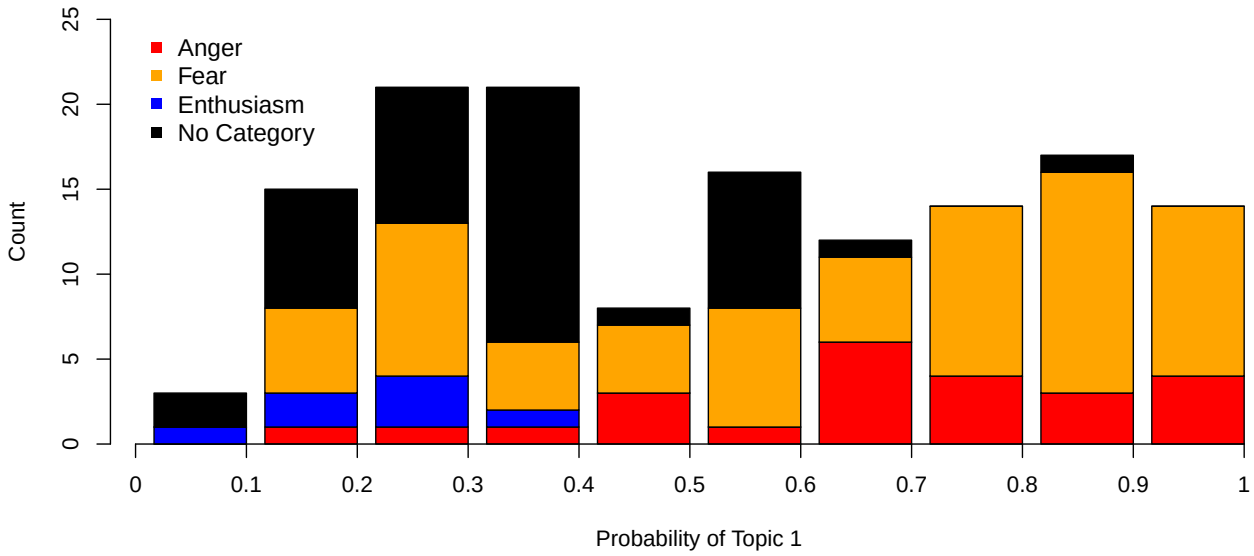



Figure 27: Coders' Classification Compared to Unsupervised Topic Proportions

## 6.3 Getting Started

In this section we discuss topics of interest to the applied user. Section 6.3.2 covers concerns about multiple testing as well as the application of pre-registration and false discovery rate methods. In Section 6.3.1, we describe the standard pre-processing algorithms we employ as well as introduce a new software tool for pre-processing and working with texts. In Section 6.3.3 we introduce a second software tool which provides an interactive visualization interface for exploring the results of topic models. In Section 6.3.4 we show how to use our approach to perform mediation analysis.

### 6.3.1 Text Preparation and Custom Pre-Processing Software

In our examples below, we adopt a set of procedures which are standard pre-processing steps in computational linguistics (Manning *et al.*, 2008). First we remove punctuation, capitalization, and stop words from the open-ended responses. Stop words are terms that are so common as to be uninformative. For example, the word "and" occurs in nearly all responses. As such, a respondent's use of the word "and" conveys little to no information

about the topical content of the response. If stop words like "and" are not removed, the model may discover topics that are described primarily by stop words, due to their relatively high frequency of occurrence. And even if stop words end up scattered across topics, these associations are likely spurious.

Next, the remaining words are *stemmed*. Stemming is the process of reducing inflected words to their lexical roots, or "stems." All the terms that share a common root are folded into a single term. This term need only be unique from all others; it does not need to be a real word. For example, in some implementations (including the one we use), the words happiness, happiest, and happier are all reduced to the stem "happi." Because the inflected variants of a common stem are probably all associated with a common topic, stemming generally improves performance with small data sets. Gains from stemming decrease with larger data sets, as the the model is less likely to spuriously assign variants of a common stem to different topics. In our analysis, we employ the canonical Porter Stemming Algorithm (Porter, 1980), which operates by algorithmically stripping away the suffixes of words and returning only the roots, or stems, of the original words.

The remaining terms are then transformed into a vector-space representation called a term-document matrix (tdm). For example, let $T$ index the unique terms in the data and $D$ index the documents. Let $M$ be a tdm. Then $M = [m_{t,d}]_{T \times D}$, where cell $m_{t,d}$ is the number of times term $t$ occurs in document $d$.[22] Column $m_{*,d}$ is then a vector representation of document $d$, which can be analyzed with standard techniques from multivariate analysis. All statistical analyses are computed on the tdm. To facilitate we wrote a multi-featured `txtorg` text management interface discussed below.

Notably, by discarding the structure and linear ordering of the words in the documents, the tdm makes the simplifying assumption that documents can be represented as an unordered "bag of words." While perhaps counter-intuitive, this approach yields a reasonable approximation while greatly improving computational efficiency.[23] We emphasize

[22] Occasionally, cell values are indicators for whether or not term $t$ appeared in document $d$.

[23] Typically we use the space of single words, called unigrams. We can easily incorporate a level of word order by incorporating bi-grams or tri-grams. Unfortunately the number of terms grows extremely quickly. The consensus in the literature is that unigrams are sufficient for most tasks (Hopkins and King,

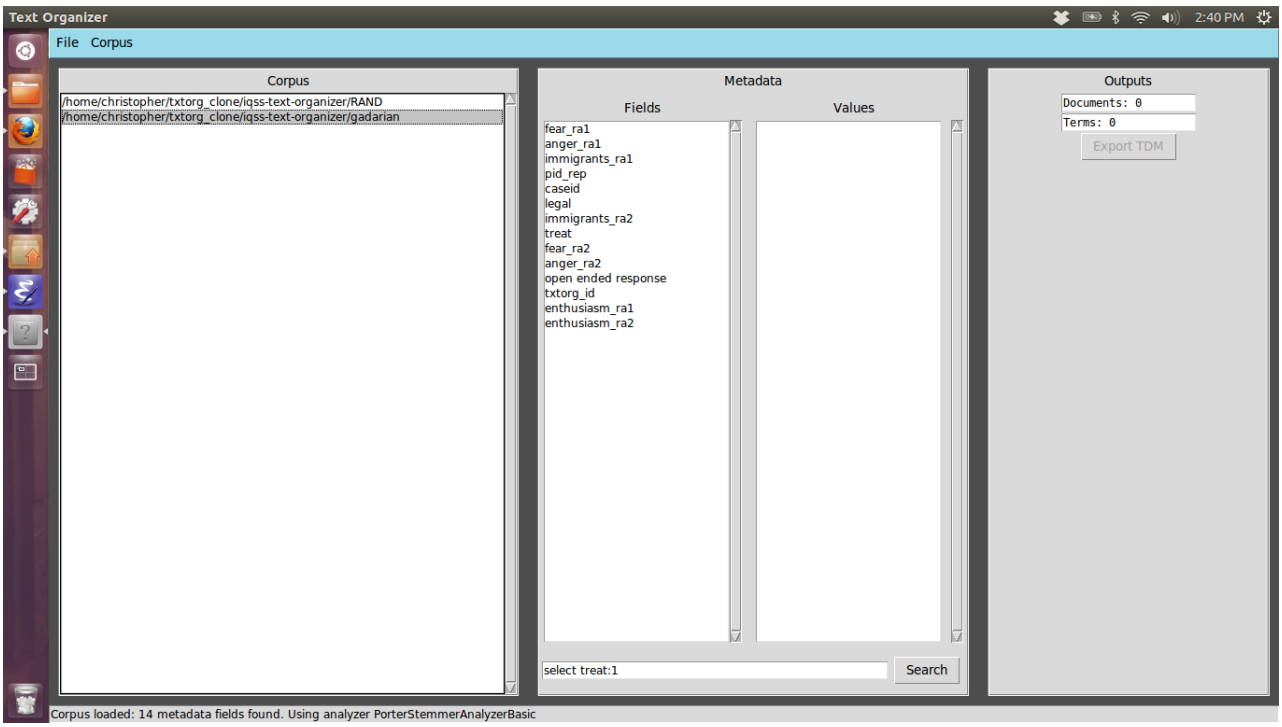



Figure 28: Screenshot of txtorg

that these data preprocessing choices are separate from our method in the sense that the analyst can choose to apply them or not based on their particular case.

To facilitate the proposed workflow we developed a software package `Txtorg` [Blinded For Review]to prepare our data for analysis. `Txtorg` is a scalable Python-based tool with a full-featured graphical user interface that makes the management of textual data simple and intuitive for everyone. By leveraging the power of Apache Lucene (The Apache Software Foundation, 2013), `Txtorg` enables users to store and search large amounts of textual data that would otherwise be unwieldy. A screenshot of the `Txtorg` user interface is presented in Figure 28.

`Txtorg` uses Apache Lucene to create a search index, which allows for local storage and quick searching, even with large corpora. Search indices, like those created by `Txtorg`, allow queries to return the relevant documents without scanning the entire corpus. Moreover, the index created by `Txtorg` is compressed to a fraction of the size of the original

2010)

corpus, and from it, full corpora can be recovered.

A typical workflow with `Txtorg` might then proceed as follows. Steps 4 - 7 may be repeated infinitely without repeating earlier steps.

1. **Collect data:** Here, `Txtorg` is agnostic, as the program can be used with a wide range of data types, ranging from survey responses to Chinese poetry written in Mandarin.

2. **Organize data:** Depending on the original source from which the data were collected, reformatting may be necessary. Specifically, `Txtorg` can import corpora stored in one of three formats - a field (column) in a csv file (common for data generated by a survey), a directory containing the individual documents as `.txt` files, or a csv with a field containing the file paths to the individual documents in the corpus. If the data are imported as a field in a csv file, the remaining columns in the csv are imported as metadata. For example, a researcher might upload responses to an open-ended survey question stored as a field in a csv. The remaining columns may be responses to closed-ended questions, such as sex, age, and education. If the data are not imported as a field in a csv, metadata may be imported separated as a csv.

3. **Import the data:** Once the data are properly formatted, this step is trivial. One need only launch the `Txtorg` user interface and select the import data option from a drop-down menu. The corpus is then indexed and the index is saved locally.

4. **Preprocessing:** Next, the user may choose to implement any of several forms of preprocessing supported by `Txtorg`. Most importantly, `Txtorg` supports stemming, stop word removal, and phrase indexing. `Txtorg` implements the canonical Porter stemming algorithm, which users may simply turn on and off at any time. Similarly, users can turn stop word removal on and off as they please, and may change the list of stop words at any time. Phrase indexing is somewhat less flexible. By default, the `Txtorg` indexes all single-word tokens (unigrams). If a researcher wishes to search and index tokens longer than a single gram (for example, "border security"),

these grams must be imported at step 3, when the user imports the data. If after importing the data, the user wishes to index phrases longer than a single gram, the corpus may be re-indexed with an updated list of phrases.

5. **Searching the corpus:** Next, the user may search and subset the index by the terms in the individual documents or by document metadata. The search function is sophisticated, allowing users to string together multiple searches. For example, a researcher might be interested in the set of documents written by individuals in specific treatment conditions, or by men between the ages of 18 and 30 who are either Hispanic or black mentioning "border security" or "national security." `Txtorg` can implement searches of this form and complexity, as well as more simple searches. Finally, users may also select the entire corpus.

6. **Exporting the data:** After selecting the documents of interest, users can export the data as either a term-document matrix or as entire documents. `Txtorg` supports two primary TDM formats, the LDA sparse matrix and the standard, flat csv. Documents exported in their entirety are written to a directory as txt files.

7. **Analysis:** Finally, the user may conduct any sort of analysis on the exported data, including those based on the structural topic model explained in this paper, but also analyses of other forms, from LDA to manually reading the documents in their unprocessed entirety.

Perhaps `Txtorg`'s most compelling feature is its speed, especially with large corpora. Sequentially searching through all responses to subset on a specific term is incredibly slow, and literally impossible in some cases. By creating an index, search time decreases by orders of magnitude. One can easily analyze just those responses containing "immigration reform," then seconds later export and analyze the TDM for responses containing "border security".

A second selling point is the program's ability to subset and sort data without editing the original index. Perhaps a researcher conducts an initial analysis on responses from college students, then in a subsequent project, wishes to examine racial minorities. Or

one might want to examine the responses mentioning "President" separately from the full corpus. Subsetting responses with a simple script would be tedious in cases where the data are small, and impossible where they are large. With `Txtorg`, after the corpus is indexed, researchers can quickly pull from it as many unique TDMs as they like, at any point in time.

### 6.3.2 Multiple Testing

There is reasonable concern that with large groups of topics multiple testing will cause the researcher to find spurious relationships between covariates and topics. Put intuitively the idea can be stated that as the number of topics grows, iteratively testing treatment effects over all topics in the model will cause a false positive significant result. We discuss why this concern might not be as troubling as it first appears, and then discuss how certain methods can be used to address it.

On one view, standard p-values are misleading in this setting because the null distribution is not particularly believable (Gelman *et al.*, 2012). That is, we chose to incorporate a covariate into the model because we believe that there is an effect on word choice and it is unlikely that any effect would be exactly zero. Thus the question is really more about estimating effect sizes correctly rather than rejections of a sharp null. While this view is appealing, it still does not address the nagging concern that comparison over many topics will tend to cause us to see spurious relationships where there are none.

Even if there is a spurious relationship found by the model, it does not mean it is a substantively interesting topic. For the researcher to come to an incorrect conclusion there must be a significant covariate on a topic that has a semantically coherent and intellectually relevant meaning. For those who still have concerns, we conducted a permutation test in the previous section. There we showed that even when searching for the maximum effect in the permuted data, the true effect with the real treatment indicator was an enormous outlier. This suggests that spurious effects are not as common as people might fear.

Still for those willing to expend additional planning and effort, it is possible to further hedge against the risk of false positives. In some experimental settings, researchers limit

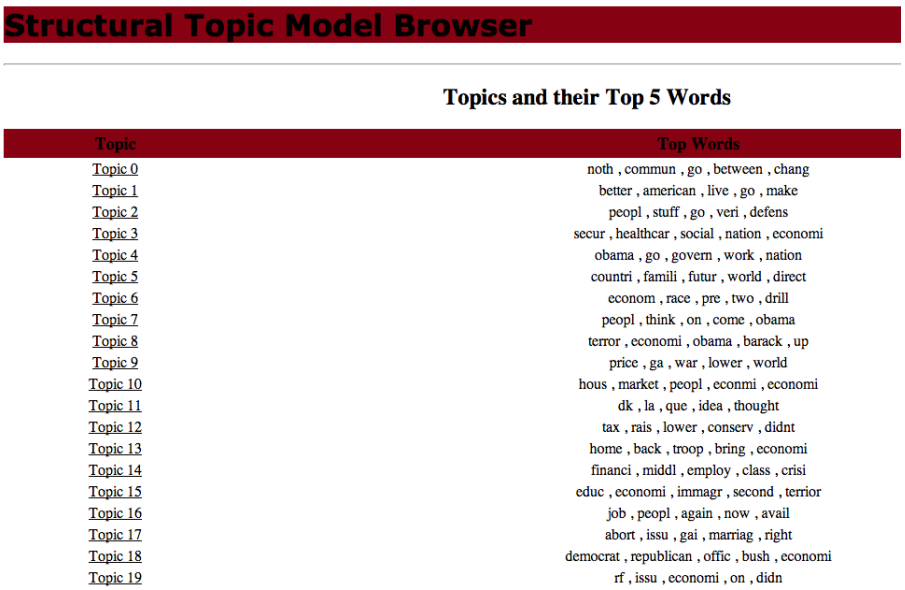
their risk by adopting pre-analysis plans. These plans specify which relationships they will test and what they will expect to find. In the topic model setting, researchers can specify word groups that they expect to co-occur and the effect they would expect to find on such topics. This could be as informal as expecting to find that issues about the 'economy' are likely to appear in the ANES most important problems and the likely relationship with household income. On the more rigorous side, analysts could specify a list of word stems beforehand and only consider topics which contain a certain percentage of those key words.

An alternative approach is to condition on the model results but augment testing with False Discovery Rate methods (see Efron (2010) for an overview). Conditioning on the model we treat the topic proportions as observed data and apply standard corrections as with any other dataset. These methods have been shown to be effective in genetics where there are often tens of thousands of tests; by contrast, we will typically have far fewer topics.

### 6.3.3 Post-analysis visualization

For those unfamiliar with topic modeling the output from standard software can be overwhelming. Models typically contain thousands of parameters which must be summarized for user interpretation. Typically an analyst will develop custom functions which summarize parameters of interest for their specific application. In addition to a rich set of functions for performing the types of visualizations reported in this paper, we have also been developing a more general visualization and model summary tool which builds on recent efforts in computer science (Gardner *et al.*, 2010; Chaney and Blei, 2012).

Here we highlight our ongoing work on the Structural Topic Model Browser (STMB). The STMB visualizes output from the model in a standard web browsing format. It incorporates information at the corpus level as well as individual views of each document. We believe that browsers of this sort serve a crucial role in connecting the end user to the original text, and help to make good on the promise of topic model as a method for *computer-assisted reading*. The software will be made available as an open-source project on Github which will allow easy access for the end user as well as other developers who



**Structural Topic Model Browser**

**Topics and their Top 5 Words**

| Topic | Top Words |
|---|---|
| Topic 0 | noth , commun , go , between , chang |
| Topic 1 | better , american , live , go , make |
| Topic 2 | peopl , stuff , go , veri , defens |
| Topic 3 | secur , healthcar , social , nation , economi |
| Topic 4 | obama , go , govern , work , nation |
| Topic 5 | countri , famili , futur , world , direct |
| Topic 6 | econom , race , pre , two , drill |
| Topic 7 | peopl , think , on , come , obama |
| Topic 8 | terror , economi , obama , barack , up |
| Topic 9 | price , ga , war , lower , world |
| Topic 10 | hous , market , peopl , econmi , economi |
| Topic 11 | dk , la , que , idea , thought |
| Topic 12 | tax , rais , lower , conserv , didnt |
| Topic 13 | home , back , troop , bring , economi |
| Topic 14 | financi , middl , employ , class , crisi |
| Topic 15 | educ , economi , immagr , second , terrior |
| Topic 16 | job , peopl , again , now , avail |
| Topic 17 | abort , issu , gai , marriag , right |
| Topic 18 | democrat , republican , offic , bush , economi |
| Topic 19 | rf , issu , economi , on , didn |

Figure 29: ANES Structural Topic Model Browser Index Page

may wish to adapt the code or add features.

To use the browser, the user passes the original documents and covariate information as well as the model output to a visualization script written in Python. In addition, the user specifies the number of words to be displayed for a topic. In this example, we're exploring a structural topic model of the ANES data that combined all of the most important problem questions with 55 topics and 5 words displayed per topic.

The Index Page for the website lists each topic and the most probable words under that topic. Figure 29 shows the Index Page of the browser for the ANES dataset.

From the browser page, each of the topic listings, e.g. "Topic 10", function as links to individual Topic Pages. The Topic Page contains a number of important metrics about the topic's fit and relationship to documents. In addition to the top words as displayed on the index page, it contains the top FREX-scored words, as described in Section 6.1.4.

The Topic Page contains the topic's expected frequency across a corpus, which is the mean proportion of words across the documents that are assigned to this topic. Furthermore, the Topic Page displays the topic's semantic coherence score, a measure of the topic's fit within documents, where larger numbers indicate a better fit (Mimno *et al.*,

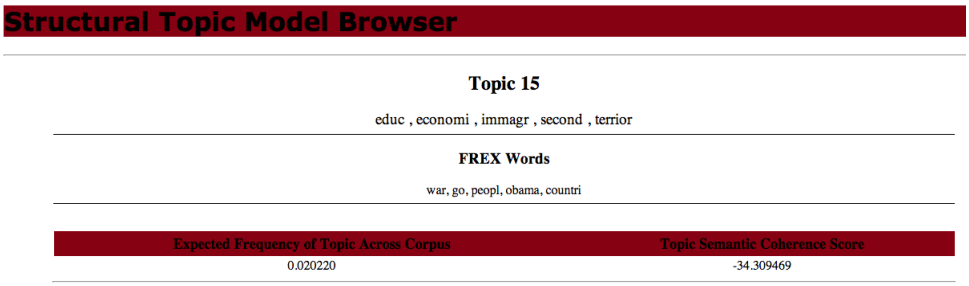



Figure 30: Topic Page for Topic 15

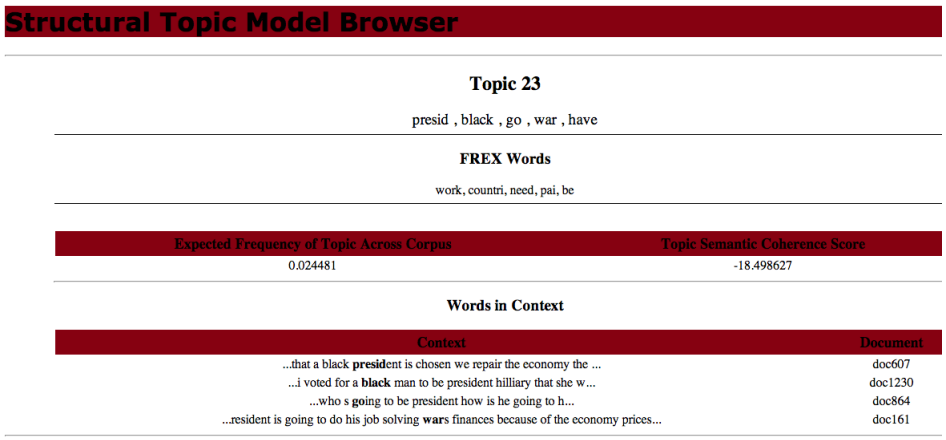



Figure 31: Top of the Topic Page for Topic 23

2011).

Below the expected frequency and semantic coherence scores, the browser displays the top words in context as in Gardner *et al.* (2010). For each top word, there is a sample from one of the documents with a high proportion of words assigned to the specific topic and with high incidence of the specific word. This sample presents the word in bold, with forty characters to each side, and the document listed to the right. This allows the user to connect the top words in the topic to an instance in which this has been used. Figure 31 shows the top of the Topic Page for topic 23, including the context section. The topic's top two words are 'president' and 'black', and we see in the context two documents describing how the participants specifically voted for a black president.

After the words are presented in context, we present the top 25 documents by the

**Top Documents**

| Document Name | Proportion of Words in Document Assigned to this Topic |
|---|---|
| doc564 | (0.917) |
| doc806 | (0.868) |
| doc1340 | (0.866) |
| doc1000 | (0.860) |
| doc1397 | (0.790) |
| doc849 | (0.773) |
| doc332 | (0.752) |
| doc995 | (0.693) |
| doc1599 | (0.688) |
| doc1813 | (0.679) |
| doc607 | (0.641) |
| doc161 | (0.630) |
| doc427 | (0.629) |
| doc548 | (0.629) |
| doc500 | (0.600) |
| doc1888 | (0.600) |
| doc1253 | (0.595) |
| doc1842 | (0.594) |
| doc920 | (0.570) |
| doc50 | (0.563) |
| doc864 | (0.555) |
| doc1230 | (0.541) |
| doc74 | (0.505) |
| doc371 | (0.493) |
| doc1882 | (0.471) |

Figure 32: Top Documents for Topic 23

proportion of words assigned to this topic. Each document name is also a link to that document's page. This allows the user to see roughly the level at which documents that use this document heavily pick words from this topic, as well as providing links to the individual document pages, described below. Figure 32 shows the 'Top Documents' section of the Topic Page for topic 23.

The last element on the topic page Topic Page is a set of graphs of the documents' expected topic proportions plotted against the covariate values for that document. This allows the user to visualize how the covariate impacts the topic's fit to individual document and to spot any patterns for further analysis. Figures 33 show the covariate plot for topic 23 with respect to the Age.

Each document has its own page, linked to by the Topic Pages. The Document Page contains a list of the Top 10 topics associated with this document, the proportion of words in the document assigned to that topic, and the topic $n$ words in that topic. Below, the browser displays the entire text of the document. Figure 34 is an example of a Document Page from the ANES topic model. Document 1586 scores highly in Topic 15, described above. We see from the text and topic fit, particularly the use of topic 12, that this is an example of a conservative response to the ANES questions.

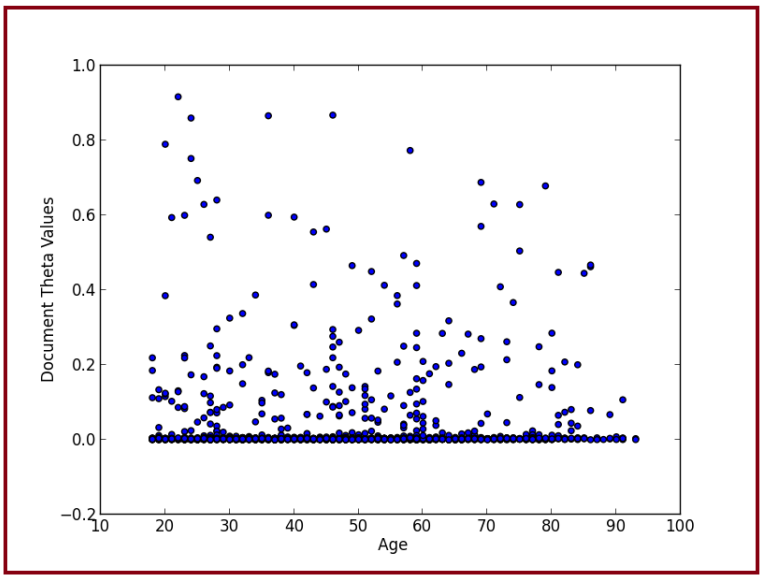



Figure 33: Age Covariate, Topic 23

## Structural Topic Model Browser

**doc1586**

**Top Topics in this Document:**

| Topic | Topic Weight | Topic Words |
|---|---|---|
| Topic 15 | 0.807 | educ economi immagr second terrior |
| Topic 12 | 0.031 | tax rais lower conserv didnt |
| Topic 17 | 0.031 | abort issu gai marriag right |
| Topic 52 | 0.027 | know don dont realli on |
| Topic 36 | 0.007 | militari moral economi oversea re |
| Topic 09 | 0.006 | price ga war lower world |
| Topic 20 | 0.005 | debt peopl much govern problem |
| Topic 02 | 0.003 | peopl stuff go veri defens |
| Topic 04 | 0.003 | obama go govern work nation |
| Topic 26 | 0.003 | keep everyth peopl person be |

**Document Text:**
smaller government lower taxes strong defense conservative values judges dont legestrate from the beach economy moral issues upholding constitution no abortion second amendmment first amendment oposing to the fairment documentx terriorism economy no dont know of any

Home

Figure 34: Document 1586

Overall, the Structural Topic Model Browser provides a way for the user to use the output of the topic model to explore the relationship between text and the covariates. By mixing graphical representation, samples from the original texts, and quantitative measures of fit, the browser facilitates broader understanding of the analysis and intuitive ideas about patterns observable in the text.

### 6.3.4 Mediation Analysis

The output of the structural topic model can easily be combined with standard analysis techniques. In this section we show how this can be done with mediation analysis, though we stress that this application does not jointly estimate all parameters which would be useful future work. In this application we examine how ex-post strategy choice descriptions mediate the relationship between Rand *et al.* (2012)'s experimental intervention of encouraging intuitiveness or deliberation. In particular, for every subject we know their treatment assignment, their explanation of strategy choice, and their contribution. We can then take their explanation of their strategy choice and analyze it using the STM, utilizing the information we know about their treatment assignment. Then, for each subject, we have an estimate of $\theta$ for each of the topics (proportion of words from that topic). We can then apply standard mediation analysis using this $\theta$ as the mediator.

To conduct the mediation analysis we utilize the `mediation` package in R. This entails fitting two parametric models. The first is a regression of the topic representing intuitive thinking on the treatment condition. The second is a regression of normalized contributions $(0, 1)$ on both the treatment and the intuitive topic proportion. We expect the mediation effect to be positive, with the treatment positive impacting the intuition topic, and the intuition topic positively influencing contributions. Use of a tobit model for the contributions equation produced similar results.

Table 35 presents the average causal mediation effect along with 95% confidence intervals. We observe a positive mediation effect of approximately .09. An alternative form of analysis could use the open ended text that respondents wrote when first responding to the encouragement to write about an intuitive topic. This would avoid the problem of using ex-post rationalizations as a mediating variable.

# References

Anandkumar, A., Foster, D., Hsu, D., Kakade, S., and Liu, Y.-K. (2012). A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 25*, pages 926–934.

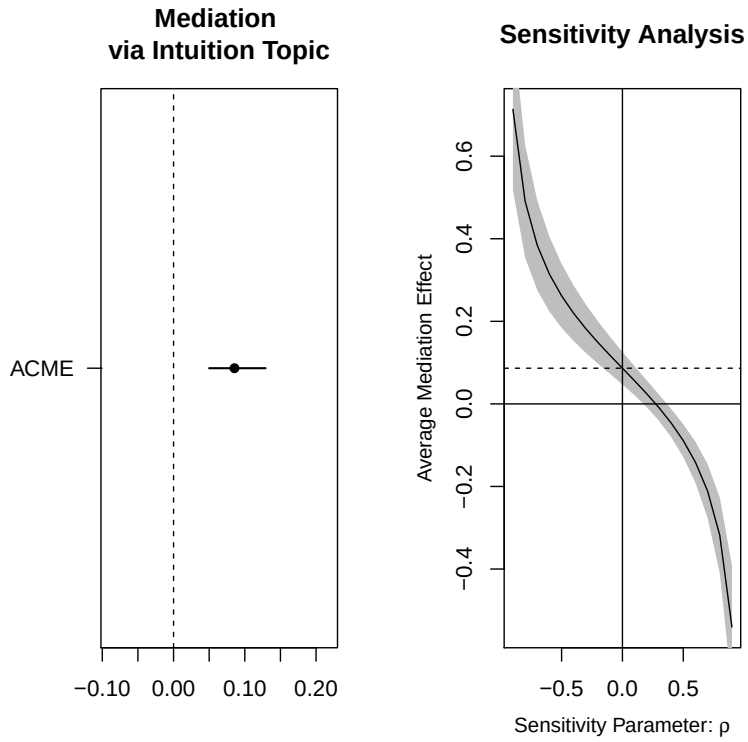



Figure 35: Left side of presents the average causal mediation effect (ACME) of intuition topic on normalized contributions using intuition encouragement design. Right side presents a formal sensitivity analysis as described in Imai *et al.* (2011).

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596.

Bischof, J. and Airoldi, E. (2012). Summarizing topical content with word frequency and exclusivity. *arXiv preprint arXiv:1206.4631*.

Bishop, C. (2007). *Pattern Recognition and Machine Learning*. Springer, Cambridge, MA.

Blackwell, M., Honaker, J., and King, G. (2011). Multiple overimputation: A unified approach to measurement error and missing data.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, **55**(4), 77–84.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *AAS*, **1**(1), 17–35.

Blei, D. M., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *JMLR*, **3**, 993–1022.

Buntine, W. L. and Jakulin, A. (2005). Discrete component analysis. In *SLSFS*, pages 1–33.

Buot, M. and Richards, D. (2006). Counting and locating the solutions of polynomial systems of maximum likelihood equations, i. *Journal of Symbolic Computation*, **41**(2), 234–244.

Chaney, A. and Blei, D. (2012). Visualizing topic models. *Department of Computer Science, Princeton University, Princeton, NJ, USA*.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *NIPS*.

Druckman, J., Green, D., Kuklinski, J., and Lupia, A. (2006). The growth and development of experimental research in political science. *American Political Science Review*, **100**(4), 627.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Eisenstein, J., Ahmed, A., and Xing, E. P. (2011). Sparse additive generative models of text. In *Proceedings of ICML*, pages 1041–1048.

Gadarian, S. and Albertson, B. (2013). Anxiety, immigration, and the search for information. *Political Psychology*.

Gardner, M., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., and Seppi, K. (2010). The topic browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization. MIT Press*.

Geer, J. G. (1988). What do open-ended questions measure? *The Public Opinion Quarterly*, **52**(3), 365–371.

Geer, J. G. (1991). Do open-ended questions measure "salient" issues? *The Public Opinion Quarterly*, **55**(3), 360–370.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*, volume 3. Cambridge University Press New York.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, **5**(2), 189–211.

Gerring, J. (2001). *Social science methodology: A unified framework*. Cambridge Univ Pr.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, **18**(1), 1.

Grimmer, J. (2011). An introduction to bayesian inference via variational approximations. *Political Analysis*, **19**(1), 32–47.

Grimmer, J. and King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, **21**(2).

Hainmueller, J., Hopkins, D., and Yamamoto, T. (2012). Causal inference in conjoint analysis: Understanding multi-dimensional choices via stated preference experiments.

Hopkins, D. (2012). The exaggerated life of death panels: The limits of framing effects on health care attitudes.

Hopkins, D. and King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, **54**(1), 229–247. http://gking.harvard.edu/files/abs/words-abs.shtml.

Hornik, K. (2005). A CLUE for CLUster Ensembles. *Journal of Statistical Software*, **14**(12).

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, **105**(4), 765–789.

Iyengar, S. (1996). Framing responsibility for political issues. *Annals of the American Academy of Political and Social Science*, **546**.

Jain, A. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, **31**(8), 651–666.

Kelley, S. (1983). *Interpreting elections / by Stanley Kelley, Jr*. Princeton University Press, Princeton, N.J. :.

King, G., Schlozman, K., and Nie, N. (2009). The changing evidence base of social science research. *The Future of Political Science: 100 Perspectives*, pages 91–93.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Sage, New York.

Krosnick, J. A. (1999). Survey research. *Annu. Rev. Psychol.*, **50**, 537–567.

Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, **97**(2), 311–331.

Lazarsfeld, P. F. (1944). The controversy over detailed interviews - an offer for negotiation. *Public Opinion Quarterly*, **8**, 38–60.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational social science. *Science*, **323**(5915), 721–723.

Lombard, M., Snyder-Duch, J., and Bracken, C. (2006). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human communication research*, **28**(4), 587–604.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

Mimno, D. and Blei, D. (2011). Bayesian checking for topic models. *Empirical Methods in Natural Language Processing*.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of*

*the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Paisley, J., Wang, C., and Blei, D. (2011). The discrete innite logistic normal distribution for mixed-membership modeling. In *AISTAT*.

Papadimitriou, C. H. and Steiglitz, K. (1998). *Combinatorial optimization: algorithms and complexity*. Courier Dover Publications.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.

Quinn, K., Monroe, B., Colaresi, M., Crespin, M., and Radev, D. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, **54**(1), 209–228.

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, **489**(7416), 427–430.

RePass, D. E. (1971). Issue salience and party choice. *The American Political Science Review*, **65**(2), pp. 389–400.

Schuman, H. (1966). The random probe: a technique for evaluating the validity of closed questions. *American Sociological Review*, pages 218–222.

Schuman, H. and Presser, S. (1996). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. SAGE.

Simon, A. and Xenos, M. (2004). Dimensional reduction of word-frequency data as a substitute for intersubjective content analysis. *Political Analysis*, **12**(1), 63–75.

Slapin, J. B. and Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, **52**(3), 705–722.

Sontag, D. and Roy, D. (2009). Complexity of inference in topic models. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*. Citeseer.

The Apache Software Foundation (2013). Apache Lucene.

Treier, S. and Jackman, S. (2008). Democracy as a latent variable. *American Journal of Political Science*, **52**(1), 201–217.

Wang, C. and Blei, D. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM.

Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*.

Wolpert, D. and Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transaction on Evolutionary Computation*, **1**(1), 67–82.

Zou, J. and Adams, R. (2012). Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems 25*, pages 3005–3013.