Three Essays on the Study of Nationalization with Automated Content Analysis

Joseph L. Sutherland

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee of the
Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

ABSTRACT

Three Essays on the Study of Nationalization with Automated Content Analysis

Joseph L. Sutherland

In three papers, I consider two questions of nationalization in American politics, and one question of the methodology necessary to study them.

Nationalization is the process by which local politics become more like national politics on the basis of political issues and electoral engagement. It is usually measured using the difference in presidential and state-level electoral returns over time. To expand the study of nationalization, I use automated content analysis to derive new measures for the phenomenon's study based on political text. In particular, I apply automated content analysis via latent dirichlet allocation to code for salient topics in text from national political agenda speech, local agenda speech, and state laws.

The primary source for these local agenda codes is a novel database of State of the State addresses, which are like presidential State of the Union addresses, but are delivered by governors. I developed the database over the past seven years as part of this dissertation; it draws from all 50 States, and the earliest captured addresses date to the year 1893. The secondary sources for these codes are the State of the Union addresses and a corpus of laws passed by state legislatures. I utilize the codes from these naturally distinct text corpora to study the nationalization of the political agenda, and how nationalized elections relate to the production of salient laws.

The comparison of naturally distinct texts, however, is problematic and requires further examination. To that end, the first paper, "A Theory and Method for Pooling Naturally Distinct Corpora" concerns the theory and method for why we should be able to use, pool, and compare the computer-generated codes from these naturally distinct text corpora to study nationalization. I propose a theoretical framework with which the researcher may defend the pooling of corpora, and introduce an empirical approach to testing for absolute comparability, the *delta-statistic*. While statistics like the Akaike Information Criterion (AIC) and penalized log likelihood can help the researcher to determine if a model fits the pooled corpora better than the corpora separately, the delta-statistic relies on a strong theory of latent traits to evaluate the absolute quality of a pooled model. This is especially important when it is impossible to evaluate ground truth fit because some data are unlabeled.

The second paper, "Have State Policy Agendas Become More Nationalized?" examines whether the nationalization of state policy agendas is related to the nationalization of gubernatorial elections. The analysis shows that State agendas, as laid out in the State of the State addresses, have become more similar to each other over time. It also shows that State agendas have become more similar to the national agenda, as laid out in the State of the Union addresses. Finally, I demonstrate an increasing relationship between the similarity in the agenda and the nationalization of elections. The findings suggest that the nationalization of the agenda is a significant and related factor to the nationalization of elections.

The third paper, "Can States Govern Effectively When Politics Are Nationalized?" considers the question of whether electoral nationalization moderates the relationship

between divided government and legislative productivity in the states. I find a null effect of divided government on salient lawmaking ability, and that nationalization of state legislatures has generally *decreased* the production of salient laws. The result holds even though nationalization is unrelated to the ability of our state governments to take action on salient issues during times of divided government. The findings suggest that behavioral factors driving lawmaker decisions may be more to blame for lawmaking defects than institutional ones.

Taken together, the essays demonstrate the value of text analysis to the analysis of nationalization and other research topics in American politics.

Contents

List of Figures

List of Tables

For Sarah.

Preface

This three-paper dissertation represents the culmination of nearly seven years of data collection and research that came about as a result of a National Science Foundation research proposal I prepared in 2013. My initial idea was that I could study the power of American executives by using data from gubernatorial agenda speeches, also called "State of the State" addresses. I quickly realized upon socializing the project with other scholars that these speeches may also be used to study a number of other topics in American politics. One of those topics was nationalization; its importance became quite apparent during my time at Columbia, which overlapped with an intensifying environment of polarized politics during the Obama and Trump administrations.

The study of nationalization generally examines correlation in national and local patterns of political issues and electoral engagement. When we study nationalization, we therefore pool national and local data together to make inferences and draw conclusions. When we use text to capture those patterns and look for correlations, we are also pooling national and local data together. The data I intended to use in order to study nationalization were, rather than electoral returns, data of a textual sort.

Justin Phillips, Bob Shapiro, and Suresh Naidu were the first to point out that there exists no best practice approach to doing such an analysis. When may we pool,

and compare, different sources of text? How may we compare them? There was no obvious answer, and so I embarked on an effort to place a framework around why we might expect to be able to compare different corpora. As it turns out, the choice to compare (or pool) corpora is a subset of the general problem of the comparison of disparate observations, addressed by researchers like Larry Bartels and Chris Achen in their studies of joint scaling during the 1970's and 1980's.

While this work is principally addressed to political scientists, it may also be of interest to other scholars in the social and computer sciences. The methods this dissertation implements borrow widely from methods developed in other fields. Text analysis is a niche subject with which political scientists are generally unfamiliar. Many, however, understand the process of measurement and the use of regression analysis to support an argument. I address this general audience, with some parts reserved for methodologists with experience in the area.

I have been fortunate to have a lot of help from some of the best methodologists and empirical researchers in the field. To name everyone who had a hand in the production of this work would be impossible. I am most thankful for the advice of Bob Shapiro, Suresh Naidu, and Justin Phillips, who encouraged me to pursue this path of research and providing endless support along the way. I am honored to have defended this work in front of a committee of scholars including Bob Erikson and Mark Hansen, who provided critical feedback included in the final revisions. I am thankful to Elliot Ash, Jon Rogowski, John Marshall, Arthur Spirling and the text-as-data community at New York University for numerous fortuitous discussions and feedback on several other projects that informed my understanding of this project and text data

profoundly. Without the diligent support of my research assistants Scott Solomon, Christian Baehr, Peter Rosenquist, Allen Hao, Sydney Greene, Jared Dauman, and Deepa Devanathan, this would reseasrch never have been possible.

I am privileged to have worked with Dan Butler, Jim Gibson, Jon Rogowski, Johannes Urpelainen, Ryan Kennedy, Don Green, David Broockman, Adam Zelizer, Michael Berkman, Eric Plutzer, Burt Monroe, and several other collaborators at Washington University in St. Louis, Columbia University, Emory University, Pennsylvania State University, and the University of California at San Diego during the course of this project. Dan Butler and I have worked together for the better part of the decade to collect and analyze the State of the State addresses, and I am deeply indebted to him for his guidance and support of the project. It is a testament to their patience that I have developed the research ability I have today. This research competed for National Science Foundation funding and was awarded an honorable mention, which provided super-computing resources with which some of this work was completed.

Finally, I would be remiss if I did not pay immense gratitude to my family and friends, who provided immeasurable amounts of support through the ups and downs during the course of my research. My everlasting love Sarah moved to New York with me, spent countless hours in our one-bedroom apartment living room prepping me for comprehensive exams, and served as a constant source of clarity in reviewing draft after draft for so many years. My family and friends stuck by us through joys, health issues, marriage, and the death of my mother, who taught me the value of perseverance and passion. It is the joy of a lifetime to have "my people" by my side.

Introduction

In three papers, I consider two questions of nationalization in American politics, and one question of the methodology necessary to study them. Nationalization is the process by which local politics become more like national politics on the basis of political issues and electoral engagement. It is usually measured using the difference in presidential and state-level electoral returns over time. To expand the study of nationalization, I use automated content analysis of political texts to derive new measures of nationalization based on text. In particular, I estimate the difference between agenda codes from the presidential State of the Union addresses to agenda codes from a new database of gubernatorial State of the State addresses, and between codes from the State of the State addresses to codes from a database of the laws passed each year by state legislatures. The State of the States database, which I developed over the past seven years as part of this dissertation, draws from all 50 States and dates since 1893. I apply automated content analysis via latent dirichlet allocation to code for agenda topics in the text documents. The comparison of these codes from naturally distinct text corpora enables me to study the nationalization of the political agenda, and how nationalized elections relate to lawmaking.

The comparison of naturally distinct texts, however, is problematic and requires

further examination. The first research question, addressed in "A Theory and Method for Pooling Naturally Distinct Corpora," is when and why the researcher should be able to compare computer-generated codes from two or more sources of text. I consider the unit of analysis for this question to be a singular political actor, who is the source of several textual documents. The documents, in turn, are each composed of several paragraphs of text, which may comprise transcribed political speech from the floor of the Congress, a governor's expressed political agenda, a newspaper article, or a statute passed into law by a state Congress. The sources (units) of text are drawn from one or more underlying populations. This creates a nested structure, which in order of decreasing magnitude is: the population or populations, the political actor (unit, source), the document, and words. The aim of the first research question is to produce a test of whether the political actors were drawn from the same population, given that we know nothing else about them other than their speech.

The question is important because published research papers make exactly that assumption, but do not provide any positive evidence for it. For example, studies in the areas of nationalization and polarization like Gentzkow and Shapiro (2010), Groseclose and Milyo (2005), and Martin and McCrain (2019) use text from speeches by Congresspeople, the News, and Presidents to create a model that codes for ideology in one population of actors and use that model to code for ideology in other units. In these works, the unit of analysis is the "political actor," but the political actor has, in each case, two different levels. In the case of Gentzkow and Shapiro (2010), Newspapers and Congresspeople are pooled together; in the case of Groseclose and Milyo (2005), interest groups and Congresspeople are pooled together, and; in the case

of Martin and McCrain (2019), Television stations and Congresspeople are pooled together. These works ultimately draw conclusions based on predicted ideology codes for one of the units (*e.g.*, the Newspapers) based on a model trained on the other (*e.g.*, the Congresspeople); however, they lack the "ground truth" scores for ideology that we would usually use to empirically test for the fit of the model on the novel unit, or in the absence of the empirical opportunity, assert a theory for why the two are related. This setup has created a circumstance where inferences are made for units based on strong, but untestable, assumptions.

This dissertation enhances our ability to test those assumptions. It does so by showing how the modeling approach just described is also an application of joint scaling, and invokes joint scaling theory to derive an empirical testing framework. Joint scaling is a method to allow for the direct comparison of the same latent political trait, such as ideology or preference, across two naturally distinct populations (*i.e.*, the House and the Senate, the Congress and the President, or the President and federal agencies). Research in joint scaling has repeatedly posited that the use of a model trained on one population to make predictions on another makes fundamental assumptions about the relatedness of the populations and the ways they generate data. For example, the comparison of Congresspeople to their district ideologies by Miller and Stokes (1963) raised exactly this issue, and generated vibrant debate about the ability of such a comparison to create numerically precise estimates of similarity, as opposed to general correlations between the two (Achen 1977). Coding for latent traits like ideology, or preferences on regulation and other issues, through content analysis is an issue of the same species.

Developing further the theory of latent traits on which both content analysis and joint scaling are presupposed allows for the derivation of a test statistic for the hypothesis that the populations are comparable—and, therefore, that any model estimated on one subsample of it is reasonable to use for prediction on another subsample. This solves the "strong, but untestable assumptions" problem. It also solves the problem of absolute model quality. While statistics like the Akaike Information Criterion (AIC) and penalized log likelihood can help the researcher to determine if a model fits the pooled corpora better than the corpora separately, the delta-statistic relies on a strong theory of latent traits to evaluate the absolute quality of a model, in addition to the rank ordering of fits. This is especially important when supervised methods are employed but a subset of the data are unlabeled, making it impossible to evaluate ground truth fit.

The second and third research questions apply to the nationalization of the American political system. One might wonder what on earth the study of comparable text sources has to do with nationalization. In fact, the study of nationalization in this dissertation *is made possible* by the use of text, which I use to code for political agendas at different levels of government. For instance, I compare the agenda speeches of Governors to the agenda speeches of Presidents, or the agenda speeches of Governors to statutes passed in their states. These sources of text are naturally distinct and therefore subject to the assumptions of which I have been quite critical. It is necessary to test for these assumptions in the course of comparing these sources of text, and as such I test the assumptions using the delta-statistic. The research then – having its assumptions tested – proceeds, with business as usual.

The second paper, "Have State Policy Agendas Become More Nationalized?" explores whether the political agenda correlates with the nationalization of political engagement and electoral outcomes. It takes as its unit of analysis the State over time. It uses a human-validated LDA model to code State agendas spanning the period 1960 to 2016, from a novel set of gubernatorial State of the State addresses, gathered over the last eight years as part of this dissertation nationally from state libraries. It uses that same model to code the National agenda using Presidential State of the Union speeches. It then compares for each unit the similarity between the State agenda and the other concurrent State agendas, or the State agenda and the concurrent National agenda using the overlap in the codes generated by the model.

The analysis shows that State agendas have become more similar to each other over time. This result suggests that, over time, States have begun to consider the same political issues. When I slice the data by party and region to look for heterogeneous descriptive effects, I find that while there generally has been a trend in the increase in similarity between agendas over time, the Southern Democrats are largely responsible for the great increase in average similarity in the 1990s.

The analysis also shows that State agendas are more similar to the national agenda (as laid out in the State of the Union addresses), and that the similarity between the state and national agendas predicts the nationalization of gubernatorial elections. Using a theory of representation, I explain how the results seriously reduce as a possibility that voters *do not engage* with some form of issue voting, in contrast to hypotheses of pure partisan teaming. This result is implied logically for individual vote choice, even though the analysis is conducted at the state level. Further, it

suggests that the nationalization of gubernatorial elections may represent a rational response to the choices that voters face. This claim is tempered, of course, by concerns of endogeneity, which I take address using an instrumental approach.

The third paper, "Can States Govern Effectively When Politics Are Nationalized?" studies the consequences of nationalization for lawmaking in the States. In particular, it considers the question of whether electoral nationalization moderates the relationship between divided government and legislative productivity in the states. It takes as its unit of analysis the State over time. Conventional wisdom holds that divided control of the government hinders the ability of our elected representatives to govern, but research on whether this is the case has been a mixed bag. In this paper, I introduce new evidence by testing whether divided government affected lawmaking in the States from 1960–2012. I then test whether nationalization of electoral behavior is a part of this problem, arguing that nationalization is related to the ability to govern because of its close relationship with polarization and the incentives of elected representatives.

I use gubernatorial agenda speeches to identify salient state political issues within each year, and then with the resulting issue salience codes, I use automated content analysis to identify salient laws passed in the states. I apply these codes to study whether nationalization and divided government affected lawmaking in the States from 1960–2012, using, defining the ability to govern as the percentage of laws passed in that year that were salient. I find a null effect of divided government on lawmaking ability. The approach shows evidence consistent with the "Mayhewvian" null findings on divided government, while adopting a similar, salience-based measurement approach of studies that do find an effect.

I also find that while nationalization is *not* related to the ability of our state governments to take action on salient issues during times of divided government, nationalization of state legislatures has generally *decreased* the production of salient laws. This finding is a somewhat troubling. It suggests that our nationalized political environment has influenced the ability of state lawmakers to govern effectively, but not through the institutional arrangements we usually consider to be the problem. This finding is a somewhat troubling. It suggests that our nationalized political environment has influenced the ability of state lawmakers to govern effectively, but not through the institutional arrangements we usually consider to be the problem. In fact, the findings suggest that behavioral factors driving lawmaker decisions may be more to blame for lawmaking defects than institutional ones.

The remainder of this introduction proceeds as follows. First, it introduces further the problematic comparison of political texts—or, to put it another way, the problematic pooling of two separate units in the same analysis. Then, it outlines the approach I take to solve these problem.

Problems in the Comparison of Political Texts

Encapsulated in the written word are the world's political ideas, records of transaction, and markers which indicate the behavior, anticipated or realized, of individuals. The *Federalist Papers*, penned by Hamilton, Madison, and Jay, offer insight into the collective action problems faced by the founders, and how they mobilized the masses to solve them (they also provide bountiful reference material to the justices of

the U.S. Supreme Court).[1] The *geniza* papers, medieval records of correspondence preserved for a millennium by Jewish law, describe the development of the silk road and the precedents of international trade, feeding plentiful research on the ancient origins of economics and commerce. Public statements of opinion on social media networks like *Facebook* and *Twitter* help marketers to persuade voters and personalize services, while helping governmental organizations prevent acts of terror.[2]

Questions that require the analysis of content motivate much of the research in political science, economics, and the social sciences, but such research is currently limited by subjective and assumptive methods of conjecture. Political scientist often examine text bearing on state policy agenda, legislative agendas, and presidential agendas, for example (the text sources I rely upon in this dissertation). It is only through the subjective analysis of content that we may tap these data, which exists primarily in *unstructured* form—more commonly known as the written word, the spoken language, or the visible reality captured in a picture. These data must be interpreted, or *structured*, by the researcher before any systematic analysis may begin. The interpretive process breathes meaning into what might otherwise be considered exhaust, but it is problematic because the process, and the means by which it is reviewed, are quite subjective.

The first paper of this dissertation helps researchers overcome the pitfalls of

[1]The first widely cited statistical analysis of text for the purpose of historical inference used automated content analysis to determine that Hamilton was the author of the previously unattributed Federalist Papers (Mosteller and Wallace 1964).

[2]Research questions such as the ones this dissertation addresses require analysis of the written word, or the analysis of text. Newspapers, political agendas, and legislation – the sources of text one might use to answer such questions – are full of data to analyze.

subjectivity by developing a theory and structure for the analysis of content. Generally, the paper is about making valid inferences from unstructured text data when a meaningful analysis depends on the comparison of one text data source to another. Particularly, this project concerns analyses that pool naturally distinct sources of text – distinct *corpora* – in an effort to draw substantive empirical conclusions.

The Pooling of Political Texts

While it is clear to most researchers that they should adjust in some way for pooled comparisons of corpora, it is unclear to most researchers what precisely to do. The modal response is to ignore major differences between the corpora, treating the documents they contain as perfectly comparable units of observation. The trouble is that this makes classical statistical inference problematic, because it omits corpus-associated parameters from any posited statistical model. In fact, pooling disparate observations is generally one of the least understood aspects of model specification.

The problem of pooling disparate observations is not novel. Social scientists make choices every day concerning which time periods (*e.g.* presidential administrations) to include in an analysis; whether different surveys of opinion should be included in the same analysis; whether householded survey responses may be compared to individual-level respondents (or a particular locale's), and; whether votes from the Senate, House, and the President (via policy agendas) may be used to jointly scale the ideology of political elites. In fact, these educated decisions to pool observations directly inform the fundamental process of induction: what entitles us to make inferences about the

behavior of a particular unit on the basis of observed behavior from some different unit? One would not train a model on elections in France and apply it to elections in Turkey, but the analogous case is often applied without second thought in text analyses. Such considerations are perhaps obfuscated further in text analysis because the comparison of text to text does not seem to involve disparate data at all. This dissertation provides reasoning and proposed models to account for just how disparate text data are.

Research papers on the subject of joint scaling have encountered this problem in the past, from two slightly different angle. The first angle is that of the substantive interpretation of jointly scaled analyses. Scholarly critiques of Miller and Stokes (1963), such as Achen (1977) and Erikson (1978), for example, emphasize the inability of pooled analyses to recover absolute distance in the study of representation, rendering some results inert.[3] The joint scaling methodology used does not allow for the direct comparison of pooled observation—merely their orders are comparable. The second angle is that of the ability of joint scaling methodologies to recover estimates of the latent dimension, or dimensions, the researcher wishes to capture. Jessee (2016, page 1122, italics mine) studies the ability of several popular datasets to recover useful jointly scaled estimates and finds incredible variability. He concludes in the discussion that:

> "[Jointly scaled] Ideal point models are a valuable tool for political scientists. But these models are not magic. There is no guarantee that an ideal point

[3]See also, however, Page and Shapiro (1983), Lax and Phillips (2012), and Clinton (2006), who propose methodologies to overcome some of the issues in Miller and Stokes (1963) (which include, among other things, problems relating to a probability sample of constituents which might be biased).

> model will find the dimension of interest to researchers when fed a set of data. *The burden is on researchers to provide clear thought about both the indicators chosen and what underlying dimension is relevant for a given application.* The selection of a measurement approach is as much a substantive or theoretical issue as a statistical one."

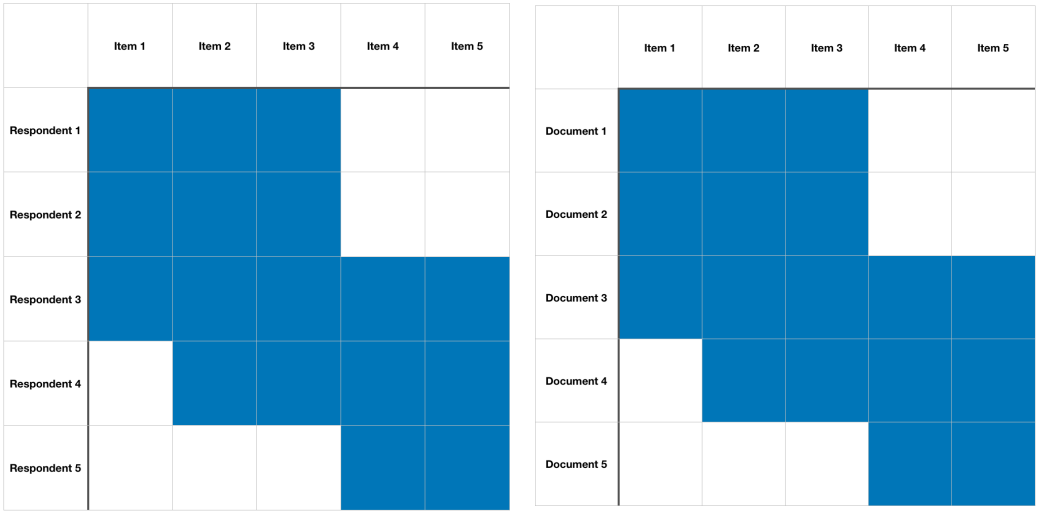
More generally, even Poole and Rosenthal (1997) exhibit great trepidation at the idea of comparing roll call votes across chambers of Congress.

The joint scaling of disparate observations within the political domain is indeed an outstanding and important issue, and one for which no quantitative solution has yet been offered.[4] It is surprising to think that researchers have progressed to the comparison of such disparate observations as casual everyday Twitter speech and corporate filings to the Securities and Exchanges Commission (Loughran and McDonald 2019, *e.g.* and references therein) when comparisons of even patently similar databases – such as the ones used in joint scaling – do not hold up. If scientists have yet to remedy the issue of joint scaling in political science, how can we have confidence in even more distant "joint scaling" efforts, using the same methods? One need only consider the ever increasingly disparate nature of the datasets being pooled together today to realize this fact (disparate datasets are being "vinculated," to use the word of Monroe 2013, to an increasing degree).

The scaling of sparse observed outcomes such as roll call votes is very similar to the scaling of sparse observed language. In fact, one might very simply explain sparsely structured text data by way of analogy: you may think of text as if it were a survey, where each document is a respondent, each column is a survey item, and the

[4]Bartels (1991) offers one solution from a Bayesian perspective, which requires the researcher to attach his or her belief in the comparability of the pooled observations to the dataset, so that the researcher's belief may be used to weight contributions to any model's estimated parameters.

Figure I1: Observed Data from Surveys and Roll Calls are Commonly Represented Similarly to Structured Data from Text



(a) Matrix from Surveys (or Roll Calls) **(b)** Matrix from Text

Note: Panels are representations of spreadsheets. Rows are observations, and columns are dimensions. In panel (a), each row is a survey respondent, and each column is an item on a survey. Alternatively, in panel (a), each row may be a Member of Congress, and each column may be a vote on a bill. In panel (b), each row is a document, and each column is a word, also known as a "text feature." Shaded cells indicate when we observe a response to the survey item or observe the presence of a word. Alternatively, shaded cells in (a) may indicate if we observe a *yay* on the bill. The figure demonstrates how the structure of the data used in both scaling (joint scaling) analyses and text analyses is similar.

cell values are the document's response to the survey on that item. In other words, we represent the choice to include a word in a document in the same way we would represent the choice to respond to a survey—the columns compose the *choice space* for the global context, and the values reflect observations of choices within that context. Figure I1 provides an exhibit of this analogy by way of visual aid. In section 1.1, I return to this analogy and consider further some of the benefits and (very real) drawbacks to thinking of text data in this way.

As in the problem of joint scaling, the core idea behind this dissertation is that text is not perfectly comparable simply because it is text. Rather, text is comparable

when the units of observation are also comparable. The core innovation herein is that we may check if units of observation are comparable by naturally linking the theory of content analysis to a practice in machine learning called topic modeling, which clusters patterns in text into groups called "topics." These machine-generated topics are, incredibly, very intuitive to interpret as if they were naturally occurring topics in everyday speech. This explores what is possible if we make the principal assumption that in political speech, machine-generated patterns in language are related to real-world patterns in political traits. The assumption that topics relate to traits is commonly made, if infrequently advertised, as discussed in appendix B.4. This assumption suggests that estimated topics may be substituted for those variables, given that the design allows for a decrease in efficiency. Although the existence of latent traits is indeed debatable (see the discussion of index measurement systems in appendix B.4), this dissertation faithfully takes the assumption that they are measurable to its natural conclusion.

More importantly, however, this assumption also suggests that we may substitute topics, based on observed text, for *unobserved variables in cases where data are missing.* This makes balance checking among units of observation possible when we lack observed data other than text.[5] The statistical framework for testing if units of observation are comparable is, then, practically mechanical. When units of observation are balanced on traits, we consider them to be comparable. By testing the balance

[5]It is fortuitous, then, that the practice of applying textual regression models estimated on one corpus of documents to another corpus of documents becomes verifiable by way of this method; the usual problem is that we have no ground truth data, such as ideology, to work with when we estimate ideology—this provides a way to test against observed data.

across those traits, we may test if the units are comparable.

When Content Analysis Goes Wrong

To introduce the need for a method that evaluates if two corpora may be compared, consider the case of an analysis published by Gentzkow and Shapiro (2010). The analysis addresses a question of longstanding and intense interest in the study of media economics: do newspapers introduce ideological "slant" into their content? If so, why, and to what end? To answer this question, the researchers generate a codebook which is meant to estimate ("label") the political ideology of a newspaper, based on what that newspaper writes. They select key political phrases using a statistical technique that identifies phrases used by Republicans and Democrats in Congress. Then, they weight those phrases based on the vote returns of the members who used them, such that highly weighted phrases contribute increasingly to the estimated ideology of the newspaper. The estimator asks, *if a newspaper where a congressperson, what would their ideology be?* They use a codebook generated on a corpus of labeled Members to label a fresh, unlabeled corpus of newspapers.[6]

The paper makes an easy-to-miss, key assumption: that the units of observation are comparable. There are three main components that go into this assumption. First, a single, unitary trait we label *political ideology* exists among congresspeople

[6]One reason this question is important is because ideologically slanted newspapers may encourage electoral polarization. The readers of newspapers with ideological slant may consume biased information under the premise that it is balanced and true, forming strong opinions based on tenuous ground. Further, slant consumers may knowingly choose to engage with biased publications, leading to further social segregation and clustered homogeneity (Putnam 2000; Dunkelman 2014).

and newspapers, and the trait is of the same nature and origin in both populations. Second, there is a codebook which may identify, or measure, from observable behavior ideology within the generating corpora (note that this implies a link between the trait and observed behavior). And third, that this codebook detects in the same way those labels, in both corpora. These components may be called the *bridge criteria*, and they are defined more generally in section 1.3.

The paper problematically does not test this assumption. Moreover, numerous other papers others like it, which employ the same methodology, have yet to critically evaluate their assumptions in a public and replicable research forum. If one *does* test the assumption, the paper's methodology – and therefore the conclusions drawn from the analysis – fall apart.

Testing the Bridge Criteria

This dissertation argues that the bridge criteria are essential to successful text analyses. One way to empirically evaluate the bridge criteria is with cross-validation. Cross-validation is a method which evaluates the predictive power of a model by training it on a subset of the data, and then testing it on another, previously "unobserved" subset of the data, sometimes called a hold-out set. If the model is the "true" model for the process that generated the data, then the parameter estimates and predicted quantities produced by each training iteration will replicate in other subsets of the data. This is the approach taken by Jessee (2016), who jointly scales the ideologies of voters and senators and validates with cross-validation; the paper finds significant

variability in subsequent estimates of ideology conditional on each model, suggesting the bridge criteria are not satisfied.

As a simple example, we could cross-validate the model from Gentzkow and Shapiro (2010) by training one model on the legislator data, another model on the media outlet data, and then comparing estimated ideologies from the criss-crossed models to the true ideologies. This is *impossible*, however, because we do not (and cannot) observe the ideology of the news outlets; indeed, it is the reason for which we took the codebook generation approach in the first place. This is one problematic element of evaluating the bridge criteria, and an element to which we will return later. The method this dissertation introduces tests the bridge criteria without the need for observed values.

Alternatively, we can explore the sufficiency of the bridge criteria by treating the estimated ideology values as observed values and re-estimating the model on the media outlet data.[7] If the data generating processes are the same, then their codebooks (regression weights) should also be the same. We may generalize this process by assigning weights to the data in the model that predicts ideology for both legislators and media outlets. If the assumption holds, then the model's parameter estimates (for partisan phrases) should not systematically vary as a function of the data from which they are estimated. In other words, the model's results, fit, and βs should not change if we assign more weight to the legislator or media outlet data. We can do this by assigning weights to the data used to train the model, just as we assign weights to

[7]The logic of the approach is similar to that invoked in the estimation of standard errors in ordinary least squares: if the estimator approaches the "true" model, then residuals may be used to estimate population variability.

respondents in the analysis of survey data. This is the approach followed in Lewis and Tausanovitch (2015), which is also interested in the comparability of distinct observations (though they do not explicitly state this fact).

To do so, construct an n by k matrix of counts W, for n legislators and news outlets and k partisan phrases selected using the χ^2-ranking procedure (Gentzkow and Shapiro 2010). Construct also the vector of estimated ideologies from the paper, $\hat{Y} = (y_1, \ldots, y_n)'$. The weighted least squares model may be specified as:

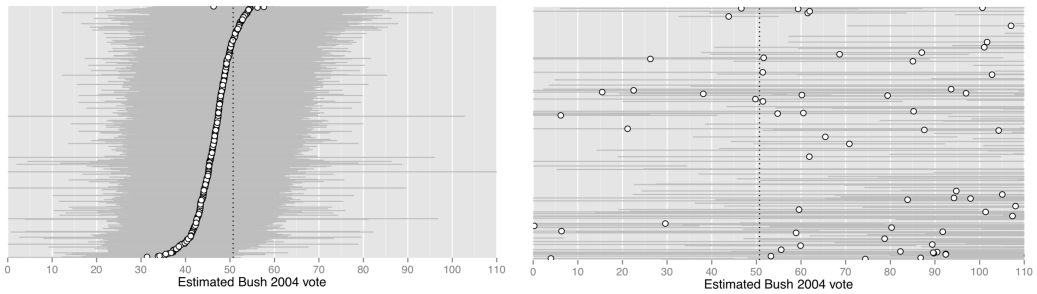
$$\hat{\theta} = \alpha + W\beta + e, \text{ where} \tag{1}$$

$$\hat{\beta} = (W'\Omega W)^{-1} C'\Omega \hat{Y}, \tag{2}$$

and Ω is a vector of n weights. At one extreme, we can use Ω to down-weight media outlet data such that media outlet estimates are conditional entirely on information from legislator data (this case is equivalent to the (Gentzkow and Shapiro 2010) case). At the other extreme, we can down-weight legislator data such that media outlet estimates are conditional entirely on information from the media outlet data. Figure I2 reports the estimated ideal points for media outlets under these two weighting schemes.

Figure I2 demonstrates that the ideal points are in no way robust to the specification of the weights. This means that the results we observe are highly conditional on the subset of data from which we derive the codebook (regression weights). This violates one of the bridge criteria. If the codebooks were the same for both subsets of the data, we would see concordance in the results. It further suggests that the data generating process that maps the population of newspaper personnel, to the text we observe

Figure I2: Media Ideal Points from Gentzkow and Shapiro (2010)



(a) Weight to Legislators **(b)** Weight to Media Outlets

Note: Points are estimated ideologies for 60 U.S. media outlets, in increasing order. Ideology is scaled to Bush Vote in 2004, per Gentzkow and Shapiro (2010). Gray lines are 80% confidence intervals. Panel (a) assigns higher weights to legislator data, while panel (b) assigns higher weights to media outlet data. Order is fixed to the ordering generated by the original model (based on legislator speech). The figure suggests that the model used in Gentzkow and Shapiro (2010) is overindexed to legislator speech. Replicated following Lewis and Tausanovitch (2015).

in their newspapers, is not the same as the data generating process that maps the population of Congress to the text we observe in the Congressional Record.

This analysis suggests the bridging criteria are not satisfied in the case of Gentzkow and Shapiro (2010). The model the authors propose is therefore inappropriate because it is over-fit to one subset of the data. The quantities it is used to estimate, including ideology, are inappropriate for use in their subsequent analysis of media bias. These results cut against the hypothesis that the codebook generated to do the measurement may be applied similarly to both corpora—that newspaper speech may be compared to Congressional speech, at least conditional on the model of language the authors posit.

My hope is that this example demonstrates to the reader how the comparison of one data source to another on unsupported grounds may generate poor results. The evidence the authors report suggests that newspapers are indeed ideologically slanted,

and that ideological slant is related to the ideology of the newspapers' markets. The broad implications of their result – biased news, increasing polarization, the persuasive transformation of the American community – are groundbreaking. The assumption of comparability upon which their analysis rests, however, is untested and problematic.

How This Dissertation Solves The Problem

The thesis of this dissertation's first paper is that the bridge criteria exist, that they are of consequence, and that it is possible to test for them with a method built on top of topic modeling. The explicit bridging criteria are a novel contribution of this dissertation, but the idea of the criteria is not new, as shown in equation (1.2), section 1.3.1. The dissertation provides theoretical consideration of why, how, and when we should compare texts. It proposes a methodology for such comparison. Then, it employs the methodology to conduct empirical research on questions of substantive interest in nationalization.

In section 1.1, I review existing research on the statistical analysis of political text. I then consider the problem of comparing corpora as a subset of the more general problem of pooling disparate observations. I review how we might test if observations may be pooled. Then, I develop a background on why we should be able to test for comparability using the link between measurable latent traits and observable text, through the lenses of content analysis and joint scaling. The chapter introduces some of the primary issues entailed in violating the bridge criteria, and extends content analysis theory to provide a solutions for how to overcome them. Section 1.5 develops

the mathematics which may be used to conduct hypothesis testing consistent with the theory. I label the statistic generated the *delta-statistic*, because it concerns the observed distance between two or more corpora, and what we would expect the distance to be if they were generated by comparable processes. The chapter empirically validates the mathematics with Monte Carlo simulation, and with real-world data from the Gentzkow and Shapiro (2010) dataset.

Having proposed a solution to the issue, the dissertation in chapters 2 and 3 switches gears. They conduct empirical research in a way that requires the comparison of naturally distinct corpora—the circumstance for which the method was originally pursued. The research uses the method essentially as a logical gate: if the method reveals that the corpora may be compared, then we proceed with the research.

Chapters 2 and 3 conduct empirical research on a novel dataset, the State of the State corpus, which I have led the collection of over the past seven years in collaboration with researchers at Washington University in St. Louis, Columbia University, and the University of California at San Diego. Chapter 2 studies nationalization by comparing the State of the Union and State of the State addresses. Chapter 3 studies divided government and the role of nationalization in it by comparing State of the State addresses with the text of passes state legislation.

Finally, the Conclusion reflects on findings from the method and research, draws conclusions, and suggests paths forward for future research. From a methodological perspective, the delta statistic is a useful tool for researchers who have text available to them and wish to evaluate the balance of two or more data subsets on the basis of that text. The method is also important because it highlights the risk of conducting

research without addressing fundamental design concerns.

Chapter 1

A Theory and Method for Pooling Naturally Distinct Corpora

This chapter proposes a solution to a problem that all content analyses must deal with, but which a significant subset of published research applying automated content analysis suffers from: the problem of pooling disparate texts in an analysis. The methodological question the dissertation addresses is, "How can we determine if two corpora of documents may be pooled together in an analysis?" Simply put, the problem is that a great deal of research papers assuming they are comparing apples to apples are actually comparing apples to oranges. The findings generated by the comparisons skip past the prerequisite step, commonly applied in the process of qualitative research and explicitly called out in the joint scaling literature, which establishes that the set of rules a researcher produces to analyze one situation, or categorize one set of documents, is valid to use for the analysis (or categorization) of another.

Consider for the purposes of demonstration the following example. A researcher wishes to test if ideological media bias is a function of the ideology of their customers, which might be expected if newspapers wish to produce content that will be favorably received by their customers. In the ideal case, the researcher would regress a vector

of newspaper ideologies on the ideologies of their customers (perhaps with some design that allows for causal identification). The main challenge the researcher must overcome, of course, is estimating ideological ideal points for newspapers, which are not readily available. Call these desired estimates for ideology $\hat{Y}$, where Y is ideology and the hat indicates that values are estimated. Because there are no surveys of newpaper editorial boards readily available, the researcher elects to employ a technique that regresses NOMINATE scores (which estimate the ideology of congresspeople) on congressional record speech, and then uses the resulting model to estimate ideology based on each newspaper's text. In doing so, the researcher *pools* and *jointly scales* the data. I will use this use case as a running example.

In this chapter, I introduce the theory and empirics for the delta-statistic. The delta-statistic is useful because it may be used to test whether two naturally distinct populations, and their generated corpora, may be pooled together. The statistic is important because it prevents unsupported inferences; researchers have in the past drawn unsupported conclusions based on pooled analyses of populations and corpora (see the introductory essay for detail). Using the running example, the delta statistic tests if the regression model used to produce $\hat{Y}$ is valid for that purpose. This test allows researchers to defend against the critique that their results may be wrought by an inappropriate application of the scaling method they use.

1.1 The Systematic Categorization of Texts

The practice of content analysis is perhaps as old as the first texts and readers, which together engaged in a process of expression and interpretation to produce new knowledge and understanding. The interpretation of religious texts and the attribution of their authorship, for example, required the ordained readers to consume the information written in manuscripts and produce inferences from the nature of their content (Krippendorff 2012). Douglas, Berelson, and Bradshaw (1940, p. 2) were the first scholars, to knowledge, to describe the process of disciplined analysis of images, recordings, and the written word as "content analysis."[1] Berelson and De Grazia (1947), in a Public Opinion Quarterly piece aimed at detecting collaboration in propaganda, developed the definition of the approach further. Berelson (1952, p. 18), in his work *Content Analysis in Communication Research*, finally proposed what is today perhaps the canonical definition of the practice: "a research technique for the objective, systematic and quantitative description of the manifest content of communication." Webster's finally included in its 1962 edition, the definition "analysis of the manifest and latent content of a body of communicated material (as a book or film) through classification, tabulation, and evaluation of its key symbols and themes in order to ascertain its meaning and probable effect."

Political communication scholars may consider through content analysis, and the content analysis of text, the issues of agenda setting, framing, priming, and persuasion. Agenda setting (McCombs and Shaw 1972) is the study of what topical and policy

[1]The citation was found in Berelson and Lazarsfeld (1948).

issues the media prioritizes in the course of its discussions. Content analysis in support of agenda setting research must recover codes for *what* is said in the course of a news article, political speech, or other text artifact. Topic modeling (Blei, Ng, and Jordan 2003) is an appropriate method with which to automatically code for the topics and issues of discussion because it assigns to any document the proportion of that document which discusses a certain topic; the methodology, including is widely used to do so (Grimmer and Stewart 2013; Martin and McCrain 2019; Roberts et al. 2014). This dissertation is concerned with the agenda setting thread of political communications research. I discuss this further in section 1.5.

The topics of framing, priming, and persuasion, on the other hand, concern not *what* is said, but *how* it is said, and how that presentation affects opinion or affect. This dissertation does not attempt to suss out the framing, priming, or persuasive effects in political texts. Instead, it is specifically interested in whether a topic is mentioned by a political actor or not. It is the simple logic of whether the topic is mentioned, and the types of words the actor uses to discuss that topic, that is of interest in the method I introduce.[2]

To scale the process of content analysis, social scientists have developed methodolo-

[2]One of the most problematic elements of content analysis is its subjectivity. Content analysis may be one thing to one form of research, and yet another thing to another form of research. Though this dissertation considers only the scope of "whether a topic is discussed" and nothing else, in my view content analysis is at once the combination of five communicative elements: first, the underlying truth – a latent trait – which prompted the generation of the text; second, the writer's ability and willingness to represent that truth; third, the world's imposition of noise which interrupts the transmission of that truth to another reader; fourth, the veil, perhaps "rose-colored glasses," which the reader applies to the content, and; fifth, the reader's ultimate communication of their synthesis to others, which starts the process all over again. I later discuss the necessity of a bounding method to determine for the researcher how much of her results owe themselves to the selection of the lens through which she analyzed the content.

gies to support the validity, reliability, and portability of qualitative analyses (Lasswell 1927; Berelson 1952; Krippendorff 1980, 2012; Neuendorf 2016). In parallel, computer scientists have over the last eight decades developed computerized methods of content analysis, which leverage sets of programmed or learned rules to annotate text at a much lower cost (Mosteller and Wallace 1964; Manning, Raghavan, and Schütze 1991). These two research arcs, focused on reducing cost and error, have at times developed on parallel but separate tracks, at the expense of extensive cross-pollination. With the emergence of cheap, accessible computing power and broad interest in the application of content analysis to new "big data" sources, a cottage industry has emerged to integrate the two of work streams together, in earnest (Grimmer and Stewart 2013).[3] Automated content analysis is now the modal approach to the scientific analysis of text documents, and as use of this new technology grows so must our understanding of its implications for inference.

In political science, it has become vastly more popular for researchers to adopt the tool set of supervised machine learning, a set of techniques that like regression use one set of data W to predict another Y, positing and estimating the data generating process $Y = f(W, \epsilon)$. A researcher can use supervised learning to reduce the cost of hand-coding to nearly zero. She does this by using the codes another researcher assigned to their text data to predict the codes for her own text data.[4]

[3]This is not to say that there has been no cross-pollination; indeed, social science research has made use of computational techniques in not only natural language processing (Nacos et al. 1991), but in dimension reduction (Poole and Rosenthal 1985; Ansolabehere, Snyder, Jr., and Stewart 2001) and Bayesian inference (Gelman and Imbens 2013), for many decades. It has perhaps not been, popularized, however, until recently.

[4]Note that this approach is compatible with, but not the same as, unsupervised machine learning (techniques like clustering), by which a researcher uses variation within the dataset itself to assign

The process usually involves two steps. In the first step, the researcher applies supervised techniques to model an observed value or class – such as ideology, or the topical matter – as a function of observed text. In the second step, the researcher uses that model to produce estimates of the value or class from another source of observed text. Supervised learning also reduces exposure to the replicability critique, as the procedure for the estimation of the new codes is fully delineated. Increasingly, research projects are taking this approach. For example:

- Martin and McCrain (2019) use a text-based measure of ideological slant to argue that corporate media interests influence the content of local news. The slant measure is trained on ideology-coded congressional record speech and applied to local televised news transcripts.
- Morgan (2018) argues that politically-driven religious sermons affect political attitudes within the church community. Morgan uses a pre-coded corpus of CNN news transcripts to predict the presence of political speech in a corpus of religious sermons.
- Martin and Yurukoglu (2017) argue that conservative media outlets increased the national Republican vote share by 0.46, 3.59, and 6.34 points in 2000, 2004, and 2008. Ideal points for media outlets are estimated from a model trained on ideology-coded congressional record speech and applied to cable news transcripts.
- Hopkins and King (2010) we develop a method that gives approximately unbiased estimates of category proportions using an estimator trained on a labeled

labels to data. Unsupervised approaches may create relative labels, but they are meaningless with respect to the supervise method until the supervised method links the relative labels to estimable quantities.

document set, and apply it to several data sets, including the daily expressed opinions of thousands of people about the U.S. presidency.[5]

- Gentzkow and Shapiro (2010) argue that local media bias is a function of local demand for biased news. Their media bias measure comes from a model trained on ideology-coded congressional record speech and applied to local newspaper transcripts.[6]
- Groseclose and Milyo (2005) measure media bias by estimating ideological scores for several major media outlets. Their ideology measure comes from a model trained on counts of think tank citations in Congress and applied to Counts of think tank citations in newspapers.

Indeed, more and more researchers are employing this methodology in order to draw substantive conclusions.[7]

Unfortunately, directly testing this assumption is not usually feasible. It is infeasible because the dimension of interest (*e.g.* ideology, as in the example used earlier) is almost always missing from one of the individuals. In a perfect world, we would observe ideology for both individuals, and then test to see if the language models we train for both individuals are comparable. In reality, we observe ideology for one individual, and are left with the estimated ideology for the other. Current approaches examine the ability of a pooled model to recover these estimates under different

[5]Jerzak, King, and Strezhnev (2018) extend the method of Hopkins and King (2010) and further validate its performance.

[6]This paper and its approach are extended in Gentzkow, Shapiro, and Taddy (2016).

[7]The boom in machine learning interest has also perhaps contributed to the recent peak in the interest in this method.

weighting schemes (Lewis and Tausanovitch 2015), but there is not yet a formal test for the independence of two or more corpora.

In this essay, I propose one such test. The test compares two text corpora – generated by what may be two distinct populations of actors – to each other on the basis of their independently modeled *topics*, which are clusters of words (and phrases) that are likely to appear together. Each topic is represented as a probability distribution over words. We would not expect to see exactly these distributions over words in the topics of another corpus if the corpora were independent. As such, the test compares the topics we observe to a set of randomly simulated topics of comparable shape to compute a test statistic.

Simply put, the test checks if the patterns in speech (or writing) of two naturally distinct sets of actors are sufficiently different to prevent a pooled analysis (that would normally be pursued if they were part of the same population). This test allows researchers to defend against the critique that their results may be wrought by an inappropriate application of the scaling method they use. In fact, the methodology in this essay should enable researchers who have until now been reticent to pool distinct corpora of texts in their analysis. Moreover, the specific theory of content I develop in this essay should, I hope, provide an analytical framework for the creation of models that better represent the observed world.

The following sections develop the test statistic and then apply it to a well-known application of the joint scaling method. Building on the linkage between content analysis and joint scaling, I restate and make more appropriate to content analysis some common assumptions made in the practice of joint scaling. I label these common

assumptions the *bridge criteria*, a set of criteria which must be assumed or satisfied in the pursuit of valid inferences based on the comparison of political texts. I then propose an empirical framework for the evaluation of bridge criteria. I demonstrate why a researcher may reasonably link latent traits to estimated probability distributions without undertaking the methodical hand-coding and interrater reliability testing usually necessary, by employing a method called topic modeling. Importantly, I do not posit that this method may recover the true *representational* measurement structures of the latent processes implied; rather, I argue that the process may validly recover relative comparisons of *index* structures, which may be used to logically test the bridge criteria. The results demonstrate the ability of the approach to produce prima facie conclusions and outperform existing approaches.

1.2 Joint Scaling: The Comparison of Disparate Data on the Same Scale

Most research projects treat text as a measurement instrument for an underlying latent trait (Laver et al. 2003; Grimmer 2010; Quinn et al. 2010, e.g.).[8] This view is consistent with the second conception of content discussed in appendix B.4. For example, Laver et al. (2003) treat text as a measurement instrument for political ideology; Grimmer (2010) treats text as a measurement instrument for the salience of political issues. Most research in political science is interested in the underlying processes that generated the

[8]See also Benoit, Munger, and Spirling (2019), which considers the readability of text, usually treated as a property of the text itself, to map to underlying variables.

text we observe. For example, studies that use political platforms to estimate ideology implicitly assume that true, latent ideology is meaningfully linked to the text produced in platforms. In other words, most text analyses published today treat documents as if they are *measurement instruments* for the underlying, latent phenomena they wish to represent.

Ideal points also ascribe to this conception of measurment. Their original use was to evaluate the intellectual "ability" of students who have answered questions, or "items," in tests (Lord and Novick 1968, *e.g.*). Political scientists co-opted this approach to use observed political choices (akin to these "test items") to estimate the left-right locations of legislators or voters. Poole and Rosenthal (1997) applied the technique to congressional data to yield NOMINATE scores, which fundamentally changed the way congressional research is conducted.

An ideal point represents an individual as a "point" in geometric space. This point is the projection of an individual, who may have unlimited preference or characteristic dimensions, into a lower-dimensional numeric space. Individuals may have formed several thousand opinions their likes and dislikes over their lifetimes. However, it is hard to understand their preferences in any sort of meaningful way when they are so highly dimension. Ideal point estimation allows us to take those many opinions and distill them down to one or more general dimensions, which are more informative and allow for easy comparisons between individuals. In our running example, we may consider $\hat{Y}$ to be estimated ideal points that intend to represent Y.

Soon after the introduction of NOMINATE (Poole and Rosenthal 1997), which used joint scaling methodologies to produce interpretable ideological scores for legislators

within a particular chamber of each Congress, scholars began to employ ideal point estimation to a broader swath of datasets. For example, Londregan (2000) incorporated agenda information into the estimation of ideal points. In recent years, researchers have framed the estimation problem in terms of Bayesian statistical methods to allow for better handling of missing data and uncertainty (Bafumi et al. 2005; Clinton, Jackman, and Rivers 2004; Martin and Quinn 2002). These Bayesian ideal point estimation methods have been used in a variety of contexts, such as the supreme court (Martin and Quinn 2002, 2007), regulatory agencies (Clinton and Lewis 2008), the federal government (Bailey and Chang 2001; Clinton et al. 2012). Some research have even produced ideal point estimates that allow for comparison of latent dimensional estimates across different periods of time, such as Ansolabehere, Snyder, Jr., and Stewart (2001) which applies a technique introduced in Heckman and Snyder Jr (1996), along with an adjustment suggested by Groseclose, Levitt, and Snyder (1999).[9] This practice, in which different time periods, or somewhat disparate observations are pooled in order to accomplish the comparison of their scaled quantities, is known as joint scaling.

A key limitation of ideal point estimates is that they are only given meaning relatively. That is to say, the interpretation of one ideal point is only meaningful insofar as it is compared to other ideal points. This is problematic when ideal points are compared across time regimes or across contexts. For example, ideal points for members of the Senate are not directly comparable to ideal points for members of the

[9]Some approaches, such as Martin and Quinn (2002), build dynamic estimation directly into the model.

House if they are estimated separately; an ideal point for an MC in one session is not comparable to an ideal point for that same MC in another session if they are estimated separately. If compared directly, the analysis may suggest differences in ideology that do not in reality exist, simply because the ideal points are not comparable. This is frequently referred to as the problem of "common policy spaces" (Bafumi and Herron 2010).

1.2.1 The Importance of Bridging Assumptions

Joint scaling is a method which places individuals in a common space, such that their ideal points may be compared. To place individuals in a common policy space, authors make "bridging assumptions" (Bailey 2007, *e.g.*). These assumptions treat responses to certain policy items as if they were generated under the same circumstances, thereby generating ideal points in the same space. This allows for the direct comparison of the individuals on the basis of their ideal points.

These "bridging assumptions" are instances of the more general assumptions made in any comparison of disparate data. Beck (1985, page 79) suggests an obvious example, that in research using the American National Elections Studies (ANES), "survey researchers usually analyze the entire sample," but researchers comparing time series "always face a choice." A specific example of this type of survey data is non-panel evaluations of presidential performance across different presidencies, drawn from different time periods. Another example of this type of pooled survey data that is not the ANES is the comparison of stated preferences towards bigots and racists

from several countries over time. A third example used by Jessee (2016) is the pooling of individual citizens responses to ideology scales and the responses of political elites into the same database.

Even though the questions asked of these respondents in these surveys may not vary in the slightest, the populations and time periods of these exemplars vary significantly. And that creates problems in classical inference, as I discuss in the next few paragraphs. Text analysis is no different. Text looks the same, no matter who it is from, and regardless of what the question was asked—to the naive researcher, it is ripe and ready for comparison. However, we must consider the specific circumstances under which, and by whom or what, the text was generated, to overcome problems of inference.

Generally, social scientists make choices every day concerning which time periods (*e.g.* presidential administrations) to include in an analysis; whether different surveys of opinion should be included in the same analysis; whether householded survey responses may be compared to individual-level respondents (or a particular locale's), and; whether votes from the Senate, House, and the President (via policy agendas) may be used to jointly scale the ideology of political elites. In fact, these educated decisions to pool observations directly inform the fundamental process of induction: what entitles us to make inferences about the behavior of a particular unit on the basis of observed behavior from some different unit? One would not train a model on elections in France and apply to elections in Turkey. Such considerations are perhaps obfuscated further in text analysis because the comparison of text to text does not seem to involve disparate data at all.

As Bartels (1991) points out, perhaps one of the only ways a researcher may be

able to account for the sensitivity of her results to the assumption that the data may be pooled, is to also state her prior belief in the "poolability" of the data points at the outset. Indeed, the issue of poolability in models where parameters that capture political effects over time has been considered from several angles. Switching regime models (*e.g.* Hamilton 2010, and references therein) and stochastic parameter regression models (Beck 1983) are approaches which allow different slices of data (different, but compared, populations) to have heterogeneous effects, while still recovering a pooled estimate for the quantity of interest. These assumptions are infrequently tested, as Jessee (2016), Shor, McCarty, and Berry (2011), and Lewis and Tausanovitch (2015) point out. Classical statistical inference becomes problematic when population, or observation-group parameters are omitted from any posited statistical model (and the observation groups are disparate). Decisions to pool observations directly inform the fundamental process of induction: what entitles us to make inferences about the behavior of a particular unit on the basis of observed behavior from some different unit?

The problem of bridging disparate data sources in order to compare them along the same dimension is not new to political science. Miller and Stokes (1963), in their foundational study of representation, prompted decades of scholarly research propelled by the need to compare data on public attitudes with data on political outcomes—two very different data sources, each characterized by a unique choice space and measurement function. The method of joint scaling furnishes simple preference scales to reduce the dimensionality of the choice space while maintaining a generality that may be applied to different data sources, thereby reducing measurement error

and bias. Scholars have used joint scaling methods to study a wide array of topics, and those studies have developed into some of the most influential papers of the past few decades.

Most political scientists are aware of the work of Miller and Stokes (1963), a seminal study linking the voting behavior of political elites to the surveyed preferences of the masses (therefore, *jointly scaling* elite behavior and mass preference). The question they address is, "do the people we elect respond to the preferences of the people who elect them?" To do so, they collect preference data on constituents using survey data and compare it to the votes of legislators to Congress. Their key assumption is that the political positions of legislators, as measured by votes, and the political positions of constituents, as measured by survey responses, or comparable. The work reports a positive correlation between the preferences of the masses and the voting behavior of elites for the policy domain of civil rights, but finds much lower correlations on the issues of social welfare and foreign affairs; the authors suggest that representative democracy is perhaps the case for civil rights, but not for other issues areas.

Miller and Stokes (1963) are often credited with performing the first large-scale, empirical study of ideological representation. Although the approach was innovative, it was subject to several methodological criticisms. Achen (1977) criticizes Miller and Stokes (1963), pointing out that correlations between policy outcomes and preferences do not allow us to make claims about representation. The correlation coefficient only reveals directional relationships, and for that reason, we may observe a high correlation between preferences and outcomes even for cases in which outcomes are in fact very

far away from preferences.[10] In other words, Achen pointed out that just because a legislator might take a more conservative position because her constituents were more conservative does not mean that the legislator and her constituents shared the same position. This would suggest that the results of the study were correlational. He proposed an alternative measure, which he called "centrism." Centrism is operationalized as the squared distance between the legislator's position in the mean position of her constituents. The measure is meant to represent the extent to which legislators take positions that are close to the center of the distribution of their constituents positions.

The problem, however, is that measurement of centrism is usually not feasible. It's measurement is not feasible because the "survey items" used to elicit responses by both parties not comparable. Converse (1964), for example, suggested that individual preferences are much more prone to error than those of political elites, and that they may have a different structure altogether. Ansolabehere, Rodden, and Snyder (2008) also demonstrate the aggregated scales over multiple survey items will reliably estimate voter ideology than those based on one measure. Further, the choice and preference spaces of the parties compared may differ, such that the response one party would give under their present condition would not be the same that they would give under the other parties condition, ceteris paribus. In effect, the measurement of centrism is a lofty but difficult to implement ideal.

[10]For example, consider several districts, all which have a majority of constituents who support gay rights, and several representatives, some of whom vote in support of gay rights and some of whom do not. If the representatives who vote for gay rights also happen to represent the districts with the highest rates of support for gay rights, the correlation between preferences and outcomes would be very high. The absolute level of representation, however, would be very poor—there would still be several representatives who voted against gay rights, even though a majority of their constituents desired votes *for* gay rights.

The methodology of joint scaling, for which the "modern" canonical implementation is perhaps Bafumi and Herron (2010), offers a solution to the problem of proximal comparison between legislators and constituents. The paper links the responses of legislators and constituents by asking survey respondents to take policy positions on bills legislators had voted on. The paper assumes that the responses of the constituents have the same functional form as the responses of the legislators, such that they may be treated the same as roll call votes. Under this assumption, the distributions of preferences may be treated as if they were drawn from the same distribution, and therefore compared on non-correlational grounds. This method provides one solution that answers Achen's (1977) criticism.

The method of joint scaling has seen great popularity in the political science literature. In fact, in legislative politics, it has been used to compare the positions of legislators in the house to the positions of legislators in the Senate, or the positions of legislators at one time to the positions of legislators in another. Jessee (2009) and Jessee (2016) tests if voters have spatial preferences using joint scaling. Bailey (2007) compares the positions of judges to the positions of elected officials. Tausanovitch and Warshaw (2013) compares the preferences of respondents across different public opinion surveys. Bonica (2013) compares the political positions slaters to the political positions of donors. Shor and McCarty (2011) use joint scaling to compare the positions of state legislators to members of Congress. Groseclose and Milyo (2005) compare legislators to members of the media.

1.3 The Bridge Criteria for the Comparison of Texts

The invocation of Item Response Theory in content anlaysis demands further consideration of its main tenet—which the joint scaling literature have acknowledged, if not fleshed out in detail. Recall Bailey's (2007) assertion that certain "bridging assumptions" must be made when comparing individuals in a common policy space. These assumptions treat responses to certain items as if they were generated under the same circumstances, thereby generating ideal points in the same space. This allows for the direct comparison of the individuals on the basis of their ideal points.[11] In the language of Item Response Theory, there be a common *item response function* for all individuals pooled, and for all individuals for which estimate were generated.

The assumption of a common item response function relies on three components: comparable preferences, comparable choice spaces, and a valid measurement instrument. Put simply, "do they care about the same things, are they able to express how and how much they care about them, and does our apparatus do equally well interpreting that expression in both cases."

Comparable preferences are important because it suggests the pooled observations have meaningful, non-zero values on the same latent trait dimensions. For example, two individuals would have comparable preferences if they both held preferences on

[11] I have noticed in the course of my research that when a researcher makes "bridging assumptions," she, rather than defending them on the basis of a first-principles argument, usually defends them on the basis of pointing out idiosyncrasies in the data generating processes that might make them *incomparable*, and then by minimizing those idiosyncrasies. For instance, Bailey (2007) makes the point that a non-vote in Congress might mean something very different than a "non-vote" made by a Justice.

gun control; they would not be comparable if one individual held the preference on gun control but not gay rights, and the other only held a preference on gay rights.

Comparable choice spaces are important because it ensures that individuals who do share comparable preference sets are able to express measurable data about those preference sets. For example, two legislators may both hold preferences on gun control, but only one of them has the ability to express that preference, because the other legislator's party suppresses her ability to vote as she intends to. As another, perhaps simple example, two people may both hold preferences on ice cream, but be unable to reconcile those preferences because one person is trapped in a block of ice, unable to speak or gesture to indicate their preference.

Finally, a valid measurement instrument is important because it ensures that any observed discrepancy is due to fundamental differences in expression, and not due to a systematic error on the part of the researcher. For instance, racial self-censorship on vcertain items is an issue which has plagued the ANES and other surveys for decades: there are questions one can't get "good" responses for simply because the item's ability to recover an estimate is related directly to the respondent herself.

Assuming a valid measurement instrument, the item response functions for two people are comparable when they share comparable preference spaces and when they have the ability to make the same choices. This is not always the case. For example, survey respondents make snap, unreliable judgments in a low information, low stakes environment (Zaller 1991). Legislators, however, must carefully and painstakingly be informed by a trained staff about the consequences of their choices; lobbyists, other legislators, and the media constantly impinge upon them their beliefs. The

bills in which they vote are complicated, and although someone less privy to the legislative process might be forgiven for missing details in them, legislators face a much higher level of scrutiny. The choices they make have a direct link to their prestige and job security. Meanwhile, however, joint scaling methods patently assume that "respondents may be treated as guest senators stopped in to vote on a small number of issues." The choice and preference spaces between the constituents and their representatives as studied by Miller and Stokes (1963) are therefore somewhat incomparable at the outset (but can still be analyzed correlationally).[12]

Thus, for two observations to be appropriately jointly scaled, they must have similar item response functions. The tenet of comparable item response functions, by way of both analogy and theory, may equally apply in the comparison of texts. I offer three *bridge criteria*, which are repackagings of the item response function requirements, made more appropriate and interpretable for the purpose of corporal comparison.

1.3.1 The Bridge Criteria in Textual Analyses

To make clear the bridge criteria, consider again our running example, delineated in the introduction of this chapter. Without the availability of the regression model of

[12]On policy issues, it may not always be easy or straightforward to determine if the opinions of two individuals should be comparable. It is much easier to do so in the context of policy issues and surveys, however, than in the context of language. An individual's response to "Do you support a war in Iraq?" on a survey should be comparable among individuals, perhaps even if the individuals face different choice spaces and have difference global preferences, because the question is very specific and concerns a well-defined issue. In text – especially text from other "vinculated" data sources, it is usually entirely unclear if the number of times an individual says a single word (a single token) "Iraq" should be comparable to the number of times another individual says "Iraq", and what, if even, the direction suggested by any difference may be.

ideology on congressional record text, the researcher may have taken a more subjective approach. She would have collected several documents from the congressional record. The researcher, usually an expert in the topic of the analysis, would create a standard set of rules by which she will score congressional record documents on the basis of ideology, and memorializes these rules in a codebook. These rules may include topics of discussion, certain cued concepts, specific keywords, or even the emotional reaction the reader has to the text. She then reviews (sometimes, with the help of research assistants) each congressional record document and labels it with the ideology the codebook suggests. The result is a set of documents, each of which has one or more ideology labels produced as a function of the codebook. She has used the rules from the codebook as a *measurement instrument* for the ideological categories she wishes to detect (Neuendorf 2016).

Now, consider a common circumstance, in which the researcher uses that same codebook on a new set of documents—a new *corpus*. In the running example, the new corpus is the set of newspaper documents. May the researcher compare the new labels to the labels from the original corpus, as if comparing apples to apples? The communal answer is that she may, so long as she defends the ability of the codebook to recover comparable measurements, and interprets the results within the literary constraints of the time period (or, makes explicit her intent to interpret them vis-à-vis the context of the present day). Thus, we have the bridge criteria—the item response function for the analysis of text. With respect to our running example, the researcher must account for differences between how the newspapers and Congresspeople are able to address issues entailed in ideology, like setting taxes and determing environmental

policy. Congresspeople have at their disposal the largess of the federal government, of which the Department of Agriculture and the Department of Energy play huge roles (at the direction of the President and several agencies). Newpapers, on the other hand, may stake out positions but may take little individual action on them. The contexts are significantly different, and as such, must be considered with care when comparing directly the amount of time a Congressperson a newspaper spend on taxes and the environment. In this instance, an equal rate of time spent would actually imply a much greater focus on the environment for the congressmembers, simply because the talk is not cheap.

The codebook must recover comparable measurements of the latent trait the researcher wishes to study. This is important because it ensures values generated from one corpus are comparable on a ratio basis to the values generated from any other subset of the data (such as another corpus), because they are both anchored in the same data generating process (Coombs 1960). There are three criteria which underpin this requirement:

1. **Comparable Preference Spaces**: the latent trait, or traits, of interest must be present and non-zero valued in both corpora. *(Running example: the concept of a voting-relevant ideology must be available for both congressmembers and newspapers.)*
2. **Comparable Choice Spaces**: the generative relationship between any latent trait and the text one observes must cause patterns in the observed text which are related to the latent trait, and the detectable patterns must be the same in

both corporal data generating processes. *(Running example: the way that congresspeople determine or are assigned ideology – voting – must be approximately similar to the way that newspapers determine or are assigned ideology.)*

3. **Valid Measurement Instrument**: the codebook must similarly be able to recover an estimate of the latent trait from each corpora's text in the same way; no systematic relationship may exist between the ontology of the corpora and the instrument. *(Running example: the codebook or regression model is assumed to detect approximately the same effect for every word spoken in congress on ideology as there is for every word written in the paper on its ideology.)*

Let these criteria be the *bridge criteria*. In smaller, qualitative studies by a subject matter expert, these criteria are hardly impeachable.[13] For quantitative studies, however, there exists a statistical approach by which one may verify the comparison of the labels to each other, which this dissertation introduces.

These criteria lay out the circumstances under which we would expect corpora to be comparable, and make apparent a quantitative test statistic that may critically evaluate the ability of an analysis to reach conclusions consistent with our strong political science priors. The approach from the perspective of a choice space is critical: what is the underlying process by which political actors choose the language we observe? Rather than characterize the choice set over all possible types of language,

[13]Though it will not be considered at length here, the statement that the researcher is "able to interpret the results within the literary constraints of the time period" is important because a concept's meaning changes over time. The interpretation of the song, "Baby It's Cold Outside" has, for example, taken on an entirely different meaning as of the writing of this dissertation. The #MeToo movement revealed the content of the song to be suggestive of an inappropriate relationship between the male and female vocalists. The dissertation's author does not comment on whether the interpretation is the right one.

this project characterizes it as a choice over how to sample from a smaller number of latent topics that perhaps underpin all political corpora. Consistent with even the most basic of frequency-based methods, the theory suggests that corpora must share at least one latent dimension to be comparable.

Similar Concepts to the Bridge Criteria in Current Political Science Research

The spirit of the bridge criteria is not novel. In fact, Hopkins and King (2010) provides a prominent exemplar of how to consider the bridge criteria, in the context of a supervised machine learning approach to text analysis.[14] In the paper, the authors explicitly call out their assumption that

$$W^L = W^U, \tag{1.1}$$

where W^L and W^U are the matrices of features which predict the label set in the labeled and unlabeled corpora (the authors call W the conditional feature matrix; they use the notation "X"). Jerzak, King, and Strezhnev (2018) later relax this assumption, because their application only concerns population-level estimates, denoting it (with modified notation) as

$$\mathbb{E}[W^L] = W^U. \tag{1.2}$$

[14]The authors build on the point made by Hand (2006), who shows that all classifiers make the assumption of a shared support.

Equations (1.1) and (1.2) are important because they show how at a basic level, the textual measurement instrument for the latent dimension of interest must be supported by the same (potentially augmented) set of related preferences or choices—in this case, set of observed words.[15] Any estimate based on an unsupported preference an choice set from another corpus is not useful, because it has no grounded index.

Hopkins and King (2010) and Jerzak, King, and Strezhnev (2018) are somewhat unique in their treatment of the concept. It is not common, even in influential research, to state the assumption, let alone test it.[16] The mere fact that it is acknowledged by the paper, however, goes to show that the consideration of comparability is not a new idea. Moreover, it goes to show that the logic behind what might make two or more texts comparable is not a new idea. Clearly, other researchers have considered that ignorance of the bridge criteria – though perhaps not called the bridge criteria – can result in findings that are undercut after the fact.

The logic of Hopkins and King (2010) may be extended further (though the authors do not go so far as to do so). Assume that the expected set of words present in the measurement instrument is satisfactory. Researchers then often make the assumption that the presence of the word in both context implies the same *directional meaning*, or weight. In other words, that the monotonic and increasing use of the word "death tax" implies increasing levels of conservative ideology, in both contexts. This assumption is also problematic. This dissertation does not in detail consider the monotonicity of

[15]Furthermore, the words must be related to the dimension of interest *monotonically*: an increasing relationship between the frequency of the word and the dimension of interest must hold in every subset of the data.

[16]In fact, Hopkins and King (2010) and Jerzak, King, and Strezhnev (2018) *do not* test the assumption.

the weights assigned to particular words in the measurement instrument; rather, it simply looks to see if patterns in usage of words observed are similar in both corpora. The application of the method to the idea of weights (linear, non-linear, monotonic, and non-monotonic) is a future research direction.

How Automated Content Analyses Fail to Test the Bridge Criteria

One of the implications of joint scaling is that an estimator derived using data from one group may be used to estimate a comparable quantity of interest and the other group. This implication has been used quite profitably in circumstances where quantity of interest is unobservable in one of the groups. For example, Gentzkow and Shapiro (2010) use a model trained on legislator data to predict ideology for members of the media. In essence, the approach assumes that the data may be jointly scaled because they are drawn from the same distribution, or produced by the same data generating process. However, in pursuit of the estimand, that assumption is taken for granted, and we jump directly to the prediction of the estimand.[17]

The popularity of this approach in the text as data literature continues to grow. Today, however, no papers have tested the key assumption on which joint scaling arrests—that item response functions are constant, or comparable, across groups. This goes for both the canonical joint scaling applications and the text as data joint

[17]It is my opinion that the exuberance with which we have applied machine learning methods to causal political science research has prevented us from carefully considering the implications of the methods we apply. For instance, while Poole and Rosenthal (1997) are quite hesitant to compare estimates of congressperson ideology across chambers, even when the choice and preference sets of Senators and Representatives are inherently more comparable than those of legislators and constituents, we flippantly accept that we may easily compare news and speech. This seems as if the rules of inference have been applied inconsistently.

scaling applications, which are newcomers to the scene. The peril of failing to test these assumptions, especially for high dimensional text data, is that the likelihood of false hypothesis confirmation becomes exponentially more likely as the dimensionality of the data increases. In other words, if this trend goes unchecked, we risk drawing false conclusions.[18]

1.4 A Framework for the Evaluation of Bridging Assumptions

With the three bridge criteria established, how might we establish a method by which we may test them? In the ground truth scenario, we would be able to examine if a model successfully bridges two corpora because we would be able to observe the values for the latent traits themselves. The preference spaces would be comparable if we see values on the latent traits in both cases; the choice spaces would be comparable if we observe clustering of language related to those traits in both cases, and; instrument validity would be achieved through cross-validation. With simple cross-validation, we would compare the predictions from a instrument, codebook, or model trained on one

[18]The emergence of automated content analysis has made it easier to generate efficient codebooks focused on a set of statistically relevant keywords, predict category labels by weighting the contributions of these keywords within the codebook, and apply the codebook to a diversity of corpora at an incredibly low cost. As a result, subject matter expertise and resources are no longer limiting factors; anyone may conduct an automated content analysis and draw conclusions from it, because the resource gateways are no longer protected by the necessity of design. Machines allow for the application of a single codebook at astonishing scale, but how do we know if the codebook used is the right one? It is harder than ever before, as a reviewer and reader of published research, to understand the sensitivity of an automated content analysis to the choices of the researcher, and how those choices affect the validity of the inferences the researcher makes. There is a (not so) popular concept in data science, that machines do dumb things faster, and with impunity. This flies in the face of the press, who tend to fetishize how knowledge may be hidden to the human, but seen with priestly ability by the machine.

corpus to the real values for another. The problem is that we can't do this because most often values aren't observed in the other corpus.[19]

Problematically, the model fit is almost never evaluated on the out-of-sample data, and there is never much theoretical argument given as to why such an approach is valid. Usual research methods take a strict "prediction" approach without considering the theoretical implications of the approach. Indeed, the dimension ostensibly estimated might not even exist in both corpora.

For instance, in our running example, the researcher could never have observed the ideology values Y to begin with. Our mean estimated ideology $\bar{Y}$ among the newspaper units is "NA", unless we use $\hat{Y}$.

It is possible to construct a statistical approach to the evaluation of the bridge criteria. A method which estimates the nature of the preference spaces, choice spaces, and instrument validity is readily available in the form of of topic modeling. The critical assumption involved in using a topic model to estimate the underlying preference and choice spaces is that the language we observe is actually related to the latent traits of the individual, or data generating process, which produced the text.

1.4.1 Why Topic Models Can Recover the Basic Space

Computerized content analyses usually take a frequentist approach to language, representing documents as frequency distributions over words, and drawing inferences

[19]One solution, given that we cannot observe true values, is to produce values for the other corpus using human coders. This is cost-prohibitive, and in some cases impossible due to the computational complexity of the operations. Another solution is to do the usual prediction but then conduct ex-post model checks using the procedure discussed in the Introduction. This is suboptimal because it requires the time and technical knowledge necessary to do the empirical evaluation.

about the document's content from the nature of the distribution. The comparison of documents follows suit, drawing inferences about the similarities and differences between documents from the similarities and differences in their distributions. This approach has seen great success in applications where prediction is the primary concern (Mosteller and Wallace 1964, e.g.), as the frequency of words in a document tend to relate to the quantities researchers wish to predict, such as ideology (Laver et al. 2003).

Prediction, however, does not demand careful examination of the model by which language was generated; the presence of a word or phrase in one document might be used in another document but not for the same reason it was used in the first. This is principally why it would be inappropriate to state that a topic model is not necessary to compare two texts. Simply comparing two texts on the basis of the marginal frequency distributions over their words entails no transformation which would reduce the risk that we treat a word in one corpus to mean the same thing it does in another corpus.

Moreover, there is a major inferential issue which comes along with a marginal comparison of two document's frequency distributions: we may not make an inference about the comparability of two texts when there is no expectation of what the texts would have looked like had they not been comparable. Indeed, the most reasonable reference distribution for a direct comparison of political texts is a randomly generated page of text.[20] This is problematic because any two texts generated by a human process will almost assuredly result in more similar language than we would expect

[20]Some of my colleagues have suggested using "everyday" textual sources as a reference distribution, but I am unable to find any suitable published references for the practice of what I will call "quotidien benchmarking."

by chance (even the number of permutations in a fifty-character "tweet" exceeds the number of atoms in the universe).[21]

Topic modeling, introduced in earnest by Deerwester et al. (1990), is an approach to the statistical modeling of language which represents chunks of language as a function of a smaller set of topics, rather than words. The main innovation was the idea that we can model a hierarchical process, by which common shared topics among all documents, and then fit those topics to the data we observe through a maximum likelihood estimation procedure. More recent probabilistic approaches, which model latent "topics" using Bayesian methods, specify the data generating process as part of the model (Blei, Ng, and Jordan 2003; Roberts, Stewart, and Airoldi 2016). The data generating process specified posits that there is a set of latent topics which exists among the genesis of the corpus. The topics are represented as probability distributions over words. Documents are generated by sampling from the set of potential topics, and then by sampling from the distribution of words within those topics.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of K topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is given by:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_k \mid \theta) p(w_n \mid z_k, \beta), \tag{1.3}$$

[21]For more intuition on the concept of null distributions in text, it is instructive to consult the example of the Library of Babel (Borges 1941). The Library of Babel contains all texts that ever have, are, or will be generated, simply by the fact that it contains all permutations of text which are possible. In fact, it contains all knowledge and secrets, because all knowledge and secrets may be communicated in written form. The only problem is, of course, being able to find them! This is a case for rationally bounded search for knowledge.

where $p(z_k \mid \theta)$ is θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_k} p(z_k \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta. \tag{1.4}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(\mathcal{D} \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dk} \mid \theta_d) p(w_{dn} \mid z_{dk}, \beta) \right) d\theta_d. \tag{1.5}$$

In the context of the bridge criteria, the mixture θ and the set of topics $\mathbf{z}$ is of primary interest—they at a level of abstraction higher than the document and the words we observe, akin to the basic space of preferences and choices for any individual of interest. These parameters are what distinguish the topic modeling approach to the evaluation from bridging assumptions from that of Hopkins and King (2010).

If frequentist approaches are able to capture relationships between language and latent dimensions, as demonstrated in appendix B.4.1, then so too must topic models. Topic models, however, are different from n-gram approaches because they explicitly model the process by which latent traits generate observed text. As delineated in equation (1.3), topic models treat the probability of observing patterns in language as a direct function of the latent space of dimensions from which entities sample to generate text documents. This provides a very simple and intuitive approach for the modeling of latent traits, which is also compatible with the bridge criteria (section 1.3).

It merits noting that this approach to solving the problem of testing for comparable corpora does not claim to provide a method for approximating the "true distribution" over latent traits; further, it does not claim to characterize directly the joint distributions over traits among two or more corpora. Instead, this is a theoretically motivated statistical approach that intends to solve a problem in text analysis. It is in fact impossible to ever validate the ability to a topic model – or any other model – to recover the "ground truth" for latent traits, since the traits themselves are fundamentally observable. The best we can do is create "index" representations of the traits and attempt to validate them with multiple coders and good theory (Krippendorff 2012).

Specifically, if we assume that latent topics are to a significant extent driven by latent traits, then we may treat the latent topics recovered by the topic model as *measurements of* the latent trait space (with non-trivial noise). Further, if latent topics are able to sufficiently recover a diverse space of the salient latent traits expressed by an individual, then we may treat the *set of topics recovered* as estimate for an individual's basic space of preference and choice. The presence of a matchable topic may represent the preference space, and the comparable distribution of a topic to any other may represent the choice space. To further prove why modeled topics are significantly related to latent traits, consider the following example.

1.4.2 Sensitivity of Analyses to the Measurement Instrument

The final bridge criterion requires that the estimator used to perform automated content analysis – the machine-generated codebook to be used in the analysis – is produced reliably, such that the measurements it "takes" in one corpus are comparable to measurements taken in the other corpus. It is important to consider the satisfaction of this criterion because, as Denny and Spirling (2018) demonstrate, there is a significant and consequential degree of variability that may result in the application of an estimator when different pre-processing steps, and structured representations of the underlying text, vary.

Further steps beyond pre-processing, such as dimension reduction, may also affect the results the researcher observes. Depending on the researcher's definition of the corpus's support, the empirical results and performance of estimators will vary. For instance, I, and numerous others, have observed in the course of research that the performance of generative topic models, and especially deterministic techniques like non-negative matrix factorization (simplified latent semantic analysis), is highly conditional on the corpus's support.

Finally, final generated quantities of interest are very sensitive to the specification and weighting of the support in any automated content analysis. As the shared support of two documents increases in size, the estimated similarity between those two documents is biased downwards, as is the sampling variability of the similarity. This is because the influence of a change in any token rate grows smaller as more features

have the opportunity to affect the similarity. Empirically, this results in a variable which has bias in its central tendency that is a function of researcher decisions.

How many other supervised learning approaches would reveal similar results if they were subject to the same test? It is time-consuming and technically demanding to do so. The establishment of the comparability of texts *ex-ante* would enable researchers to test the assumption of comparability before pursuing research based on it; further, it would provide an objective data point which researchers could use to cultivate confidence in their findings. This dissertation therefore also proposes an *ex-ante* tool, based on theory, that researchers may use to test for comparability.

The method does not produce a test statistic which may allow the researcher to reject a result generated from a comparison of political texts. Instead, it bounds the sensitivity of the analysis non-parametrically, so that the researcher may have confidence in the significance of any observed result, and the ability of the researcher's approach to satisfy the third bridge criteria.

1.5 The Empirical Comparison of Disparate Texts

To introduce the notation used herein, consider two sets of textual documents. Let $\mathcal{D}^L$ include N^L labeled documents, and let $\mathcal{D}^U$ include N^U unlabeled documents. The generic subscript i indexes individual documents within each set, and the total number of documents is $N = N^L + N^U$. Each document may be assigned a value y_{ik}, which indicates a document's membership in each category label k, for an exhaustive but not mutually exclusive set of label categories, $k \in \{1, \ldots, K\}$. Labels y_{ik} are not observed

in the unlabeled document set.

Many researchers will apply a topic model to the high-dimensional text to produce a lower-dimensional representation of the text that will not break standard regression (and is more readily interpretable). Topic models represent word counts as multinomial observations, parameterized by a weighted sum of latent *topics*. A topic model may be considered as a lower dimensional factorization of the multinomially distributed text:

$$W_i \sim \text{MN}(Q_i, m_i), \tag{1.6}$$

where $m_i = \sum_j W_{ij}$, Q_i is a document-specific word probability vector over K latent traits v_{ik}

$$Q_i = v_{i1}\theta_1 + v_{i2}\theta_2 + \ldots + v_{ik}\theta_k, \tag{1.7}$$

and θ_k is the trait-specific probability distribution over words. The term v_{ik} is therefore the amount of the document owing to latent trait or topic k, and θ_k tells us what patterns in language indicate the presence of v_{ik}. Each document $\mathcal{D}_i = W_i$ may be a mix of topics, or it may have a single topic, but the sum over all topic probabilities for a document must be 1. Since K is much smaller than W, each Q_i is the projection of the document into a lower dimension space. The full specification of the likelihood function is available in equation (1.3).

To empirically define the problem at hand, Consider a simple two equation *text regression* model, in which the unit of analysis is the individual represented as a

set of documents, and two document subsamples are pooled within the "population-representative" sample of N units. Per our running example, one subset is the member of congress, and another subset is the newspaper. We regress a dependent variable of interest y (*e.g.*, ideology) on estimated topic proportions resulting from the model, using the logic of Bartels (1991). The equation for individual i (from the first subsample; *i.e.* the "Members of Congress" in our running example) and j (from the second subsample; *i.e.* "News" in our running example), is:

$$y_i = v_i\beta + e_i,\ i \in N, \tag{1.8}$$

$$y_j = v_j\gamma + e_j,\ j \in N,\ i \neq j, \tag{1.9}$$

where v_i is a $1 \times K$ vector over predicted topic compositions for documents i, β and γ are $W \times 1$ vectors of parameters. By convention, β and γ are usually the parameters of theoretical interest; however, in this case, they are the weights which compose the "codebook" used for the estimation of the latent trait in the sub-population. Assume generally that the models are well-specified.

How should we estimate the β and γ codebooks? At one extreme, we may estimate β and γ entirely separately using any flavor of regression. The generalized regressions equations are:

$$\mathbb{E}[y_i \mid v_i] = v_i\beta,\ \forall i \in N,\ \text{and} \tag{1.10}$$

$$\mathbb{E}[y_j \mid v_j] = v_j\gamma,\ \forall j \in N. \tag{1.11}$$

At the other extreme, we may treat β and γ as if they were the same, or similar codebooks:

$$\mathbb{E}[(y_i, y_j) \mid (v_i, v_j)] = (v_i, v_j)\alpha, \ \forall i, j \in N. \tag{1.12}$$

To keep it simple, let's consider the case of ordinary least squares regression. The pooled OLS estimator is

$$\alpha = (v_i'v_i + v_j'v_j)^{-1}(v_i y_i + v_j y_j), \tag{1.13}$$

$$= (v_i'v_i + v_j'v_j)^{-1}(v_i'v_i\alpha + v_j'v_j\beta), \tag{1.14}$$

$$\mathbb{E}[\alpha] = \beta + (v_i'v_i + v_j'v_j)^{-1}v_i'v_i(\gamma - \beta). \tag{1.15}$$

In other words, α is the matrix-weighted average of the separate parameter vectors α and γ, with a weight matrix that is proportional to the inverse of the covariance matrix of that parameter vector.

It is then hypothetically possible to test the hypothesis that the parameter vectors β and γ are identical by comparing the sum of squared residuals from the two subset regressions against the sum of squared residuals from the pooled regression. If the improvement in fit is large enough, the null hypothesis of parameter inequality will be

rejected. The test statistic

$$\frac{\text{SSR}_{\alpha} - (\text{SSR}_{\beta} + \text{SSR}_{\gamma})}{K} \Big/ \frac{\text{SSR}_{\beta} + \text{SSR}_{\gamma}}{N - 2K} \tag{1.16}$$

is distributed F with K and $N - 2K$ degrees of freedom, where $\text{SSR}_{\beta,\gamma,\alpha}$ is the sum of squared residuals for recovered estimates using each parameter vector. However, it is impossible to compute equation (1.16) using observable, ground-truth data (we only have $\hat{y}$). This suggest that we may need an alternative approach.

1.5.1 Derivation of the Delta Statistic

As earlier works readily show, equation (1.15) is more than enough to demonstrate that the pooling of two distinct sets of documents (two distinct corpora) risks biasing estimates of the codebooks β and γ. If the goal is to entirely avoid bias, then two separate regressions should always be run. This fact, however, makes Type II analyses entirely untenable, since y_j is fundamentally unobservable! One major conclusion to be drawn from this is that there is significant risk of bias in our machine-generated codebooks, but this bias is not quantifiable because we often do not have valued outcomes for a significant subset of the data.

Building on the theory developed earlier, it is possible to interpose the by-products of the topic modeling procedure with the test for if β and γ are the same, by positing that there exists a relationship between the label y and the latent traits v_k. This is commonly assumed as part of supervised LDA (sLDA), which estimates topics conditional on labels provided about those documents Mcauliffe and Blei (2008).

Formally, this suggests that

$$\text{Cov}(\beta, Y) > 0, \tag{1.17}$$

$$\text{Cov}(\gamma, Y) > 0, \tag{1.18}$$

$$\text{Cov}(Y, V) > 0, \text{ and} \tag{1.19}$$

$$\text{Cov}(Y, (\beta, \gamma)) > 0, \tag{1.20}$$

where V is the set of reduced-dimension topics estimated by a topic modeling procedure. Therefore, it is possible to construct a test for whether we would expect β to be similar to γ by using some function of the θs from their topic models,

$$\delta(\text{News}_i, \text{Congress}_j) \sim (v_i' v_i + v_j' v_j)^{-1} v_i' v_i (\beta - \gamma), \tag{1.21}$$

$$\sim f(\Theta_i, \Theta_j). \tag{1.22}$$

This suggests that the F-test in equation (1.16) – which is impossible to estimate, since we do not have ground truth values for $\hat{y}$ – may be approximated using equation (1.22). The δ-estimator may approximate a rejectable fit statistic *without the use* of y or the inestimable regression on its missing counterparts, simply by using a reduced form of patterns in language. What is required is the full-rank sparse text data produced by the distinct corpora, and a mixture model that may specify distributions over the reduced form. The following sections explain in greater detail how to use this insight to generate values for the δ-estimator.

1.5.2 Linkage of Independently Estimated Topic Models

In order to estimate $f(\Theta_i, \Theta_j)$, we first generate two independently estimated topic models—one for each corpus. One challenge resulting from the use of two independently estimated topic models is that there is no information linking the topics across corpora. This means that even in the case that the corpora have very similar topical distributions, we may observe a topic matrix $\Theta^L \neq \Theta^U$ because the row indexes are misaligned. We must re-index $\{\Theta\}^U$ to align each topic in one corpus to its most similar analog the other corpus. At the same time, we cannot allow any topic to be matched twice. To solve this challenge, we may implement a greedy matching algorithm, which compares each topic in one corpus to every topic in every other corpus, and matches the topics by optimizing the continuous correlation between them, $\rho(\cdot, \cdot)$.[22]

Linking topic models is possible for any method which produces an intermediary matrix of topics which are probability distributions over words. In fact, the method may use matrix factorization methods such as singular value decomposition (SVD, also called latent semantic indexing or LSI), Poisson factorization (PF), or other distributed operations such as local least squares decomposition. The preferred method in this dissertation will be to apply latent Dirichlet allocation (LDA), though a structural topic model (STM) or correlated topic model (CTM) with information from a set of variables may also be used. The general method approach in which two independently

[22]Other distance metrics D, some of which are discussed in equation (1.28), may be employed to a similar, but different, effect. The correlation approach compares distributions over features relative to their global averages, which has the benefit of adjusting for features outside of the shared support. Correlational approaches will outperform distance-based approaches when the concentration over features is lower (higher entropy), and underperform when the concentration over features is higher (lower entropy).

estimated topic models of the same kind are linked via a matching technique is called *tl-LDA*. In the case other techniques like non-negative matrix factorization or latent semantic indexing may also be used in place of LDA (tl-NMF and tl-LSI for short).

The first step of the tl-LDA approach is to run separate topic models on the naturally separated corpora. For each corpus $\mathcal{D}^U$, $\mathcal{D}^L$, produce the $N \times P$ sparse count matrix of features W. In the case of tl-NMF and tl-LSI (or other matrix factorization methods), the tf-idf transformation of $\tilde{C}$ may be used in place of W:

$$\tilde{W} = \text{tf-idf}(W). \tag{1.23}$$

The matching technique tl-LDA generates potential matches by finding for each topic in the model from $\mathcal{D}^L$, θ_i^L the most similar topic in the model from $\mathcal{D}^U$, θ_j^U that hasn't already been matched. The algorithm, delineated in algorithm 1 is a greedy matching algorithm where the objective to be maximized is the continuous Pearson's correlation coefficient between the distributions over tokens. For each topic in the model for corpus i, ranked by overall prevalence of the topic in i, compute a σ-matrix with all θ_j. Pick:

$$\arg\max_{\theta_j} \rho(\theta_i, \theta_j). \tag{1.24}$$

This results in the $K \times M$ re-indexing matrix I, and the $K \times M$ similarity matrix S. The algorithm for matching, using Iverson bracket notation, is reported in algorithm 1.

Finally, we re-index all Θ^U:

$$\left\{\left\{\Theta^U[i,:] := \Theta^U[I[i,m],:]\right\}_{i=1}^{K}\right\}_{m=1}^{M}. \tag{1.25}$$

There are some general effects of this matching procedure. First, it tends to match topics with similar support spaces. This is because the contribution of non-zero feature in both corpora will be greater than the similar case where the feature is zeroed out in one corpus but not in the other. This has the desirable benefit of finding topic matches for topics that have the same words (the same support space). This is doubly beneficial because the theory supporting the bridge criteria suggests that shared support space will be related to comparability.

Second, it will produce only one match for topics that appear to be related to multiple topics. This has the effect of reducing the contributions of similarity due to a dispersed, poorly estimated or "global" topic in one corpus. The drawback is, of course, that perhaps a nice topic might match to the poor topic. Future developments of the method might consider dropping dispersed topics from consideration from the outset. This alternative approach, however, must consider what the reduction in the size of the match set suggests, and how an orphaned topic match might affect variability.

Using our running example, the delta-statistic works by running two separate generative or deterministic topic models on the naturally distinct texts from the naturally distinct populations. For instance, in the example above, we would run one topic model on text from the Members, and another topic model on text from

Data: Θ_m, $\Theta_{m'}$
Result: **I**, **S**
initialize stack `sims`;
for *i in range(K)* **do**
    initialize stack `isims`;
    for *j in range(K)* **do**
        sim := $D(\Theta_m[i,:], \Theta_{m'}[j,:])$;
        push sim to isims;
    end
    push isims to sims;
end
S = matrix(sims);
initialize stack `matches`;
for *ip in range(K)* **do**
    amax = argmax(**S**);
    Kdiff = K - ip row = amax // Kdiff;
    col = amax % Kdiff;
    ii = i.pop(row);
    jj = j.pop(col);
    val = $\mathbf{S}[row, col]$;
    push (ii, jj, val) to matches;
    newi = range(Kdiff);
    newj = range(Kdiff);
    S := $\mathbf{S}[newi, newj]$;
end
I = matrix(matches);
S := $\mathbf{I}[:, 2]$;
I := $\mathbf{I}[:, :1]$;

Algorithm 1: Greedy Matching Algorithm for Linking Topics

the Newspapers. This produces two separate topic models, each of which may be considered as a lower dimensional factorization of the multinomially distributed text:

$$W_i \sim \text{MN}(Q_i, m_i), \tag{1.26}$$

where $m_i = \sum_j W_{ij}$, Q_i is a document-specific word probability vector over K latent

traits v_{ik}, and θ_k is the trait-specific probability distribution over words:

$$Q_i = v_{i1}\theta_1 + v_{i2}\theta_2 + \ldots + v_{ik}\theta_k, \tag{1.27}$$

The term v_{ik} is therefore the amount of the document owing to latent trait or topic k, and θ_k tells us what patterns in language indicate the presence of v_{ik}. Each document $\mathcal{D}_i$, represented as W_i in full-rank form, or Q_i in reduced dimensional form, may be a mix of topics, or it may have a single topic, but the sum over all topic probabilities for a document must be 1.

We then take these two separate topic models from the Members and the News texts and align their topics using a greedy matching procedure, to generate topic "matches." We do this because topic models provide no guidance as to how topics estimated in one run of the model align to topics in another run of the model. The criteria that determines the match is the response distance $R_{kk'}$ between any two topic probability distribution, where each greedy match stage uses the smallest KL-divergence available to determine a match. As a result, we get a number of topic-to-topic matches, and each match is given a response distance. We also simulate for each match a null distance that we would expect to have seen given a similar but random topic probability distribution.[23]

[23]It is noteworthy that LDA often doesn't yield unique topics. An obvious drawback of the greedy matching method is that it downweights the estimated comparability of two corpora when any one of the corpora is more likely to produce non-unique topics. I discuss this further later on.

1.5.3 Response Distance

Even with the aligned matrix of topics, the matching method simply matches the *closest* possible topics—even if the topics are relatively far away. Therefore, we must also estimate which matches are "real," by determining if they are closer than we would expect by chance. Here, the *response distance* comes into play.

Desirable properties of the distance would be the following. First, the distance could indicate how similar the choice spaces for two matched topics are; it would increase as the choice spaces are less similar, and decrease towards zero as they are more similar. Second, the distance metric may be used to rank matched topics with comparable distances, such that a larger distance implies a less similar choice space. The response distance is related specifically to any given topical pair, and not to the corpora overall (though section section 1.6 expands on the distance to allow for overall comparison).

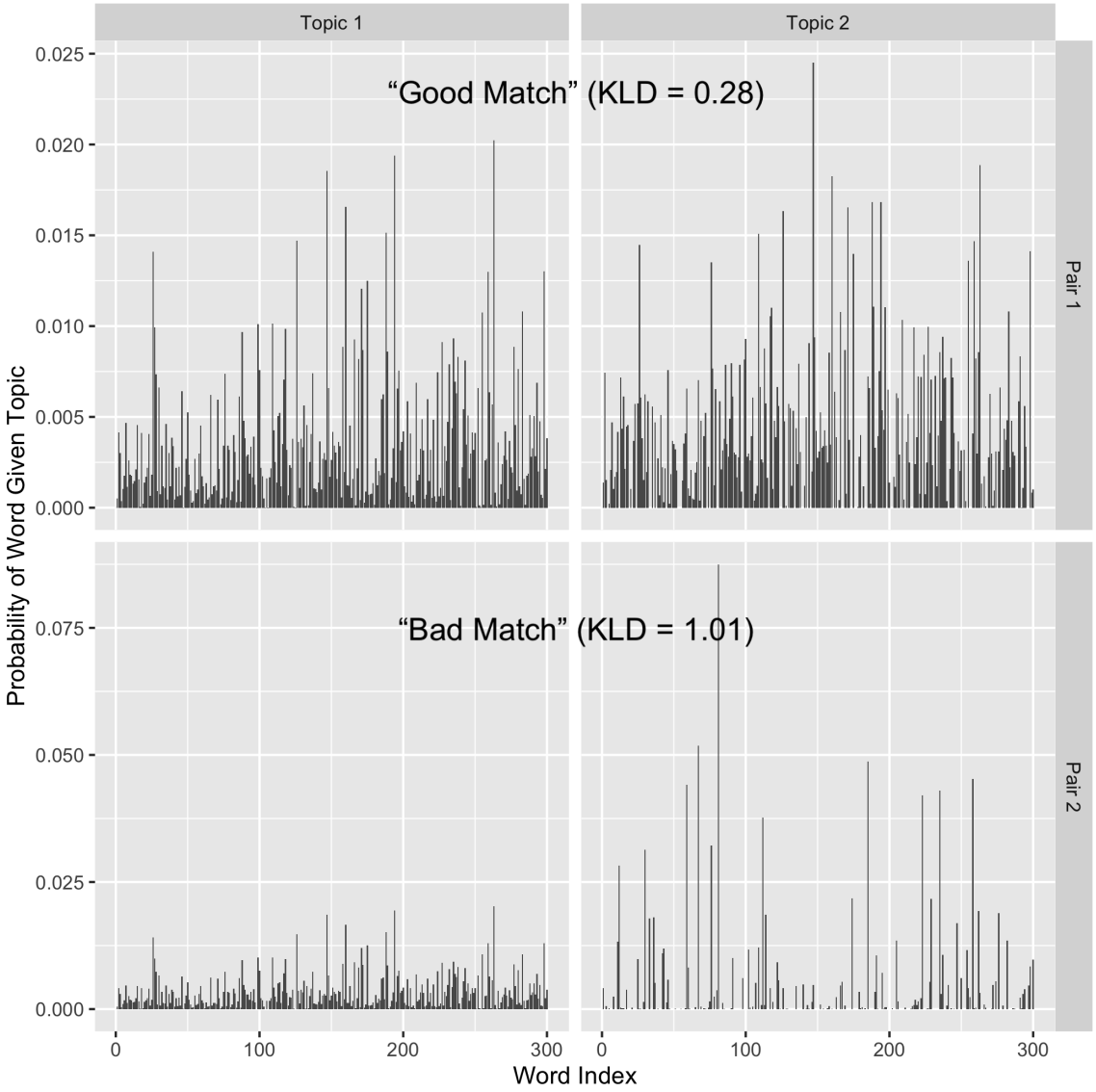
I propose that the candidate distance metric for the response distance be the KL-divergence, which is used to measure the ability of one PDF to explain the information in the other PDF. The response distance for any two estimated topics is:

$$R^{kl}(\theta_i, \theta_j) = \int_a^b \theta_i \log\left(\frac{\theta_i(x)}{\theta_j(x)}\right) dx, \text{ and} \tag{1.28}$$

where $a = \min(\theta_i, \theta_j)$ and $b = \max(\theta_i, \theta_j)$.[24]

[24]The distance may be substituted for any other distance metric appropriate for the comparison of probability distributions with overlapping support, such as the Anderson–Darling squared-errors approach, or the Kolmogorov–Smirnov largest-difference approach, though the latter has the drawback of over-indexing to the powerful influence a handful of tokens in the support space which are disproportionately different in their rates.

Figure 1.1: Response Distance Finds Similar Patterns in Language



Note: Figure reports exemplar distributions for two topic pairs. Topics on the left were estimated from one LDA model (for instance, from the News corpus), and topics on th right were estimated from a second LDA model (for instance, from the Congressional Record corpus). Each word, indexed by a number, is assigned an estimated probability of observation, given the topic appears in a document. The first pair is a "good match" (it is the same latent topic), and the second pair is a "bad match" (it is not the same latent topic). Accordingly, the KL response distances R reflect their matchability; the good match has a much smaller distance than does the bad match. These response distances are used to compute the δ-statistic.

An alternative specification of the response distance could be based on an analysis of the overlap in the support space for any two matched topics, separately or in combination with the distance between the distributional unit probabilities. This dissertation opts to treat shared support as a researcher specification issue, rather than one which may be tested under distributional assumptions.

Figure 1.1 showcases an exemplar for the response distance. The figure reports simulated distributions for two topic pairs. Topics on the left were estimated from one LDA model (for instance, from the News corpus), and topics on th right were estimated from a second LDA model (for instance, from the Congressional Record corpus). Each word, indexed by a number, is assigned an estimated probability of observation, given the topic appears in a document. The first pair is a "good match" (it is the same latent topic), and the second pair is a "bad match" (it is not the same latent topic). Accordingly, the KL response distances R reflect their matchability; the good match has a much smaller distance than does the bad match. These response distances are used to compute the δ-statistic.

1.5.4 Simulation of Null Response Distances

We also simulate for each match a null distance that we would expect to have seen given a random topic probability distribution. Simulating the null allows us to interpret the distance between two topic probability distributions. When the distance between two observed distributions exceeds that which we would expect by chance, then so too do the patterns in language we would expect to see given the presence of a common latent

dimension. Therefore, observing a null match for two bits of speech that generally use the same words suggests that even though the same words are used, *the ways in which they are used together are not as similar as we might expect.*

It is the comparison of the observed distance to the null distance that allows the delta-statistic to evaluate in absolute terms the quality of the pooled model. It adjusts the distances to account for our theorized relationship between latent traits and language patterns. The result is a residual distance, anchored as a ratio quantity to a theoretically driven threshold. On the practical side, producing this ratio quantity has the desirable effect of allowing for more normally distributed noise owing to the mutability of language. On the theoretical side, this threshold represents the uncanny valley of language—a sort of boundary value for the Turing test. It is the point at which the average person would say to a machine, "you're definitely mimicking me by using the same words I am, but you just sound a little weird." It is the point at which a person reading a newspaper might think, "this article just mentioned cap and trade, but it would read really weird if it was pushing some sort of ideology; maybe its just salient reporting."

The next challenge, then, is to estimate the expected null distance R^0 for each topic. One method for estimating R_0 would be to simulate several thousand random multinomials using a gamma distribution with default parameters, and then compare the realized value to the simulated distribution:

$$\hat{R}_k = \mathbb{E}[R(\theta_k, \text{MN}(p, 1/p)) \tag{1.29}$$

This would, however, result in comparisons that are unrealistic. Comparing a realized topical distribution to a simulated standard multinomial is akin to comparing the language we observe to all possible combinations of language, regardless of the tenability of that lingual distribution. Some combinations of words will always be more likely than others because of grammar, syntax, and context-specificity, and our simulated distributions must reflect this. Further, in a world of bounded rationality, it only makes sense that certain types of language will be more informative than others, as we have yet to search the global information space ,and likely will never be able to (Simon 1997). Ignorance of this fact results in null distributions that are far too optimistic.

An alternative method would be simulate multinomials from gamma distributions of *similar shape* to the topic we observe, and then compare the realized topical distribution to those simulated. This method allows for a more natural model of language, and also produces a much less optimistic null distribution for R_0. To do this, we must first estimate the parameters for the production of a comparable multinomial distribution. By way of maximum likelihood, we can estimate this using the realized topical distribution:

$$\hat{\alpha}^{MLE} = 1 + \mu_\Theta + \frac{\mu_\Theta + \sqrt{\mu_\Theta^2 + 4\sigma_\Theta^2}}{2\sigma_\Theta^2}, \tag{1.30}$$

where $\hat{k}$ equals p, the shape of the distribution over tokens. Then, the best estimator

for the expected similarity is:

$$\hat{R}_k = \mathbb{E}[R(\theta_k, \text{MN}(p, \hat{\alpha}^{MLE}))]. \tag{1.31}$$

Because of the mutability of language, we may wish to conservatively define the threshold at which we may say the similarity is greater than expected as,

$$\hat{R}_k^{95\%} = P_{95\%}(R(\theta_k, \text{MN}(p, \hat{\alpha}^{MLE}))). \tag{1.32}$$

For each topic k, we may say the topic is a "match" if $R_k >= \hat{R}_k^{95\%}$. These "matches" (1=True, 0=False) may be used to compute binarized corporal distance metrics, such as the Jaccard distance. The residuals from these comparison provide the contributions to the test statistic discussed in section 1.6.

This proposed approach to the estimation of the response distance has specificity because it would be incredibly unlikely that we would see the distributions of language related to gender in two disparate corpora simply by chance. The theoretical intuition for this is that language is quite mutable; a set of 30-word tweets, each using the 1000 most common words in the English language, would have more possible random combinations of words than there are atoms in the universe (30^{1000}). Indeed, the core assumption which underpins this approach is that if two separate corpora actually are generated or related to the same underlying dimension, then it would be incredibly unlikely for us to observe topical compositions that look the same in both corpora.

The formulation for R_0 also has the advantage of handicapping distributions that

give equal probability to the majority of the words—in other words, distributions that are more entropic. Entropic distributions will receive more dispersed simulations, which will cause the null distance threshold to be larger. This tracks theoretically, because when a pattern in language is less detectable, we should be more uncertain about whether another pattern in language is a match. This delta-statistic therefore discounts contributions from general topics with no distinct pattern in language. This results in a test statistic that is less sensitive to general topics—though this is balanced by the greedy matching procedure, which also ensures general topics are matched once and only once.

1.6 The Delta Statistic for Text Comparability

While the ability to check for matches and distances on a topic-by-topic basis is useful, it does not give us a definitive answer for the comparability of the corpora overall. To that end, we may wish to produce a test statistic for the null hypothesis of comparability. If we consider the estimated values for the shared trait statistic and the response distance to be realizations from a random variable, then we can test the null hypothesis that the two are comparable by observing if the observed distances are distributed with bias relative to the reference corpus. The δ-statistic packages the response distance to produce a single test for the null hypothesis that the corpora are comparable.

The test statistic is a form of G-statistic, an analog of the KL-divergence, which is appropriate for the comparison of probability distributions. This statistic is a

likelihood ratio that takes the form

$$\delta = 2 \sum_{k=1}^{K} O_k \ln \left(\frac{O_k}{E_k} \right), \tag{1.33}$$

where O_k are the observed response distances and E_k are the expected values under the null of comparability (using the simulated θ'_k. The statistic is distributed proportionally to the KL-divergence, which is appropriate for the probabilistic comparison we make:

$$\delta = 2 \sum_{k=1}^{K} O_k \ln \left(\frac{O_k}{E_k} \right) \sim 2N \; D_{kl}(O, E). \tag{1.34}$$

If the form of this statistic looks familiar, it is because the second order Taylor approximation of the G-statistic is a generalization of Pearson's chi-squared statistic, which was developed to circumvent the computation of log-likelihood ratios:

$$\chi^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}, \tag{1.35}$$

where $O_k = N f_O$ and $E_k = N f_E$ in the case probabilities are given. These statistics, which are distributed χ^2 with $K - 1$ degrees of freedom, may be used to test the null hypothesis of comparability.[25] In our case, the test statistic is therefore:

$$\delta = 2 \sum_{k=1}^{K} O_k \ln \left(\frac{O_k}{\hat{O}_k} \right). \tag{1.36}$$

[25] An alternative degrees of freedom estimate may be $K - K \cdot S(\theta_i, \theta_j) - 1$, in the case the researcher only includes distances for matched topics in the computation. This, however, will bias the test towards the null.

The test statistic δ is distributed $\chi^2(K-1)$, and therefore if $\Pr(\chi^2 \leq \delta) > 0.95$, the corpora are independent and are not comparable. Intuitively, this would happen because the aggregate of the response distances is greater than we would expect it to be if the topical distributions were similar. The G-statistic approach gives more conservative test statistics than the approximated χ^2 approach for a sufficiently large number of topics (in excess of 100). For $K < 100$ it is recommended to use the χ^2 flavor of the statistic (otherwise, the degrees of freedom should be adjusted to account for idiosyncratic differences). An example of how to use the previously stated steps to compute the statistic, using computer code, is available in the appendix.

Using our running example, we sum over the standardized residuals of the response distances to compute the delta-statistic, which is distributed χ^2 with degrees of freedom $K-1$:

$$\delta(\text{News}, \text{Congress}) = \sum_{k=1}^{K} \frac{(R_{kk'} - R_k^0)^2}{R_k^0} \tag{1.37}$$

If the statistic exceeds the critical value $\delta > \chi^2_{K-1}$, then we may conclude that the patterns in language in the two populations of New and Members of Congress are significantly different.

Stated plainly, even if the words used by Newspapers are the same as the words used in Congress, they are used in different ways, and may imply significantly different models of text for the two populations. This means that the assumption that we can

predict:

$$\hat{Y}_i = f(W_i), \tag{1.38}$$

is rejected in favor of two separate models:

$$\hat{Y}_i^L = f(W_i^L), \tag{1.39}$$

$$\hat{Y}_i^U = f(W_i^U), \tag{1.40}$$

where L and U indicate labelled (scored) and unlabelled (unscored) individuals on the variable of interest (in this case ideology). Therefore, we are in this case "back where we started," with two separate units of anlaysis from two naturally distinct populations—only this time, their distinction is demonstrable and meaningful.

It is worth noting that the delta-statistic method is related to a maximum likelihood approach that determines whether a pooled model outperforms a marginal model. For instance, the alternative approach might compare separate topic models to a jointly estimated topic model on the basis of their penalized log likelihoods or Akaike Information Criteria, and reject the joint model if a marginal model explains the data significantly better. Why is the delta-statistic method any better than this approach? The AIC and associated statistics are limited in that they only provide relative tests of model quality, and may not provide insight in absolute terms. The delta-statistic method is derived from a strong theory that allows for meaningful and absolute evaluations of model fit. Indeed, while the AIC will select a pooled model

for a minimal set of 100 words over another model using 100,000 if it fits better, the delta-statistic specifically considers the ability of the model to recover meaningful patterns in words. The purpose of this analysis is to answer the question of "what are we substantively doing when we do text analysis," rather than the question of "what model has less error?"

1.6.1 Verification by Monte Carlo Simulation

We now evaluate the delta-statistic by conducting formal Monte Carlo simulations. To do this, we first simulate a numerical corpus using the data generating process posited in the latent dirichlet allocation topic model, equation (1.6) (Blei, Ng, and Jordan 2003). Recall the notation from equation (1.6). The simulation first requires that the K topical distributions over words are generated from a symmetric Dirichlet with concentration parameter α:

$$\theta_i \sim \text{MN}(\alpha, \#W). \tag{1.41}$$

Expected topic mixes v_i, and realized words from the topics, are then drawn for each document W_i given Q_i and a sister symmetric concentration parameter:

$$v_i \sim \text{Softmax}(\text{MN}(\alpha', K)), \tag{1.42}$$

$$W_i \sim \text{MN}(Q_i, m_i), \tag{1.43}$$

where m_i is the document's length, generated around a central tendency μ:

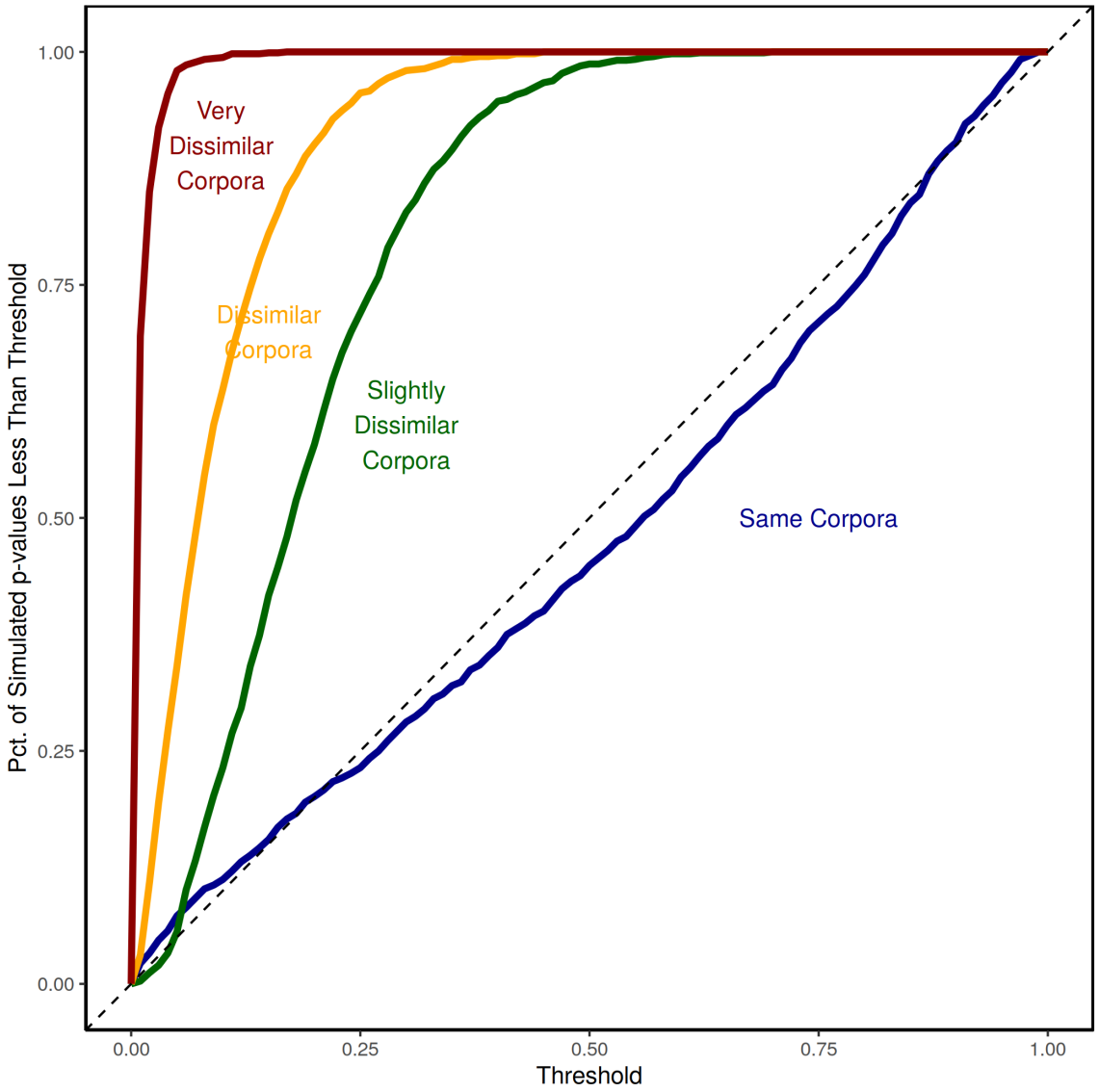
$$m_i \sim \text{Poisson}(\mu, N). \tag{1.44}$$

We randomly split this generated corpus into two equally sized subsets of $N/2$ to create two corpora generated by the same data generating process. We then draw 1,000 samples from this data generating process, run the delta-statistic on each each pair, and obtain a χ^2 p-value. Specifically, I run the test with topics $K = 30$, documents $N = 500$, unique words $\#W = 1000$, and average document length $\mu = 100$.

Since the corpora are generated using the same set of latent parameters, a formal indication of whether the delta-statistic works properly would be that the p-values are distributed uniformly. Figure 1.2 graphs the cumulative distribution of these results (plotting the percent of p-values less than each given value of the probability). The uniform distribution of p-values does indeed show up in the simulations and is reflected in the figure by the blue line closely approximating the 45 degree line.

We may also evaluate the delta-statistic further by introducing different levels of dissimilarity into corpora, and observing whether the test deviates as it should from uniform. We may introduce dissimilarity by varying the concentration parameters α and α'. I run the delta-statistic for $\alpha, \alpha' = \{0.4, 0.5, 1.0\}$ and report the results in figure 1.2. All dissimilar corpora simulations have p-values above the 45 degree line, indicating that the p-values are not uniformly distributed; indeed, as expected, the greater the dissimilarity, the more the results are skewed more toward lower, less uniform, p-values. Higher levels of dissimilarity have a higher p-value line because the

Figure 1.2: Monte Carlo Verification of the Delta-Statistic



Note: Figure reports the cumulative distribution of p-values from the delta-statistic for sufficiently similar corpora, and for increasingly dissimilar corpora. Areas under the curve (AUC) for the lines are 0.99, 0.91, 0.81, and 0.48, from left to right.

test has more power to detect these inter-corpora differences.

Figure 1.3: Overview of Text from Congressional Record and Newspaper Corpora Used in Gentzkow and Shapiro (2010)

(a) Congressional Record (2005) **(b)** Newspapers (2005)

Note: Panels are wordclouds generated by the two corpora used for the analysis in Gentzkow and Shapiro (2010). The first corpus, in Panel (a) is a set of Congressional speech documents from the Congressional Record in the year 2005. The documents are at the level of the Member. The second corpus, in Panel (b) is a corpus of newspaper articles from 433 U.S. newspapers in the year 2005 (the coverage of the newspapers is about 75 percent of the total U.S. readership. The documents are at the level of the newspaper. . The size of each word is proportional to its frequency in the corpus.

1.6.2 Empirical Validation of the Delta Statistic

We now turn to further validation of the methodology using real-world data instead of Monte Carlo simulations. In particular, consider two corpora used for the analysis in Gentzkow and Shapiro (2010). The first corpus is a set of Congressional speech documents from the Congressional Record in the year 2005. The documents are at the level of the Member. The second corpus is a corpus of newspaper articles from 433 U.S. newspapers in the year 2005 (the coverage of the newspapers is about 75 percent of the total U.S. readership. The documents are at the level of the newspaper. Figure 1.3 reports word clouds generated from the two corpora, to give a high-level overview of the language used in each. Words are weighted relative to their frequency

Table 1.1: Topic Summaries from Cong. Record and Newspaper Topic Models

Congressional Record (2005)				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
postalservice	illegalalien	lowincome	tradeagreement	africanamerican
postoffice	illegalimmigration	headstart	fretrade	civilright
committegovernmentre	bordersecurity	foodstamp	centralamerican	rosapark
Newspapers (2005)				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
taxrate	collegestudent	professionalsport	driverlicense	saddamhussein
communitydevelopment	businessowner	stpaul	gulfcoast	wariraq
affordablehousing	courtjudge	bottomline	hurricanekatrina	europeanunion

Note: Table reports unordered and unmatched topic clusters from the separate latent dirichlet allocation (LDA) models estimated on the Congressional Record and Newspaper corpora from 2005. The topics are then fed into the tl-LDA algorithm, which matches them against each other and determines if the match is valid. The results of the matching exercise are reported in table B.2.

in each corpus. These corpora are used for the validation exercises that follow.

If the data generating processes for congressional speech and newspaper text are similar, then the delta statistic should fail to reject the null hypothesis of comparability. As usual, we start with the ill-advised approach. The naive approach would be to compute the similarity between the two marginal corporal token distributions, $D_C(f_m, f_{m'}) = 0.576$, and treat the uncertainty of the estimate as we would a difference in means. Taking this approach, the analysis reveals that distributions for Gentzkow and Shapiro (2010) have a cosine similarity of 57.6 percent. This is, however, misleading. Words used more frequently as a function of the time period in which the congressional record speech and news were generated will upwardly bias the similarity estimate. We are not interested in the driving effect of time, or anything else spurious; we are interested in how ideology is linked to speech patterns in both corpora. We will now apply the tl-LDA approach just introduced to produce a test statistic for comparability.

Following the approach of section 1.5.2, consider two separate topic models run on the corpora, linked via tl-LDA, and tested using the delta statistic's difference from expected response distances. Table 1.1 reports the top terms associated with LDA-estimated topics for the Congressional Record (CR) and Newspaper corpora, before linking and validating. Table 1.2 reports the summaries for five of the matched topics after the tl-LDA procedure was completed, along with the estimated distributional differences and their δ-statistic contributions. The table shows how the δ-statistic can be computed simply by summing up the χ^2 style contributions. The model outputs and intermediate quantities are reported in table B.2. The test statistic for the comparison of the corpora is 82.46, and the χ_2 critical value for $df = K - 1 = 29$ is 42.6. The test statistic exceeds the critical value, and therefore we reject the null hypothesis of comparability. Substantively, this suggests fewer language patterns are shared across corpora than we would expect, had the corpora been generated by the same lingual process. Logistically, this means that the corpora are not comparable.

Further analysis of table B.2 reveals two insights. First, the table generally does not have a high number of matches over the number of topics. While some topics are clear matches – and therefore, estimated to be shared traits among the texts – the degree to which the texts share comparable patterns in text, as operationalized by their probability distributions, is questionable. In other words, it might appear that the topics share similar words, but the ways those words are used are very different across corpora. Such findings are consistent with confounded results.

Table 1.2: Calculation of Delta Statistic for Cong. Record and Newspaper Models

Congressional Record (2005)	Newspapers (2005)	Distance (R)	δ-Contribution $\left(\frac{(R-D^0)^2}{R^0}\right)$
Topic "African American"			
africanamerican	africanamerican	0.45	1.38
tuskeegeeairmen	civilright		
blackcaucus	lowincome		
Topic "Natural Disasters"			
hurricanekatrina	hurricankatrina	0.04	0.30
gulfcoast	federalemergencymanagement		
naturaldisaster	littlerock		
Topic "Environment"			
naturalresource	naturalga	0.46	0.03
climatechange	wildliferescue		
foreignoil	naturalresource		
Topic "War on Terror"			
warterror	nursinghome	0.61	7.4
globalwar	collegestudent		
globalwarterror	nationalguard		
...	...	.	.
Topic "Stem Cell Research"			
stemcell	courtjudge	1.58	28.49
cellresearch	seniorcitizen		
embryonicstem	courtappeal		
		$\sum_i \delta_i =$	82.46 ($\chi^2_{cv} = 42.6$) ($p < 0.01$)

Note: Table reports five matched topic clusters, using the tl-LDA method, from the separate latent dirichlet allocation (LDA) models estimated on the Congressional Record and Newspaper corpora from 2005. The topics were fed into the tl-LDA algorithm, which matched them against each other and determined if the match is valid. The response distance (R) is the difference between the probability distributions over tokens for the topics. The δ contribution is the square of the difference between observed difference and the expected distance under the null hypothesis that they were drawn from the same data generating process, divided by expected difference. At the bottom of the table, I tabulate the sum of the δ contributions to show how the δ-statistic is computed. Results are truncated for presentation. The full results of the matching exercise are reported in table B.2. The matched topics were on average not high quality matches (low distance), which resulted in a rejection of the null hypothesis of comparability.

Application of Delta Statistic to Same Corpus

The test statistic rejected comparability for the comparison of the congressional record speech corpus to the news corpus used in Gentzkow and Shapiro (2010). Does the statistic reject comparability if we run for each corpus against itself? To execute this analysis, I first split each corpus into two randomly selected subsets and treated them as two separate corpora. I then ran tl-LDA, treating the two as if they were separate corpora.

Following the approach of section 1.5.2, consider two separate topic models run on the corpora, linked via tl-LDA, and tested using the delta statistic's difference from expected response distances. The model outputs and intermediate quantities are reported in table B.3 and table B.4. For this draw of the subsamples, the test statistic for the comparison of the congressional record corpus is 38.1, and the critical value for $df = K - 1 = 29$ is 42.6. The test statistic does not exceed the critical value, and therefore we fail to reject the null hypothesis of comparability. The test statistic for the news corpus is 15.11, with the same critical value, and therefore we fail to reject the null hypothesis of comparability. Substantively, this suggests more language patterns are shared across these "corpora" than we would expect by random (indeed, they are the same corpus). In other words, the test statistic successfully provides evidence that each corpus is comparable to itself.

Table 1.3: Summary of Results

Query Corpus	Reference Corpus	Method	Test Statistic	Reject?	Conclusion
Cong. Record	Cong. Rec.	Subsampling	38.17	No	Compare
News	News	Subsampling	15.11	No	Compare
Cong. Record	**News**	**Subsampling**	**82.46**	**Yes**	**Don't Compare**

Note: The table reports estimated δ-statistics for the comparison of the Gentzkow and Shapiro (2010) corpora to each other and themselves. The results suggest that the statistical test is able to recover inferences consistent with our priors.

Summary of Validation Results

Table 2 summarizes the results reported above. The results suggest that the test statistic is able to detect when corpora are comparable under the ground truth use case, where the corpora are being compared to themselves. The statistic does not allow for the comparison of the congressional record and the news, but it does allow for the comparability of the congressional record and the news to themselves. This makes intuitive sense. Though the models used to detect language patterns vary, the patterns detected show up repeatedly when run on the same corpus.

1.6.3 Limitations of the Test Statistic

The delta statistic has several limitations. The first of these limitations originates from the assumption of independence that is required to invoke the proper distribution for the test statistic: it assumes that estimated topics are drawn independently of each other. This may not always be the case. For instance, correlated topic models and other topic models that include variable information when estimating topics (such as structural topic models) explicitly allow topics to be related to each other by way of the variables used to estimate them. In such cases, the statistic is less efficient,

because the contribution of any "group" of correlated topics results in a degrees of freedom that is less than the one used by default. An appropriate solution is to decrease the estimated degrees of freedom used to conduct the test.

Further to that point, it is noteworthy that LDA often doesn't yield unique topics. An obvious drawback of the greedy matching method is that it downweights the estimated comparability of two corpora when any one of the corpora is more likely to produce non-unique topics. This is perhaps an issue of princple, but it has a nice benefit to it. Its effect is to decrease the probability that we find a corpus to be comparable when many of its topics are essentially the same. The net result of the alternative would be an overindexing to a singular topic between the corpora. This technically satisfies the first bridge criteria.

Second, the statistic depends on the simulation of nulls using the MLE gamma parameters from the reference topic. The null response distances the procedure simulates is directly implied by the nature of the simulation. For this reason, even the test itself may be subject to sensitivity to researcher decisions if it is ever changed.

Third, the method is directly sensitive to hyperpaprameter specification and other researcher-based pre-processing decisions. Researcher decisions over hyperparameter specification may include the number of topics, or priors over α. Pre-processing decisions may include choices over the support to be included (inter-sectional feature selection, outcome-based feature selection, and Winsorization), transformation over those support values (regression weights, tf-idf), and lexical analysis (feature engineering), to name a few (let alone document standardization and unit of analysis specification). Indeed, as Denny and Spirling (2018) demonstrate, topic models are

quite unstable conditional on pre-processing decisions.

The solution this dissertation offers follows in the next section. With respect to the delta statistic, this last limitation is perhaps less problematic than the others. The statistic will be valid so long as the assumption of the mapping function, which was introduced in section 1.1, is satisfied. Section 1.7 develops further the sensitivity analysis proposed to satisfy the third bridge criterion.

1.7 Sensitivity Analysis and Multimodel Inference

The third bridge criterion requires that the measurement instrument constructed – via a textual "codebook" – recovers in all cases, and comparably across corpora, the same latent dimension. This bridge criterion is perhaps the most difficult to test. It is a given that if the mapping function $m_i(\cdot)$, discussed in section 1.1, is not applied consistently across documents (or corpora), there is immediately a violation of the bridge criterion. Given that the obvious things are accounted for, however, it is due to the power of the researcher's argument that the instrument does what it says it does. It is doubly difficult because in most cases the ground truth is also unobservable.

Research is in general too variable and context-specific to provide field-wide edicts on how to pre-process one's data. The solution for which this dissertation advocates is similar to the Denny and Spirling (2018) approach of simulating outcomes under a variety of specifications. For instance, the solution promulgated for hyperparameter

selection is to conduct a grid optimization over several hyperparameters, selecting the one which maximizes a fit statistic of some sort, such as perplexity (Mimno et al. 2011). The solution offered for sensitivity analysis is to conduct analyses across the spectrum of known pre-processing methods – each of which, as detailed in this dissertation, has its own null expectation of how it will affect the result – and then create bootstrapped confidence intervals for any effect by simulating over all possible modeling decisions.

The approach this dissertation takes to satisfying the third bridge criteria is encouraging the researcher to report empirical confidence intervals based on the many different research choices she could have made. The innovation the dissertation introduces is to show that certain types of choices, such as the tf-idf transformation to re-weight cell values, have predictable effects on downstream results.

1.7.1 Reducing Sensitivity to Researcher Specification

There are two general sources of sensitivity in the text analyses. The first source is model and hyperparameter selection. Hyperparameters such as the number of topics estimated during a dimension reduction procedure, the restrictiveness of regularized regression parameters, or the aggressiveness of the Winsorization procedure, can affect the variability and quality of the matrix $\tilde{C}$ used on the right hand side of downstream equations. Topic modeling in particular is notoriously difficult to objectively validate; the ease of interpretability among a community of peer reviewing researchers has been what has propelled its popularity. I experiment in this chapter with several methods of topic model validation, including automated topic summarization.

The second source is pre-processing. Pre-processing decisions directly affect the support of any analysis – the features available to use in it – and the quality of the matrix $\tilde{C}$ used in downstream analysis. One of the greatest pre-processing decisions the researcher makes in the comparison of political texts is the support basis on which the researcher will make inter-corpus comparisons. Should the researcher limit the analysis on the basis of shared features? Or, should the researcher include all features in the analysis? In addition to the pre-processing decisions involved in lexical analysis, which are discussed in detail in appendix B.5, such decisions can greatly affect the variability and quality of $\tilde{C}$. Transformations such as the tf-idf, least squares weighting, and other changes to matrix values can also greatly affect $\tilde{C}$.

The section proceeds as follows. First, it refreshes the notation. Second, it delves into proposed methods of hyperparameter selection. Third, it expands on the proposed methods of pre-processing. Finally, it demonstrates how to combine the variable estimates in order to produce an ensemble confidence interval.

1.7.2 Notation for the Bounding Method

Consider again set of text features W_{nj} contained in a document n from corpus j. The first goal of this exercise to identify a function $m : W \rightarrow \tilde{C}\ \forall j$, where C is the lower-dimensional set of features that vary meaningfully with the outcome or quantity of interest. In the parlance of section 1.1, g is the "codebook" that contains the instructions humans would use to map from texts to a meaningful latent

representation.[26] The specification of $\tilde{C}$ is important because it can affect the size and significance of downstream estimates. The question of interest in this section is how the properties of estimates derived under different $\tilde{C}$ affect the magnitude and precision of ultimate effect estimates, $\hat{\beta}$.

1.7.3 Sensitivity Owing to Hyperparameter Specification

First, consider sensitivity owing to hyperparameter specification. This dissertation offers an approach that optimizes a fit metric, called the UMass Coherence, by simulating outcomes across a grid of hyperparameter values and analyzing which fit metric achieves an optimum. In particular, it considers cases in which a topic model is used to produce $\tilde{C}$, which means that the researcher must specify the number of topics, and appropriately clean the data using a *Winsorization* procedure to ensure smooth and coherent topics. This section uses as its exemplar data a corpus of State of the State speeches, which are used further in the substantive papers of this dissertation.

Choosing the Number of Topics

Generally, the political science literature advocates a philosophy of thoughtful, science-based, *a priori* modeling (Anderson and Burnham 2004). One first develops a global model, and then derives several other plausible candidate models fit to the data at hand. This is incredibly difficult to do for text analyses, and is in fact not the advised best practice (Grimmer and Stewart 2013, *e.g.*), because in many cases there are no

[26] As a reminder, "codebook" is a content analysis handbook. See Krippendorff (2012) or Neuendorf (2016) for detail. Whereas a codebook is written by the researcher to map from texts to a latent dimension, g is a machine-learned set of instructions to do so.

modeling priors on which to rely. It is too risky for a scholar to stake her claim on a single model that might not produce interpretable results, or which will later be shown to be inadequate. Moreover, the computing tens, let alone hundreds of models, can be prohibitively slow. Usually, the methods used in the published research using text analysis involve a subjective determination of the ideal number of topics, and with only one reported pre-processing protocol.

In contrast to the usual practices, I produce several hundred candidate topic models and select the one with the best held-out fit. For each tuple in a grid of progressively increasing Winsorization lower bounds (Ruppert 2004) and topic number parameters (k), I fit an LDA to the text data and compute a held-out fit metric, where the Winsorization lower bounds are:

$$\mathbf{B} = \{0.25,\ 0.30, \ldots,\ 0.60\},$$

and the topic numbers are:

$$\mathbf{K} = \{20,\ 40,\ \ldots, 120\}.$$

Winsorization is a process by which extreme values are removed from a dataset, to reduce the effect of possibly spurious outliers. The Winsorization bound is the distributional threshold beyond which I exclude values. The method works by creating a rank-ordered empirical cumulative density function from the feature frequencies, and then trimming away the features which fall below the lower bound. The method

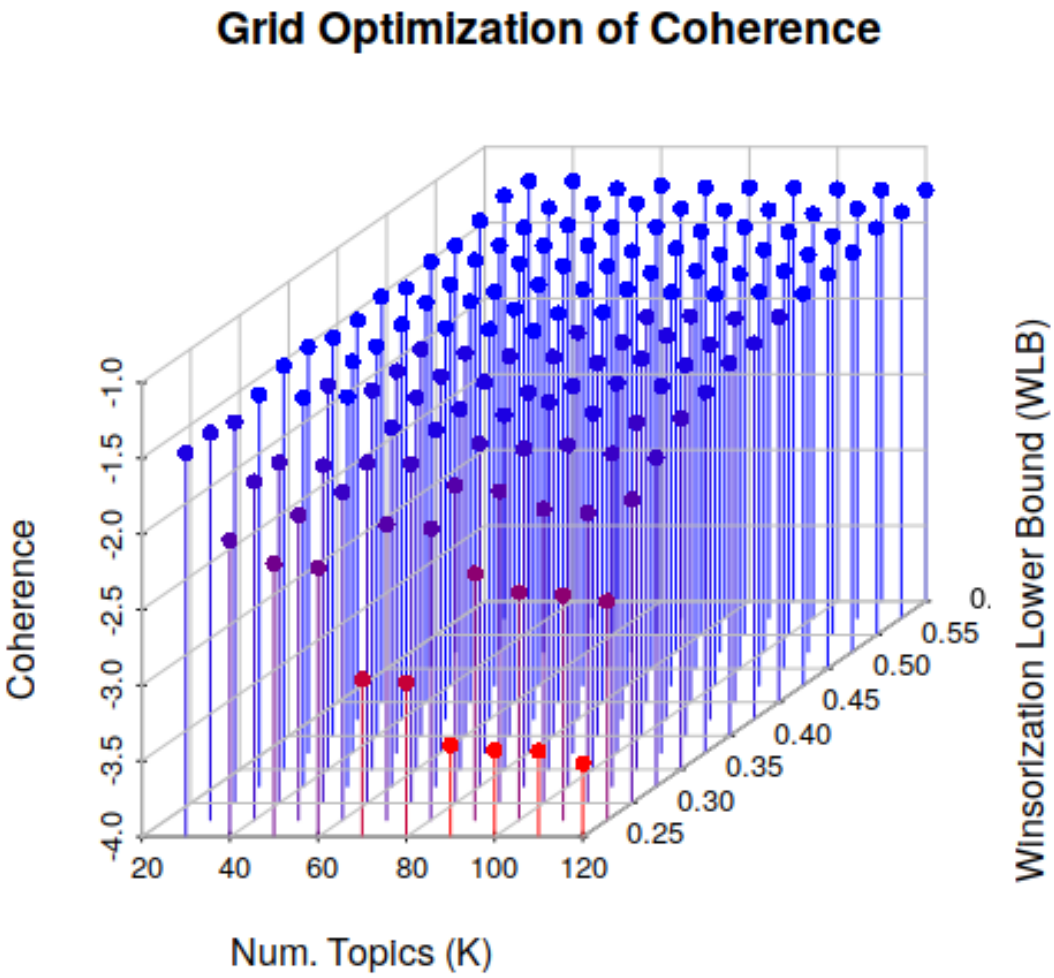
is also known as using "trimming bounds" in the political science literature (Manski 1990, *e.g.*). The application of Winsorization to topic modeling yields topics which are less over-fit: the topics produced are less sensitive to idiosyncratic language used only in the corpus to which the model is fit. The effect of Winsorization is to produce more generalizable topic models, which may be used to compare models estimated on different corpora.

Taddy (2012) summarizes three primary methods with which the researcher may learn the appropriate number of topics from the data: cross-validation, non-parametric mixture priors, and marginal likelihood.[27] Although cross-validation is by far the most common choice (Grimmer and Stewart 2013; Grimmer 2010; Hornik and Grün 2011, *e.g.*), it is not a scalable solution because it requires costly repetitions of model-fitting; it also lacks easy interpretability in terms of statistical evidence, including sample error (Hastie, Tibshirani, and Friedman 2009, 7.12). Some work, however, has attempted to clarify the interpretation through the lens of causal inference (Egami et al. 2018). Teh et al. (2005) provide a model-based alternative, which treats each document as a vector of weights over an infinite number of topics.[28] The standard Bayesian solution is to maximize the marginal model posterior. Proper marginal likelihood estimation, however, is still limited to approximation because of its computational complexity (Griffiths and Steyvers 2004). Mimno and Lee (2014) offer an alternative methodology for the estimation of the number of topics, which sets "warm" topic

[27] See also Airoldi et al. (2014) for a short survey and comparison.

[28] The downside of this approach is that it is quite sensitive to the priors set (very sparse priors can effectively result in a pre-ordained number of topics anyway), and the method necessarily entails very costly inference about high-dimensional term-topic memberships.

Figure 1.4: Optimization Surface for Topic Model



Note: Figure reports the coherence grid for topic models estimated when fitting models to both the SOTS and SOTU corpora. Each value is the UMass statistic Mimno et al. (2011) computed for a model at a particular winsorization bound and topic number. The grid assists with the selection of model hyperparameters—particularly, the number of topics to be used in the final, selected model. The procedure provides an algorithm with which the researcher may select model hyperparameters.

priors based on latent embeddings of tokens within the corpus (effectively running principal components analysis, or PCA, to optimize K independently of the topic model, and then feeding the output of PCA to the model).

The number of topics is also often a topic of great consternation. Ample research suggests that the number of topics selected can significantly affect the results of LDA, which, in turn, affects the predictions the model makes. In the present case, instability in predictions is equivalent to instability in the predicted policy agenda. Even when the number of topics is fixed, the resulting topics can be conditional on the initialization strategy or random seed (Roberts, Stewart, and Tingley 2014; Arora et al. 2013). This problem is not specific to LDA, and has been considered extensively with respect to several clustering techniques (Lange et al. 2004, *e.g.*). One solution to this is to repeatedly fit models until the results converge on a general trend. I integrate this into my strategy, which also maps across several different Winsorization lower bounds.

More generally, there are two types of topic model instability. The first is instability of replication. Even when the hyperparameters such as the number of topics or the expected concentration parameter, are unchanged, topic models can estimate topical distributions in one run that may not be estimated again in another run. This is problematic because the conclusions of analysis like the present on are drawn based on topic distributions. This type of instability is not usually seen in deterministic methods such as latent semantic analysis, or LSA, which solve for a maximum likelihood estimate over all known, observed data. The second is instability over the number of topics. Common sense would suggest that as the number of topics increases, the

fit of the model should also increase, simply because there are more parameters over which to fit the model to the data. This makes model selection more of a balancing act between the fit to the observed data and generalizability to unobserved data. In reality, however, many times, the fit of the model may not change monotonically as the number of topics increases.

We use the UMass coherence measure (Mimno et al. 2011), named "UMass" as homage to the institution at which it was developed to evaluate the fit of the topic model. I define topic coherence as

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \left(\frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \right), \tag{1.45}$$

were $D(v)$ is the document frequency of word type v (*i.e.* the number of document with at least one token of type v), $D(v, v')$ is the co-document frequency of word types v and v' (*i.e.* the number of documents containing one or more tokens of type v and at least one token of type v'), and $V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$ is the list of the M most probable words in topic t. Coherence scores topic models on the basis of their quality of their semantic clusters. The score corresponds well with human coherence judgments and makes it possible to identify semantic problems in topic models without human evaluations or external reference corpora.[29]

[29]The fit statistic I employ here differs from pure information-theoretic approaches. These approaches, such as the one taken in Taddy (2012), which applies the logged Bayes factor as the quantity of interest, consider the ability of the model to make accurate predictions on held-out data. The method of coherence adheres to this principle, but intends to construct a measure which also holds up to human validation. I use it here *because of* its balanced focus on both fit and validity. Other methods, such as perplexity (Asuncion et al. 2009), also attempt to optimize validity from a theoretical point of view, but have been shown to optimize for topics which make little sense when subjected to human validation. Moreover, the results are consistent with other approaches which avoid "word intrusion," which is the probability that a word that should not be present in the

The model which optimizes the fit metric – $K = 120$, $B = 0.25$ – is the model I use for prediction. Figure 1.4 visualizes the candidate model fits, demonstrating how the use of the metric lends itself to a smooth distribution which may be searched. The smoothness of the fits is encouraging, because it suggests any observed minimum or maximum has true meaning as a model of best fit, instead of model which randomly achieved the best possible fit (which would lead to the selection of a model which had a great fit simply by chance).

Sensitivity Owing to Researcher Presentation of Results

To evaluate the performance of topic models on their face value, I use automated topical summaries. To further extend the example of the State of the States corpus, introduced more fully in chapters 3 and 4, I compute automated topical summaries using a range of methodologies, specifically the metrics phi (Blei, Ng, and Jordan 2003), lift (Taddy 2012), relevance (Sievert and Shirley 2014), and FREX (Bischof and Airoldi 2012):

$$\phi(\text{Phi}) = \prod_{i=1}^{K} \frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\Pi_{r=1}^{V} \Gamma\left(\beta_r\right)} \frac{\Pi_{r=1}^{V} \Gamma\left(n_{(\cdot),r}^{i} + \beta_r\right)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)}, \tag{1.46}$$

$$\text{Lift}_{ik} = \frac{\phi_{ik}}{\sum_{i=1}^{n} w_{ik} / \sum_{i=1}^{n} w_i}, \tag{1.47}$$

$$\text{Relevance}_{ik} = \lambda \log(\phi_{ik}) + (1 - \lambda) \log\left(\frac{\phi_{ik}}{\sum_{i=1} w_{ik}}\right), \tag{1.48}$$

$$\text{FREX}_{ik} = \left(\frac{w_{ik}}{F_{ik}} + \frac{1 - w_{ik}}{E_{ik}}\right)^{-1}, \tag{1.49}$$

subjective, exogenous topic is included in the model's estimation of the topic.

where w_i is the corpus frequency of the token i. For equation (1.46), phi is a matrix of size k by r, where k indexes topics and r indexes words. $n^i_{j,r}$ is the frequency of token r in the jth document which are assigned to the ith topic. β_r is the prior weight over token r (usually small, to promote sparsity). In equation (1.49), F is the frequency score given by the empirical cumulative density function of the word within in its topic distribution (ϕ_k). Exclusivity, denoted E, is computed in two steps: first, I column-normalize the topical phi matrix to produce the conditional probability of seeing the topic, given the word; second, I transform that matrix by taking the empirical cumulative density function of each row (each topic). High values in E suggest that the word is more frequently associated with the topic.

Phi is simply the estimated conditional probabability of the token, given the topic. It is drawn directly from the values produced by LDA, which estimates the probabilities generatively—the values are the probabilities which optimize the model likelihood.

Lift is a transformation of phi. To compute lift, I divide phi by the empirical term probability. Larger values of lift suggest that the probability of the token is greater given the topic than it is given the corpus—*i.e.*, that the word is more likely to appear if we know the topic, than if we were to simply look at its frequency in general.

Relevance is a weighted sum of the probability of the token within the topic and lift. Setting $\lambda = 1$ ranks topics solely by their estimated probability, while $lambda = 0$ ranks topics solely by their lift. Values $\lambda \in (0, 1)$ allow the researcher to adjust the influence of either measure to determine relevance. Relevance is similar to FREX, which is discussed next.

FREX (frequency–exclusivity) attempts to find words which are both frequent in and exclusive to a topic of interest. Bischof and Airoldi (2012) note that "balancing these two traits is important as frequent words are often by themselves simply functional words necessary to discuss any topic, while completely exclusive words can be so rare as to not be informative." In this sense, FREX is similar in concept to the term frequency–inverse document frequency score (tf-idf score), a well-established metric in the natural language processing literature.

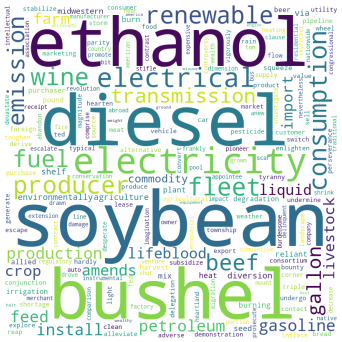
Automated topical summaries provide a principled way with which to label topics, whereas subjective labeling can result in the introduction of diversity of researcher bias. For example, researcher bias can be conditional on the party examining the topics; an economist may see the topic "particular, tax_credit, tax_refund, group" and label it "tax incentives," while a political scientist may see the same topic and label it "interest groups".[30] Although no method is perfect, automated topical summaries help to circumvent issues associated with subjectivity. Presenting a multitude of automated topical summaries, based on different factors, also helps to reduce labeling risk.

Figure 1.5 reports automated summaries of nine topics visually, in the form of wordclouds. The topics were extracted using the optimum model from the method just described. The size of each word is proportion to the weight assigned by the automated summary method. Each wordcloud contains the top 200 phrases in the topic. The method used to determine top phrases for the wordclouds is the lift method. To determine the top phrases, I rank the words within each topic by their lift score,

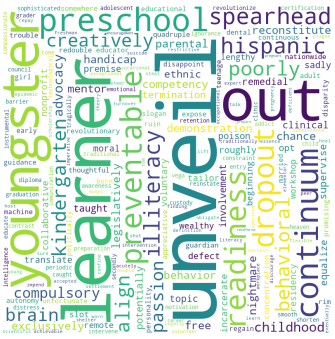
[30]Researcher bias may also result when a researcher labels several topics at once; because of the recency of a label that has already been assigned, the researcher may be less likely to assign the label again, when in reality the it should be.

Figure 1.5: Wordclouds from Selected Topics

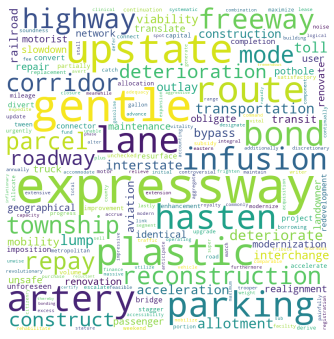
Topic 1



Topic 9

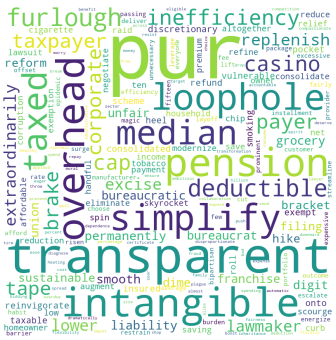


Topic 62



Topic 16

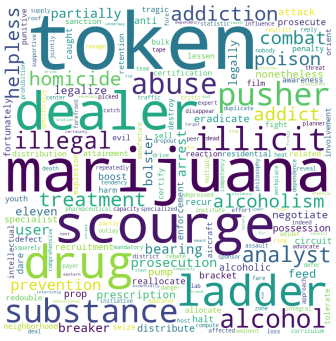
Topic 5



Topic 6

Topic 65

Topic 74



Topic 77

Note: These are the automated summaries of nine topics. The topics are estimated using the optimum model from the methodology described *supra*. Each wordcloud contains the top 200 phrases (based on their lift score) in the topic. Words are scaled relative to their lift score (Taddy 2012).

and then take the top 200.

For example, consider figure 1.5, topic 1. Topic 1 assigns high scores to words such as *hazardous, recycle, toxic, pollute, environ, and aquifer*. Such words would refer to discussion of the environment. Many governors speak about the environment and environmental reform in their State of the State addresses, and it is a common political agenda issue. Consider further topic 5. Topic 5 assigns high scores to words such as *tort, fault, severity, justifiable, and punitive*. Such words would refer to discussion of torts and tort reform, which, again, is a topic of considerable popularity among state governors; in fact, it is a legal area reserved especially for the states, and it is a common political agenda issue.

Table B.5 reports automated summaries of each topic. Unlike the wordclouds, the table reports automated summaries based on all four methods. Each table entry contains the top ten phrases from the topic, as estimated by each method. The comparison is useful because it demonstrates how each method produces a slightly different representation of the topic. It is important to report all of the summaries to reduce bias from the researcher and subject the researcher to validation from reviewers, as discussed earlier; however, it is also important to report all of the metrics to highlight which topics have automated summaries that differ significantly. If an automated summary seems to present information that is markedly different from the information the others offer, the researcher should exercise caution in making inferences. The data reported in creftab:topics does not suggest any reason for concern.

For example, consider table B.5, topic 1. Topic 1's lift method summary assigns high scores to words such as *hazardous, recycle, toxic, pollute, environ, and aquifer*.

However, topic 1's relevance and phi methods include other words, such as *solid, site, and waste*. This demonstrates how the different methods produce different results: waste, for example, is a word that would be included in other topics that focus on wasteful spending, or wasteful social behaviors, and therefore, a method that penalizes words that are used in other topics – such as lift – would not include the word in its high-rank set.

Table B.6 reports automated codes for a random sample of 10 text chunks from the State of the States corpus. Text chunks are in their "normalized" form, which the model must consume to produce predictions consistent with the training data.[31] The codes are generated by the selected model as "predictions" of the topical content of the text. Near each topical code, I have indicated a "quick and dirty" label for the topic, to indicate the topic to which the machine believed the text belonged. We also include a column which estimates the rate at which the text includes the topic—*i.e.*, the percentage of the text which is devoted to the topic. These predictions – the rates at which each State of the State and State of the Union document includes the topics – are ultimately the values I use for the similarity analysis. These values may be considered as if they were the number of words in the chunk devoted to a certain topic of speech. For example, a 100-word document which is indicated to be 25 percent focused on topic 3 might have 25 words which belong to topic 3.[32]

[31]The principle is the same for linear regression; the researcher should not use a log-standardized transformation of income as an input to a pretrained model if the model was trained on the variable without the log. Now, whether the model should have included log-standardized data to begin with is a different issue (it should have).

[32]This is, of course, a simplification; words might belong to several topics, and so the computation will downweight the contribution of any word which belongs to multiple topics which estimating the topical proportion for a given passage of text.

For instance, consider table B.6, row 1. The chunk of text is from the State of the State speech given by Governor John Patterson from Alabama in the year 1963. The chunk of text appears to be in general discussing the construction of a waterway dock to encourage trade, so that the state may remain modern and competitive. The chunk is estimated to discuss topic 42, which has an automated topic summary of "development, economic, industrial, area, trade." The topic appears to be focused on economic development of industrial areas and trade. This checks out on its face. The topic is estimated to be discussed at a rate of 27 percent. This actually appears to the reader as if it is an underestimate; however, it could be that the speaker also discussed funding options and specifically focuses on the waterways as part of the economic development initiative. This underestimation perhaps indicates how the automated content analysis method I use can pick up on the general sense in which the speaker espouses an agenda item, without the need to account for the exact context in which she applies it; specifically, in this case, the speaker might use consider economic development as an important part of her agenda, and apply it to a waterways development project.

One of the most exciting things about the method of analysis is that it enables the quick indexing and access of policy documents. We conducted a small secondary source historical analysis to further investigate the ability of the method to discovery policy facets. Not being an expert on the history of Alabama, I conducted a quick search on John Patterson and found an encyclopedia entry for him, which discusses his critical role in improvements to Alabama's docks and waterways (Civil Rights Digital Library 2019). The topic to which the model attributed most of his speech in that

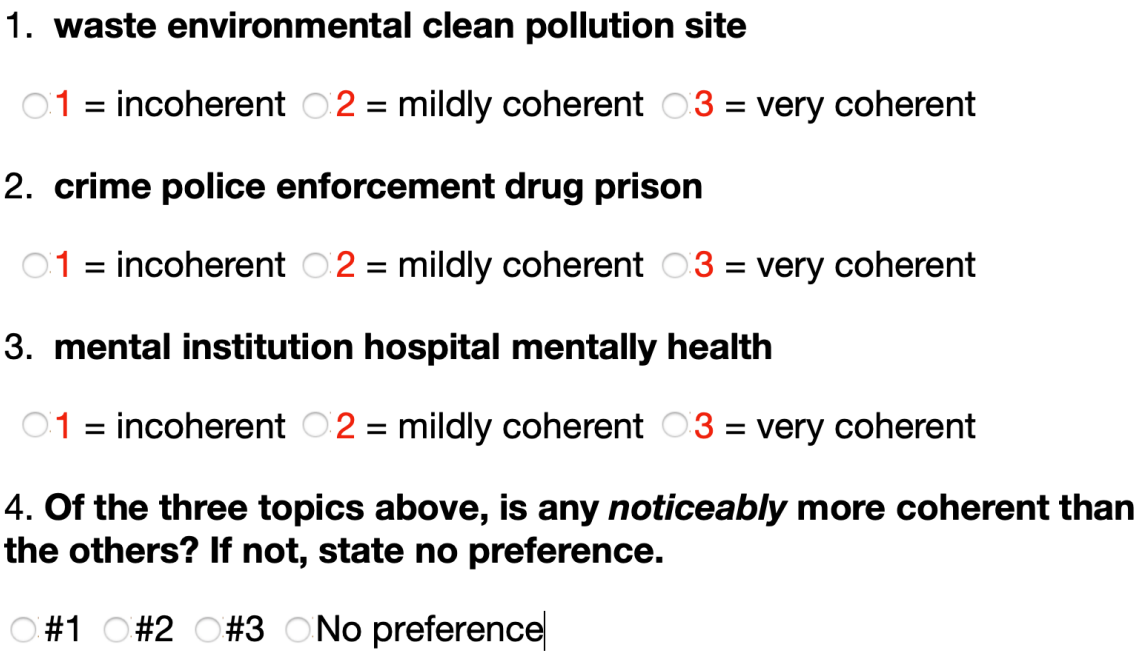
speech passage was exactly that – the topic of economic development, in the context of the docks! I imagine the method of automated content analysis I employ here may be extended to consider idiosyncratic policy elements of more general political agenda items (for instance, "economic development" would be a general political agenda item, while "of the waterways" would be the specific context).

The individual qualitative assessment of the topics derived from the automated content analysis is incredibly useful as a tool to engage with and understand the inner workings of the process. It is also a "quick and dirty" means of validating the ability of the model to produce predictions and inferences to which we attribute confidence. Still, however, individual qualitative assessment is no substitute for a principled approach by several coders working together to achieve a sense of the reliability of the prediction to recover a valid measurement of the latent variable (Krippendorff 2012). In the next section, let us consider the validity of the topic model and its predictions using such an approach.

Empirical Validation of Topic Model Results by Human Coders

Often, researchers report a smaller subset of topics in the body of their papers. They usually choose the most prevalent topics, or cherry pick the topics on which they construct tables and figures (word clouds); indeed, this is what I do in the previous section. We – and most scholars – do this because space prohibits the simple communication, in that form, of the information from hundreds of topics. Empirical validations such as the ones that follow are important because they present a fuller, simplified view of the performance of the model estimated. Moreover, such validations

Figure 1.6: Survey Instrument for Scoring Coherence with Human Coders



Note: Figure is a screenshot from the survey instrument used to measure perceived topical coherence by research assistants. Research assistants were asked to score the topics as coherent or incoherent, on the basis of their top words. Top words were determined by computing "lift" scores for each word and then ranking the words in descending order by their lift score (Taddy 2012). Training instructions were given to each research assistant to help them determine what a coherent topic looks like. Research assistants were also asked to provide a coherence ranking for each trio of topical summaries.

discourage any obfuscation or bias that might result from cherry-picking topics. To validate the quality of the model and the topics we estimate from it, let's examine the hand-coded coherence of the model's estimated topics.

We consider the coherence of the estimated topics. The validation design follows Airoldi (2014). I present research assistants with several tasks. In each task, I present a random sample of three topical summaries, from the table of topical summaries (120 topics $\times$ 4 metrics = 480 summaries) to code the coherence of each summary, using the top 5 words produced by the summarization metric. They had the option to code each summary as incoherent, mildly coherent, or very coherent. They also ranked the three

summaries as more or less coherent than each other. Three research assistants were assigned to each task, which allows for the computation of an inter-coder reliability coefficient. The results suggest that the topics are very coherent (72.1 percent coherent; Krippendorff's $\alpha = 0.88$, SD $= 0.12$). See figure 1.6 for a screenshot of the coherence task.

The results of the empirical validation effort can suggest that the method of automated content analysis one employs is effective. While it is not perfect – and perhaps no method is – it provides a good measure of the topics covered in the addresses.

1.8 Discussion

In the appendix, I propose an analytical typology of text analyses, in order to target for the reader the types of analyses for which satisfaction of the bridge criteria are necessary. The typology divides text analyses into two types. Type I analyses test for relationships between a corpus of text and a variable of interest. Type II analyses predict labels for unlabeled sets of documents. It is the second type which invokes the bridge criteria; it is the second type which necessitates testing for the criteria.

I demonstrated in this essay that a bias parameter exists in Type II analyses, which involve the estimation of two or more codebooks on two or more sets of texts. This bias is fundamentally untestable because Type II analyses do not have observed label values for all documents. The theory developed in section 1.1 clears the way for another estimator which covaries with the difference between the codebooks, but which

does not require outcome data to be estimated. The estimator compares independently estimated topic model parameters, and is called the δ-estimator. A test statistic based on the estimator (the δ-statistic) follows the well-known χ^2-distribution and may be employed to check if two corpora should not be compared.

To test for the satisfaction of the third bridge criteria, I propose a non-parametric method for estimating sensitivity in text analysis. The method is important because, as Denny and Spirling (2018) show, the outcomes of text-based analyses are highly conditional on researcher decisionmaking. I demonstrate mathematically and empirically that it is possible to generate lower and upper bounds for downstream effect estimates, thereby quantifying the sensitivity of an analysis to researcher-induced bias. The bounding procedure produces non-parametric estimates of uncertainty (confidence intervals) using an ensemble approach to multi-model inference.

The next two chapters now conduct empirical research in American politics. They apply the delta-statistic and the nonparametric bounding method to show that the State of the Union addresses are comparable to the State of the State addresses, annual gubernatorial addresses which, like their presidential counterparts, lay out the agendas of political actors in the state theater; it also shows that the State of the State addresses are comparable to the legislation passed by state Congresses. On the basis of this validity, the analyses proceed, testing theories related to nationalization and divided government.

Chapter 2

Have State Policy Agendas Become More Nationalized?

In recent decades, national factors (such as the president's popularity) have become increasingly predictive of election outcomes in both congressional elections and state elections. In this chapter, I examine whether the nationalization of state policy agendas is a factor related to the nationalization of state-level elections. I do so by collecting and analyzing the governors' State of the State (SOTS) addresses from the 1960 to 2016, and then comparing them to the presidential State of the Union (SOTU) addresses in concurrent years. I study these addresses because they are where platform agendas are laid out by the governor and the president. My analysis of the SOTS address shows that state agendas have become more similar to each other over time. It also shows that state agendas are more similar to the national agenda (as laid out in the SOTU address), and that the similarity between the state and national agendas predicts the degree to which national factors are influencing state level elections. Overall, the evidence suggests that state policy agendas have nationalized, and are a significant part of the nationalization story.

2.1 All Politics is. . . National?

The conventional wisdom states that "all politics is local." This adage, famously associated with former House Speaker Tip O'Neill, refers to the idea that voters care about local issues, and for that reason, politicians focus on local concerns rather than national issues. The management of such local concerns have historically been the purview of local party organizations, which were so effective at marshaling political support on the basis of local problems that Cotter et al. (1989, page 41) concluded, "what happens in the 3,600 county [party committees] determines the politics of states and the nation."[1] Despite the considerable focus on the power of locality, recent trends in elections have led many researchers to conclude that all politics is national (Burden and Wichowsky 2010; Abramowitz and Webster 2016). The geographical political distinctiveness scholars like Dahl (1961) and Key (1950) encountered a half-century ago has all but disappeared.

Substantial evidence suggests that congressional and state-level elections have become more nationalized with congressional election results now moving in near lock-step with those of presidential elections (Fiorina 2017). The enduring incumbency advantage that protected members of congress for nearly three decades has attenuated (Jacobson 2015). Further, national elections also strongly predict state-level elections. The relationship between midterm gubernatorial election results and the preceding presidential vote share is the strongest it has ever been (Hopkins 2018); and in the past decade, presidential approval has become a stronger predictor of state legislative

[1] See also Katz and Eldersveld (1961).

election outcomes than approval of the state legislature itself (Rogers 2016). What are the consequences and related factors of this increasing nationalization of state-level elections? What does this trend imply about voters and representation?

Changes in the partisan attachments of voters is perhaps the most prominent explanation for the increasing nationalization of elections. Research shows that party identification is a strong identity for voters and it appreciably affects their voting decisions (Campbell et al. 1960; Green, Palmquist, and Schickler 2002). In recent decades this attachment has become even potent, with Americans expressing intense dislike towards opposing partisans (Iyengar, Sood, and Lelkes 2012; Rogowski and Sutherland 2016; Webster and Abramowitz 2017). The increase in affective polarization may be causing voters to vote a straight ticket, leading to the increase in the nationalization of elections (Abramowitz and Webster 2016).

An alternative explanation is that the options facing voters in state and national elections are now more similar than they were before. There are two parts to this explanation. First, party labels are informative signals about what a politician will do once in office; if you know the politicians' party, you likely know their policy positions. Second, the state policy agendas have become more nationalized. In other words, the issues that are important in state politics now more closely mirror the issues that are important in national politics. If national issues are dominating state elections and politicians are taking distinct, polarized, partisan positions on those issues, then voters should be more likely to vote a straight ticket in order to choose the candidates who best represent their preferences.[2]

[2]Alternatively, it is possible researchers are now better able to characterize the relationship

These two explanations paint different pictures for democratic accountability. If voters are purely driven by affective polarization then they may evaluate candidates on things that have little to do with their position on the issues. In contrast, the second explanation is a more hopeful possibility, suggesting that voters are considering the issues and politicians' likely actions when deciding how to vote. I focus on testing the plausibility of the second argument—that the nationalization of the policy agenda and voter responses to it are related to the nationalization of elections.

While previous research has documented that partisanship has become a strong and informative signal about politicians' positions on the issues (Poole and Rosenthal 2007; Theriault 2008; Binder 2016; Levendusky 2009), very little work has examined how the state policy agendas have changed. In a notable exception to that rule, Hopkins (2018, Ch. 7) analyzes state party platforms and shows that state boundaries are less likely to explain differences in platform positions than they were in the past. I extend that work by testing whether the state policy agendas have become more nationalized and whether this nationalization is a contributing factor to the nationalization of gubernatorial elections. I thus directly test whether the nationalization of the policy agenda is related to the nationalization of elections.

Data limitations have been an obstacle to studying the nationalization of state agendas. I solve the problem of limited data by collecting a new database of annual State of the State addresses, which are given by U.S. governors and are analogous

---

between issue preferences and voting behavior. Ansolabehere, Rodden, and Snyder (2008), for example, aggregate survey responses over several items and find that voter preferences have a history of consistency, instead of noise. The authors argue that we observe noise at an individual preference level because of measurement error.

to the presidential State of the Union (SOTU). I use the State of the State (SOTS) addresses to determine whether state policy agendas have become more nationalized.

I perform automated content analysis to code for policy issues in the documents. I fit a topic model to the SOTS and SOTU addresses using cross-validated and grid-optimized latent dirichlet allocation, and then validate the model's issue topics and predicted codes using research assistants (Lowe and Benoit 2013; Quinn et al. 2010; Roberts et al. 2014). The results of this process are delineated in section 1.7 in lieu of including a duplicated appendix. I use the resulting codes to compare the SOTS and the SOTU on the basis of their *Topic Similarity*, which uses cosine distance to capture the degree to which the addresses talk about the same issues.

I carry out the analysis in three steps. First, I compare the governors' SOTS addresses to each other. If state agendas have become more similar over time, then I would expect to see an increase in the similarity between SOTS over time. On average, there has been a 70 percent increase ($p < 0.01$) in the similarity of SOTS addresses to each other, over the period beginning in 1960 and ending in 2016.

Second, I test whether governors' SOTS have become more similar to the President's annual State of the Union (SOTU) address. I use the same approach to measure the topics covered in the SOTU. I then compare the similarity between the SOTU and SOTS over time. I find that there has been a rapid increase in the similarity between the SOTS and SOTU over time. On average the similarity of SOTS and SOTU addresses over the period beginning in 1960 and ending in 2016 has increased by four-fold ($p < 0.01$).

Third, I test whether the nationalization of the state policy agenda corresponds

to an increase in the nationalization of gubernatorial election results. I estimate the nationalization of gubernatorial election results by examining the degree to which the presidential election results in the state predict the gubernatorial election results. I then estimate the degree to which the nationalization of the agenda – measured by the degree of topic similarity in the SOTS and SOTU addresses – moderates this nationalization of elections.

I find that the nationalization of state policy agendas moderates the nationalization of gubernatorial electoral results. Specifically, I find that when the governor's SOTS matches the president's SOTU, the governor's vote share is rapidly nationalized at a rate of 1.58 percentage points in off-cycle years and 4.10 percentage points in presidential years. On the other hand, diverging from the president's SOTU is associated with a decrease in the rate of nationalization.

Politicians at the state level are increasingly focusing on the same issues that are on the national agenda, and because they are focusing on the same issues, it is not surprising that voters are voting for candidates of the same party at both the state and national levels. The nationalization of state policy agendas and state elections have grown together.

Some of my secondary findings include a descriptive analysis of party and regional trends. These secondary findings suggest while there generally has been a trend in the increase in similarity between agendas over time, the Southern Democrats are largely responsible for the great increase in average similarity in the 1990s.

These results suggest that nationalization is not simply a result of partisan teaming. Instead, nationalization is strongly related to agenda speech. Politicians at the state

level are increasingly focusing on the same issues that are on the national agenda, and because they are focusing on the same issues, it is not surprising that voters are voting for the same candidates at both the local and national levels. The nationalization of state policy agendas and state elections have grown hand-in-hand.

I stop short of making any casual assertions. It is possible that voters take divergence from the national agenda as a signal which suggests they should not vote for the party's local candidate, as opposed to an opportunity to carefully consider the policy alternatives offered to them. However, it is clear that nationalization is a story that includes nationalization of the agenda; it is not solely a group-based response.

The paper proceeds as follows. First, I review and develop the two competing theories which would explain nationalization. Second, I introduce my methodology for the collection of the data and the automated content analysis produced by the topic model. Third, I analyze the SOTS data alone, to show how SOTS have become more similar to each other over time. Fourth, I analyze SOTS and SOTU data to show how they have become more similar to each other over time. Fifth, I combine the SOTS–SOTU data and nationalization data to demonstrate a strong relationship between the two. Finally, I review the results and suggest several interpretations of them.

2.2 The Nationalization of Elections

Research on the nationalization of politics has primarily focused on elections. Analysis shows that national factors are becoming stronger predictors of both congressional

elections and elections for state office. Congressional elections, for example, have been increasingly likely to move in tandem with national forces. These national forces have increased the frequency of apparent wave elections (*e.g.*, 2006 for the Democrats and 2010 and 2014 for Republicans).

Fiorina (2017), for instance, documents several important ways in which the trend towards nationalization manifests itself in congressional elections. First, there has been a decrease in the estimated size of the incumbency advantage (Jacobson 2015). In previous decades, politicians focused on local issues of interest to the district as a way to insulate themselves from voters who are dissatisfied with national politics. The diminished incumbency advantage suggests that Members of Congress are either doing this less, or they are being less successful in their efforts because politics is more nationalized.

Second, split-ticket voting has decreased to the lowest levels observed since researchers started tracking this information 60 years ago. Split ticket voting refers to situations where voters vote for one party for president and another party for Congress. In the 1970s and 1980s, 25 to 30 percent of voters were splitting their ticket. Now only about 10 percent split their tickets. Voters are simply more likely to vote a straight ticket.

Third, the president's popularity in the district has become increasingly predictive of congressional midterm elections. Fiorina shows that over time the president's vote share in the district has become more predictive of the congressional election in the district two years later. At the same time, the congressional election from two years earlier has become less predictive. Since 2006, "one can better predict the winner's

vote in a congressional district using the district's previous presidential vote than its previous House vote" (Fiorina 2017, page 10).

State elections have also nationalized. Hopkins (2018) shows that, among other things, the correlation between gubernatorial and presidential election returns at the state level has dramatically increased since the 1970s, state parties have become more homogeneous in their platforms, and individual American have come to prioritize national identity over state or regional identity.[3] Even media coverage has come to focus on national issues over local issues (Martin and McCrain 2019, see also). Finally, Rogers (2016) looks into what predicts state legislative elections. In his analysis, Rogers includes both presidential approval and approval of the state legislature. While he finds that both factors are predictive of state legislative races, Rogers finds that presidential approval has three times the impact of state legislature approval. State elections are now strongly tied to national trends.

Finally, Rogers (2016) looks into what predicts state legislative elections. In his analysis, Rogers includes both presidential approval and approval of the state legislature. While he finds that both factors are predictive of state legislative races, Rogers finds that presidential approval has three times the impact of state legislature approval. State elections are now strongly tied to national trends.

[3]Hopkins (2018) dissects the phenomenon of nationalization from several different perspectives, demonstrating that in addition the increase in the association between local and national electoral returns, "Americans' engagement with [and interest in] state and local politics has waned relative to their engagement with national politics" (page 257).

2.3 Two Explanations for the Nationalization of Election Outcomes

The nationalization of elections has thus been characterized by an increase in how well partisanship predicts election outcomes in each state and congressional district. At the most basic level, this appears to be driven by voters becoming increasingly likely to vote along party lines (Fiorina 2017).

One explanation for the increase in partisan-line voting is the increase in affective polarization. Voters have always had a strong social attachment to the parties with which they identify (Green, Palmquist, and Schickler 2002). However, over the last several decades, Americans have come to increasingly dislike opposing partisans (Iyengar, Sood, and Lelkes 2012; Rogowski and Sutherland 2016; Webster and Abramowitz 2017). One view is that this increase in affective polarization is responsible for the increase in straight ticket voting – by causing these group identities to more strongly influence voting decisions – that is behind the nationalization of elections (Abramowitz and Webster 2016).

Proponents of the ignorance theory argue that the nationalization of elections is a sign that voters are relying on their partisan attachments to make decisions, rather than a metered consideration of the issues and candidates. At the most basic level, elections are nationalizing because voters are more likely to vote for their chosen party, regardless of the level of government; electoral results at the national and state level are confounded by the strengthening of partisan identity. In summary, the essential argument is that group identity affects how individuals vote, and the increase in

affective polarization means that group identities have become stronger and thus more influential on individuals' voting decisions.

An alternative explanation is that the voting decisions that voters face at different levels of elections are now more similar. There are two parts to this argument. First, partisanship has become a more informative signal of candidate positions. Voters can use partisanship (and other cues) as a signal to make more informed decisions (Popkin 1994; Lupia 1994). In recent decades, party labels have become more informative signals about a candidate's positions on the issues as politicians have sorted into polarized ideological camps (Alvarez 1998; Ahler and Broockman 2018; Poole and Rosenthal 2007). The disappearance of moderate Republicans and Democrats means that there is little to no overlap between the two parties in Congress (Poole and Rosenthal 2007; Theriault 2008; Binder 2016) and at the state level (Shor and McCarty 2011). This increased polarization means that the party labels are more informative than they were 40 to 50 years ago. When voters go to the polls, party labels to provide credible information about where candidates stand on the issues and what they will do once in office (Grynaviski 2010). Evidence for this theory would suggest that elections are nationalized because voters "rationally" search for candidate platforms which support their positions at several levels of government.

Second, the policy agenda has become more nationalized. The claim that "all politics is local" is rooted in the idea that voters and politicians are focused on the policy agenda items of importance locally. The policy agenda matters because it affects the considerations that voters use to evaluate candidates (Schattschneider 1960). Therefore, this would suggest that when the state and national policy agendas differ,

voters should use different criteria to evaluate the candidates for governor than they use for the candidates for president. However, if the state and national agendas focus on similar issues, then voters should use similar criteria to evaluate candidates at both levels (especially when partisanship is a strong signal about candidates' positions), which should lead to a stronger correlation between election outcomes across offices.[4]

The nationalization of Congressional elections can be explained by a similar logic. In recent decades, Members of Congress have sorted into distinct and separate ideological camps. The increased ideological sorting means that voters have clearer choices and that the choice they face when voting for president is more similar to the choice they face when voting for their Member of Congress. Consistent with this possibility, scholars have noted a strong drop in split ticket voting in recent years (Fiorina 2017).

These two features together can explain the increased nationalization of elections as a rational reaction. The nationalization of state policy agendas matters because it is affecting the issues that voters are considering. The nationalization of the state policy agendas means that voters are voting on similar issues at both the state and the national levels. If polarization means that Democrats and Republicans are now giving voters a clearer choice, then perhaps it is not surprising that there is a stronger correlation between presidential and state level election outcomes.

[4]As the state policy agenda diverges more from the national policy's agenda, the correlation between vote totals in the presidential election and the state level election will become weaker because this divergence in policy agenda will lead the voters to consider different dimensions in the two races. The reverse may also be true. As state policy agendas become more nationalized, the factors affecting voter preferences towards the president and the v presidential candidates should become more predictive of how they vote in state elections.

2.4 Have State Policy Agendas Become More Nationalized?

While previous work has established that partisanship now provides a stronger signal about candidates' positions, I know little about the nationalization of the issue agenda. In this paper, I test whether the nationalization of the issue agenda can help explain the increase in the nationalization of election outcomes. To explore this possibility, I construct measures of the agenda by looking at the topics covered in the State of the State (SOTS) and State of the Union (SOTU) addresses. I then use the topics covered to create a similarity index that measures the similarity between two addresses (more on the measurement and methods below). I use this data to carry out three tests.

First, I test whether the issue agendas laid out in the SOTS have become more similar over time. If the states are focusing on national issues more over time, then they should have more issue topics in common over time.

Second, I test whether the issue agendas in the SOTS have become more similar to the SOTU over time. If the policy agendas have become more nationalized, then agendas at the state level should be increasingly similar to the agenda at the national level. Every year, the president lays out the political agenda for the upcoming year in his State of the Union (SOTU). Governors get to do the same thing for their individual states with their State of the State (SOTS). If all policy agendas are local, then there should be little correlation between these two; however, if the states are increasingly taking on the same issues dealt with at the national level, then I would expect to see an increasing similarity between the SOTS and SOTU over time.

Finally, I test whether the nationalization of the state policy agenda corresponds to an increase in the nationalization of gubernatorial election results. In particular, I test whether the presidential election results better predict gubernatorial election results when the SOTU is more similar to the SOTS in that state.

It is important to not caricature my argument or my tests. I do not argue that the nationalization of state policy agendas is the sole factor responsible for the nationalization of state elections. I expect that complementary forces are at work. Most notably, voters have strong incentives to use party labels as a short cut when deciding how to vote. Over time, politicians have sorted into ideologically consistent parties (Levendusky 2009), causing party labels to become a stronger signal about a candidate's position. While using the party label as a shortcut to infer a candidate's positions will occasionally cause voters to make the wrong decision (Dancey and Sheagley 2013), voters will make the correct inference much of the time. I think that using the party labels as a short cut is likely working in tandem with the increased nationalization of the issue agenda. As the state political agendas become more nationalized, voters have stronger incentives to rely on the party label when deciding how to vote.

2.5 Building the Corpus of State of the State Addresses

While there is evidence that polarization has led the party label to become more informative of candidate positions, there is limited evidence of whether the state level policy agendas have nationalized. Hopkins (2018), in one study, analyzes state party platforms to show that state boundaries are less likely to explain differences in platform positions than they were in past. Additional work is limited, though. Perhaps little work has been done on agenda nationalization because it is so hard to find the appropriate data and method with which to test it.[5]

I solve the problem of limited data by introducing a new database of annual State of the State addresses, which are given by U.S. governors and are analogous to the presidential State of the Union. I use the State of the State (SOTS) addresses to determine whether state policy agendas have become more nationalized.

To study whether state agendas have nationalized over time, I collected data on the annual State of the State (SOTS) addresses. There are several advantages of studying Governors' SOTS addresses. Most importantly, the gubernatorial SOTS address – like the State of the Union – is a prominent speech that governors use to discuss the policy agenda for the upcoming year. As a result, the SOTS address is likely to capture the issues that will be used for voters as they evaluate the candidates running for office. Further, the SOTS addresses are typically given on a regular basis

[5]There is a well-established tradition in previous research papers, which have used state of the state speeches to represent the gubernatorial agenda (Kousser and Phillips 2012; Herzik 1983; DiLeo 1997; Barth and Ferguson 2002), or to represent the governor's ideology (Weinberg 2010).

and are scheduled far in advance. This allows us to have a way to reliably measure the agenda of governors in a way that is consistent across states and over time.

Much of the State of the State address is used to discuss the policy goals for the upcoming year. In doing so, the address is likely to capture the issues that will be used for voters as they evaluate the candidates running for office. I am neutral about why this relationship holds. On one hand, it could be that governors are focusing on a given set of issues because those are the issues that voters care most about. In other words, they may simply be responding to voter demand. On the other hand, their focus on these issues may lead voters to evaluate candidates in light of the issues that they put on the agenda (Schattschneider 1960). Governors are in a position to affect the agenda because they hold a prominent political post to which state news media give ample attention. Voters are very likely to know who their governor is (Jennings 1996), and they are much more likely to hear the messages they share (Bennett and Iyengar 2008). Governors, because of their positions, are simply much more likely to be heard by voters than other politicians in the state. And it could also be a combination of these two processes working together. For my purposes, it is important that the SOTS is a good measure of the issues that voters will use when deciding how to evaluate candidates.

The study of the SOTS addresses is fortuitous because they are mandated speeches. While the content of the address may be affected by voter preferences, the *decision to give the address* is less subject to concerns about endogeneity. The SOTS address is a scheduled presentation, with dates typically set by distant, or even contrarian, state legislatures. The fact that this address is scheduled means that I have a way to

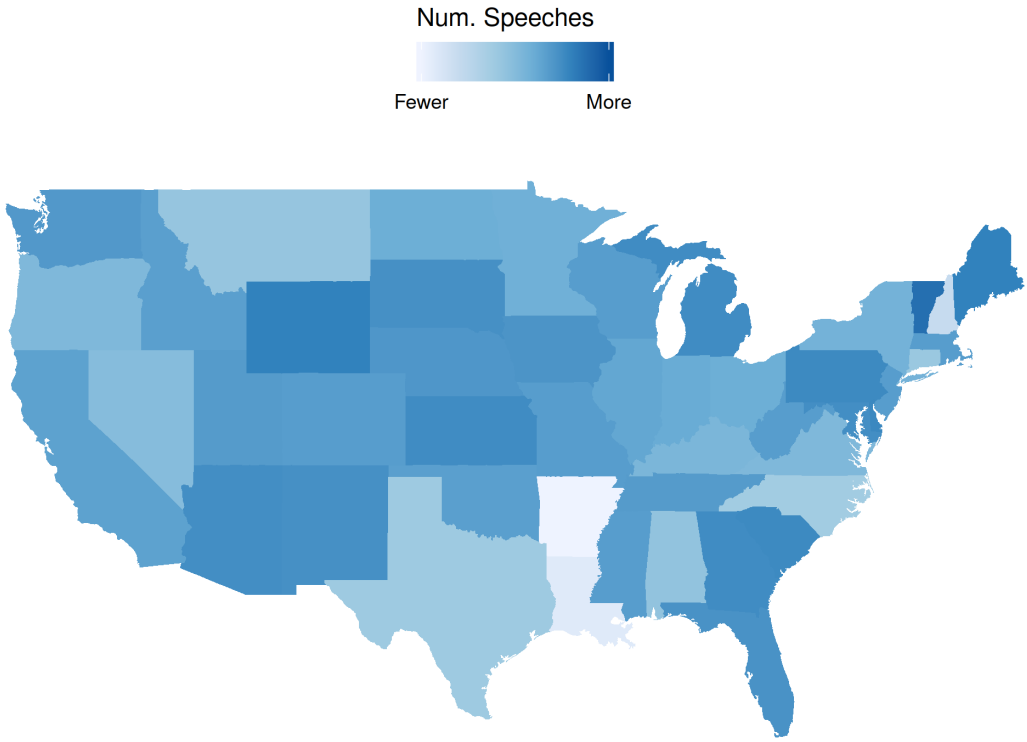
reliably measure the agenda of governors in a way that is consistent across states and over time.

I built the corpus of the annual SOTS addresses for all states as far back as I could, yielding addresses originating from the late 1800's for some states. For the analyses in this paper I use the addresses from 1960–2016 because Hawaii and Alaska became states in 1959. This smaller, more complete subset yields comparable data for all 50 states over this period. The corpus used for the present study comprises 2,236 speeches.

Efforts to collect and analyze gubernatorial speech data has been hampered by the difficulty in acquiring clean data. The very aspect that makes the study of state-level text data attractive (the diversity that comes from using 50 states) makes the study difficult to carry out because each state must be collected separately. Only a few states have the speeches readily available. For this reason, I directed my team of research assistants to systematically contact each state and request all gubernatorial speeches from the year 1960 to present. For each state, the research assistant collaborated with state legislative librarians and archival sources to diligently collect every available SOTS address. If states were unable to provide the collection services for us, I traveled to the state legislative libraries to source and scan the speeches. See table 2.1 for detail on the coverage of each state, and see figure 2.1 for visual representation of the number of speeches gleaned from each state.

Once collected, the disheveled state of the data itself presents another obstacle (to this project and others like it). Few, if any, documents are machine-readable—they are not in a format that can be directly used for text analysis. Optical Character

Figure 2.1: Nationwide Coverage of State of the State Addresses, 1960–2016



Note: Figure reports coverage of the State of the State addresses collected, from 1960–2016. Each state is shaded proportionally to the number of speeches collected from the state. Darker shades indicate that we found more speeches, and lighter shades indicate that we found fewer speeches. For example, in Louisiana, we were unable to find many speeches. Shades are relative. The shades are not adjusted for cases in which more than one speech was given in a year (*i.e.*, states in which more speeches were delivered will have a darker shade than states in which more speeches were delivered). Alaska and Hawaii are omitted from the plot. The plot suggests that coverage is even over the states eligible as part of the study.

Table 2.1: Coverage of Collected State of the State Addresses by State, 1960–2016

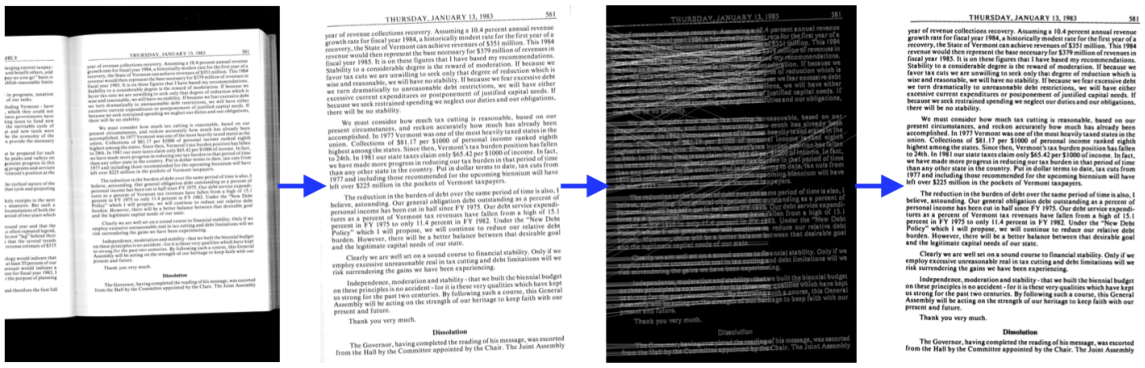
State	Coverage (%)	State	Coverage (%)	State	Coverage (%)
Alabama	46	Louisiana	25	Ohio	71
Alaska	88	Maine	100	Oklahoma	82
Arizona	96	Maryland	96	Oregon	64
Arkansas	23	Massachusetts	84	Pennsylvania	100
California	80	Michigan	98	Rhode Island	84
Colorado	84	Minnesota	70	South Carolina	100
Connecticut	50	Mississippi	84	South Dakota	95
Delaware	100	Missouri	84	Tennessee	86
Florida	93	Montana	54	Texas	50
Georgia	98	Nebraska	89	Utah	86
Hawaii	68	Nevada	61	Vermont	100
Idaho	82	New Hampshire	32	Virginia	62
Illinois	77	New Jersey	84	Washington	88
Indiana	73	New Mexico	95	West Virginia	84
Iowa	91	New York	68	Wisconsin	84
Kansas	98	North Carolina	48	Wyoming	100
Kentucky	64	North Dakota	73		

Note: Entries are coverage percentages, calculated by dividing the number of years for which a speech was collected by the number of years we would expect to have been able to collect for the state. Some states only do State of the State addresses biennially. In these cases, the proportions use the biennial denominator. The table suggests coverage of the speeches collected is good.

Recognition (OCR) software does not remedy the problem: often, document scans are of such poor quality that they cannot be processed using state-of-the-art OCR technology. Scans are skewed, multiple pages are in the frame, or someone's hand made it into the image, resulting in inaccurate character recognition. Manual transcription of these documents is more accurate, but it is cost-prohibitive and takes a long time. I searched extensively for software to extract the text at a low cost, but found that there exists no programmatic solution to extract text from poorly scanned speech images. This problem has become increasingly apparent as political science research using text data has increased.[6]

[6]Ban et al. (2017), for example, encounter the issue of poorly extracted text and manually apply a battery of regular expressions to fix apparent issues. Though doing so makes text appear cleaner,

Figure 2.2: Application of Machine Learning to Isolate and Extract Text



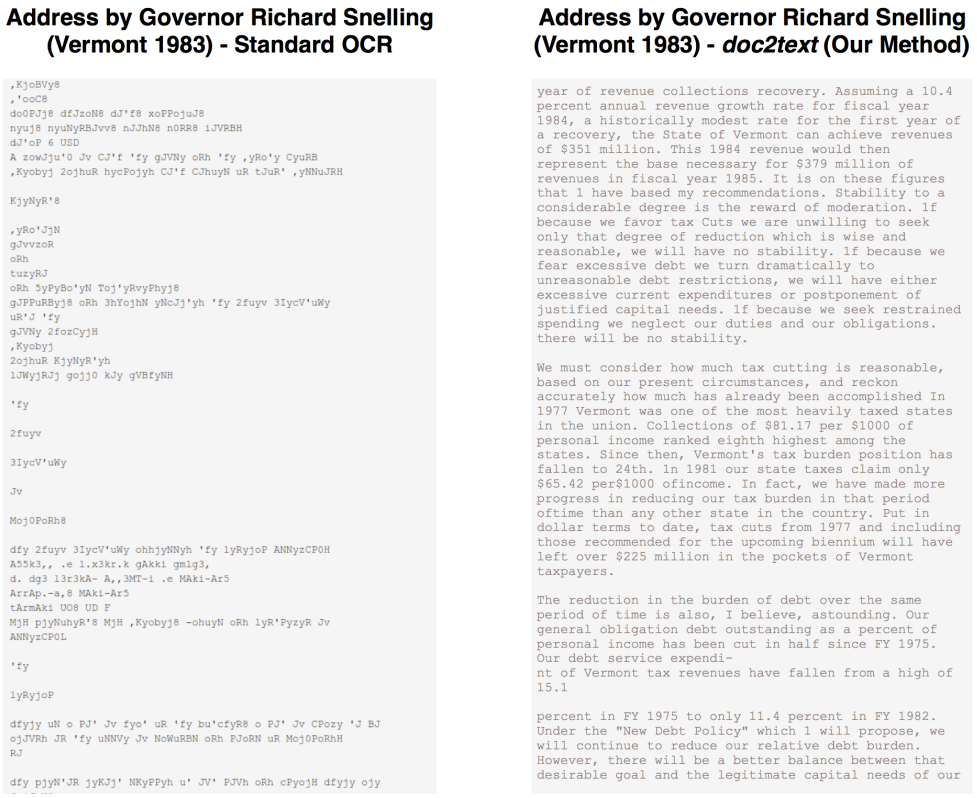
To solve the issue, I developed a software package called *doc2text*, which applies machine learning to isolate, enhance, and extract text from hard-to-read documents.[7] *doc2text* is different from other packages because it corrects for document layout problems with a novel statistical approach before it runs OCR, resulting in a significantly higher level of accuracy. With *doc2text*, my team was able to extract text from more than 1,500 documents that would have otherwise required manual transcription, significantly reducing costs and speeding up the data development process. See figure 2.2 for a detail of the process. The text extracted using this program are also of higher quality than those extracted using traditional OCR, yielding higher quality text without the introduction of researcher bias. Figure 2.3 demonstrates that the extraction quality from poorly scanned documents was greatly improved.

To evaluate the quality of the text extraction, I employed research assistants to score a simple random sample of 500 documents as "readable" or "not readable."

it also introduces researcher bias into the text, because the decision to remove some features of the text and not others may affect downstream analyses (Denny and Spirling 2018).

[7]After its release in 2016, *doc2text* became one of the most popular open-source projects on Github, placing it in the top tenth of a percent of all projects available on the website. The software has been used for instructive purposes at Pennsylvania State University. *doc2text* is available for immediate public use at http://github.com/jlsutherland/doc2text.

Figure 2.3: *doc2text* Extracts Higher Quality Text



Documents were scored three times each. The design had high interrater reliability (Krippendorff's $\alpha = 0.92$; SE $= 0.01$). RAs coded 99.8 percent of the documents as readable, where I consider a document to be readable if all scores for that document were "readable." The software did a good job of extracting machine readable text from the images.

2.5.1 Text Preparation and Document Construction

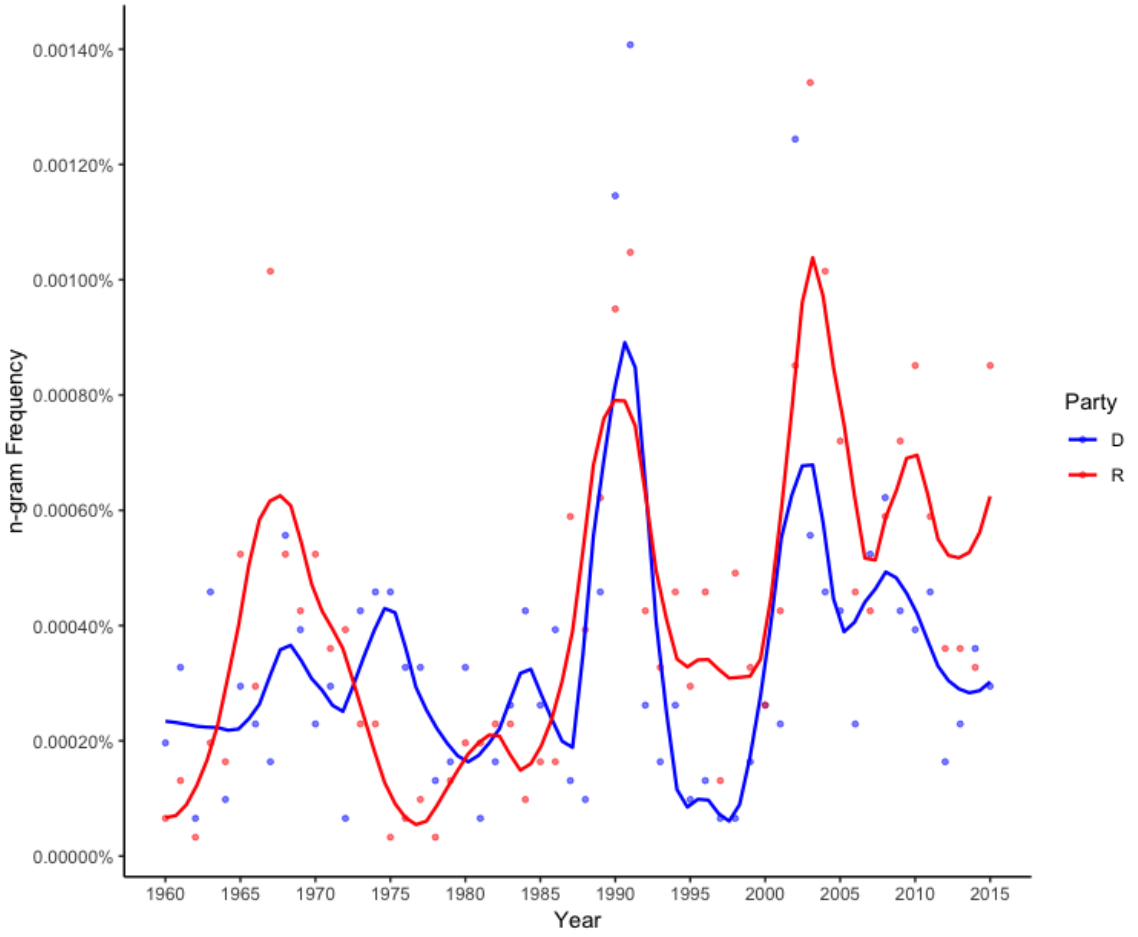
In my study, I use two databases of political speech: the State of the States and the State of the Union addresses. The raw text for the State of the States corpus comes

from the database I collected. The raw text for the State of the Union corpus comes from the database provided by (Benoit et al. 2018). Starting with the raw text, I followed the basic procedures that are used to process text that have become relatively standard in the literature analyzing political texts (Grimmer and Stewart 2013; Denny and Spirling 2018, and references therein).[8] As part of my study, I parsed, stopped noun phrases from the documents (using a procedure similar to the procedure applied in Handler et al. 2016). I then lemmatized the lowercased constituent pieces using WordNet. Punctuation and numbers were cleaned from the text.

The phrasing and lemmatization process I implement generates features with much greater utility than the features generated by the bag of words approach, because the features it constructs are embedded with contextual meaning that would otherwise be lost. The simple bag of words approach assumes a constant meaning for any given 1-gram, regardless of context. For example, the bag of words can only compare the three phrases "credit card", "tax credit", and "cap-and-trade credit" on the basis of "card", "tax", and "cap-and-trade", because it assumes the word "credit" has a fixed meaning. In the context of the State of the States, the distinction between such phrases is important, because, for instance, a credit issued to curb carbon emissions implies a set of political circumstances and ideologies that is different from the set implied by credit card reform. I also filtered procedural phrases using the approach of Gentzkow, Shapiro, and Taddy (2016), which uses Robert's Rules of Order (Robert

[8]Numerous studies apply text regularization and transformation to the study of political and economic phenomena. To name a few: Gentzkow and Shapiro (2010), Quinn et al. (2010), Jensen et al. (2012), Gentzkow, Kelly, and Taddy (2017), and Jelveh, Kogut, and Naidu (2015). See Grimmer and Stewart (2013) for an excellent overview. See Denny and Spirling (2017) for a review of standard text pre-processing procedures.

Figure 2.4: Discussion of "war" in State of the State Addresses



Note: Figure reports the usage rates of the token "war" by both parties in their State of the State speeches, over time. Usage rates are n-gram frequencies, which are computed by dividing the number of times the word "war" is used in SOTS addresses in a year by that party, by the total number of times any word was used in any SOTS address in that year by that party. The blue points are frequencies for Democrats and the red points are frequencies for Republicans. The blue and red lines are predictions from locally weighted regressions. The data show that time periods around the Vietnam, Gulf, and Iraq wars see increases in mentions of "war" relative to other features used in the same year. The figure demonstrates how increases in the salience of war in agenda speech map to the involvement of the United States in armed conflict. The average governors of both parties change their usage of use of the word at similar rates.

1915) to form a dictionary of procedural phrases, which are then removed.

As a *prima facie* validation of whether the State of the State addresses are covering the agenda, I investigate the extent to which US involvement in international conflicts have led to increased attention to the topic of war in the governors' State of the

State addresses. I focus on war because this is exogenous to the governors' SOTS address; governors are not the ones choosing whether to go to war. At the same time, however, war is important enough to affect elections (Grose and Oppenheimer 2007) and therefore should show up on the agenda. If my approach is effectively measuring the political agenda I would expect an uptick in mentions of war in the SOTS during the duration of major armed conflicts.

Figure 2.4 reports the frequency with which the word "war" is mentioned over time by party. The blue points are frequencies for Democratic governors and the red points are frequencies for Republican governors. The blue and red lines are locally weighted regression lines. The values are the n-gram frequencies for "war" at several points in time. n-gram frequencies are widely used to study the distributional properties of text features (Grimmer and Stewart 2013), and from those properties, scholars commonly draw inferences about societal trends (Michel et al. 2011). An n-gram is a feature derived from sequence of n contiguous items in a sequence of text. The frequency f_{kt} for an n-gram k in period t is

$$f_{tk} = \frac{\sum_{i=1}^{N_t}[w_{it} = k]}{N_t}, \tag{2.1}$$

where w_i is the ith token in the vector of tokens present at time t, and N_t is the size of the vector of tokens present at time t. For example, f_{tk} for the 1-gram "war" in the year 1960 would be the number of times the 1-gram "war" occurs in 1960 divided by the total feature count for 1960.

The data show that time periods around the Vietnam, Gulf, and Iraq wars see

increases in mentions of "war" relative to other features used in the same year. The figure demonstrates how increases in the salience of war in agenda speech map to the involvement of the United States in armed conflict. The average governors of both parties change their usage of the word at similar rates. Figure 2.4 demonstrates that the data set captures variation in the relative salience of important topics (such as war) over the 1960–2016 time period. As such, I would expect the corpus to be a good way to measure the policy agenda over this period.

2.5.2 Measuring the Similarity Between Speeches

My empirical tests use measures of the similarity between different speeches. This includes comparing the SOTS to each other, and comparing the SOTS addresses in each year to the presidential SOTU delivered in that year.[9]

While there are numerous ways to measure the similarity between speeches, my theoretical interest lies in understanding how changes in the agenda affect election outcomes. One of the most important ways that the agenda can affect how individuals vote is by affecting what issues they vote on. Many voters make decisions based on a single issue that they think is most important in the given election (McCombs and

[9]For some of my similarity estimates, I compare one corpus of documents – the SOTS – to another, separately generated corpus of documents – the SOTU. Traditionally, analyses of text data in political science and economics do not compare documents from multiple corpora. Lewis and Tausanovitch (2015), and this dissertation, demonstrate that comparing across corpora can be problematic if the corpora are not generated by the same process (or, at least, a similar process). I use the delta statistic, developed in Section 1.5, to test the alternative hypothesis that the corpora were drawn from incomparable data generating processes. The statistic compares the observed distance between matched topics (probability distributions over their tokens) in each corpus to the expected distance under the null hypothesis that they were drawn from comparable data generating processes. The test fails to reject the null, which suggests the corpora are generated by similar data generating processes. I proceed by comparing the two corpora on the basis of their topics, jointly estimated.

Shaw 1972). Elections are thus often determined by which issue wins the conflict of conflicts to become the dominant issue in the eyes of voters (Schattschneider 1960). This gives politicians incentives to think about what issues they should emphasize in order to win elections (Riker et al. 1996; Aldrich and Griffin 2003; Druckman, Jacobs, and Ostermeier 2004; Dragu and Fan 2016). Because of my theoretical interest in election outcomes and agenda-setting, I measure similarity by looking at the topics covered in the SOTS speeches. Topic modeling is perfectly suited for our purposes because it allows for a consistent measure of how much each speech focuses on each topic.

In this case, I estimate the proportion of each speech dedicated to each topic using latent dirichlet allocation. My grid optimization procedure delineated in section 1.5 determined the appropriate number of topics by maximizing held-out topic coherence Mimno et al. 2011. The optimal model estimates 120 topics. Section 1.5 provides a detailed discussion for how I arrived at the model parameters and validated its outputs. I use this model to estimate the proportion of each SOTS speech dedicated to each topic (Quinn et al. 2010; Grimmer 2010).

I use the information on the proportion of each speech devoted to each topic to estimate the similarity between speeches. In particular, I employ the cosine similarity metric to estimate similarity between speeches.[10] The cosine similarity of any two speeches is akin to the ideological similarity between any two ANES respondents;

[10]Readers may be more familiar with distance represented as a distance metric instead of a similarity metric. I choose cosine *similarity* instead of cosine *distance* because it makes the results more interpretable. For instance, in a regression of cosine similarity on vote returns, a positive coefficient for vote returns would suggest a positive relationship between agenda similarity and vote returns; if cosine distance is used, the coefficient would be negative. The cosine distance is simply 1 minus the cosine similarity.

the similarity between two individuals scoring the same on every policy item will be greater than the similarity between two individuals who have different scores for any policy item.[11] Similarly, when two speeches cover the same exact topics in the same exact proportions, the topic similarity equals 1. If there is no overlap on the topics covered in the speeches, then their topic similarity equals 0. When any topic shares the exact same proportion across a pair, the contribution of that topic to the similarity of the pair is maximized.

Mathematically, the cosine similarity of two speeches is the dot product over their proportion vectors, normalized to a scale between 0 and 1 (I will never see negative values, because I cannot observe negative proportions of topics). For each pair of speeches s, I estimate a *Topic Similarity*, $\hat{\theta}_{ss'}$ as a function of the correlation between their 120 topic proportions β_k:

$$\hat{\theta}_{ss'} = \frac{\sum_{i=1}^{K} \beta_{ks}\beta_{ks'}}{\sqrt{\sum_{i=1}^{K} \beta_{ks}^2}\sqrt{\sum_{i=1}^{K} \beta_{ks'}^2}}, \ \forall s, \ s \neq s'. \tag{2.2}$$

I estimate a *Topic Similarity* score for each pair of SOTS and SOTU ($N = 2,236+56 = 2,292$) speeches. This results in a total of 2,623,194 similarity scores, excluding cases where a speech is compared to itself. From these 2 million scores, I generate the datasets I use for analysis: a table of 44,743 State1–State2–Year dyads, which allows for analysis of interstate trends over time (this dataset excludes comparisons of any speech to its own state's speeches), and; a table of 2,236 State–Year dyads, which allows for for analysis of SOTS–SOTU trends over time.

[11] Similar measures are used in non-parametric matching techniques; see, e.g., Iacus, King, and Porro 2019, and references therein.

2.6 The Changing State Policy Agendas Over Time

I begin by looking at whether the SOTS have become more similar to each other over time. To do so, I create an annual, dyadic dataset. In other words I create an observation for Alabama–Wyoming in 1960, an observation for Alabama–New York in 1960, and so on, while also creating a dyad for each year so that I have observations for Alabama–Wyoming in 1960, 1961, 1962, and so on. In a handful of cases where multiple speeches were given within a year, and I generated multiple similarity scores, I took the unweighted average of the scores to produce a single score for the dyad. Dyads without scores are treated as missing data. For each dyad, I estimate the *Topic Similarity*, $\hat{\theta}_{ss'}$ for the speeches.

Table 2.2: Topic Similarity in SOTS Addresses Over Time

	Agenda Similarity	
	(1)	(2)
Time Trend (1960=1)	0.004***	0.004***
	(0.0001)	(0.0001)
Copartisans		0.001
		(0.002)
Intercept	0.157***	
	(0.002)	
N	44,743	44,743
State FE	No	Yes

Note: Entries are ordinary least squares regression coefficient estimates and standard errors, with clustering by year. The dependent variable is the Topic Similarity between State of the State addresses given in the same year. *p<0.1; **p<0.05; ***p<0.01.

I test whether the agendas have become more nationalized over time by regressing

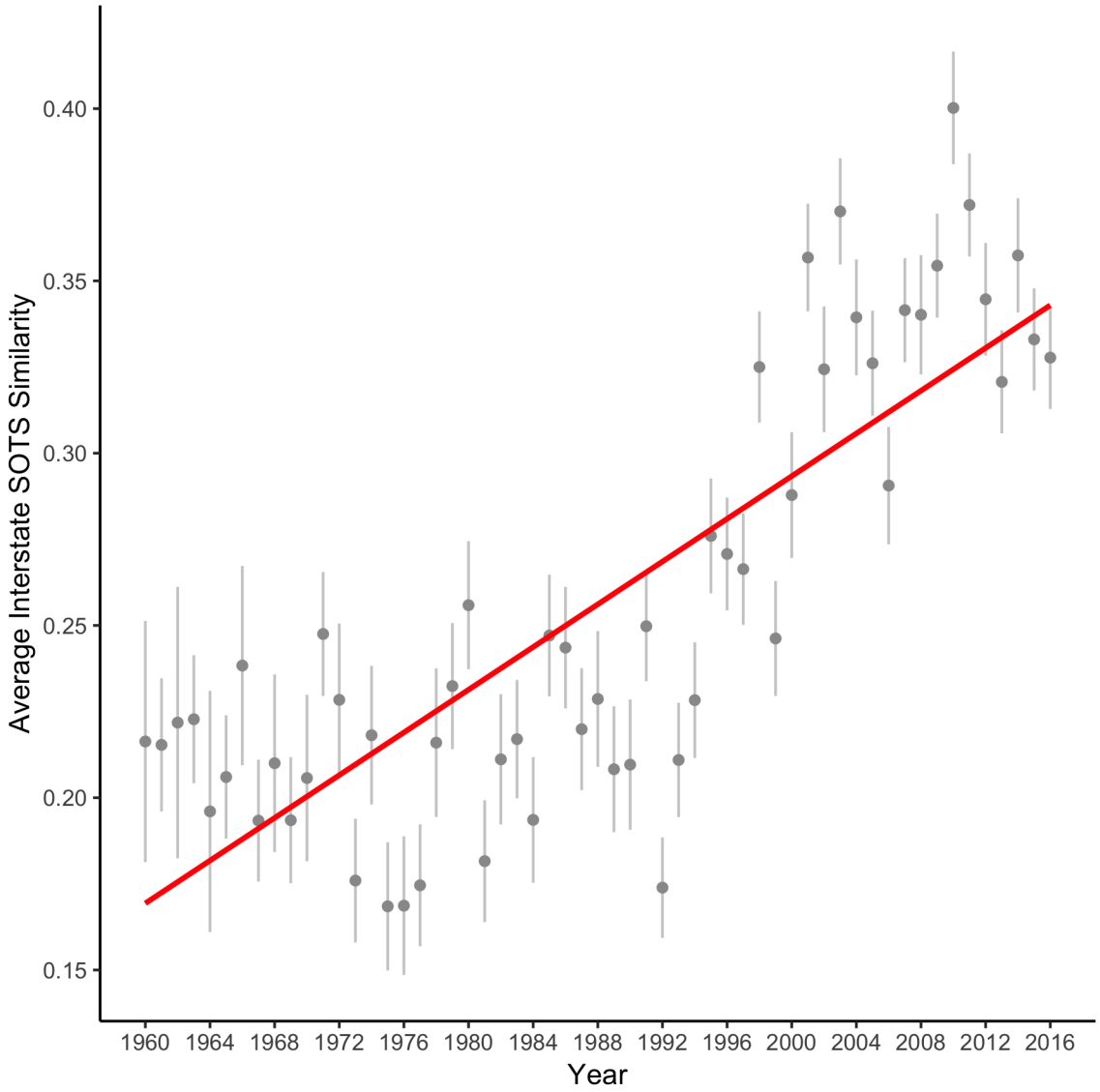
Topic Similarity on a time trend that I set equal to 1 in 1960. I use an OLS model for the estimation. To account for the fact that I have multiple observations each year, I cluster the standard errors on year. The results of the regression are presented in Table 2.2. Column 1 presents the results of just including a linear time trend. Column 2 includes state fixed effects and a dummy variable for whether the two governors giving the addresses are copartisans. In both regression models, the coefficient is highly significant and equal to 0.004. Because my dataset covers more than 50 years, the 0.004 coefficient suggests an increase of nearly 0.20 (because $0.004 * 50 = 0.20$). This is a very large increase and matches what is shown in Figure 2.5.

The results are also shown in Figure 2.5. Figure 2.5 gives the average *Topic Similarity* scores over time, with the thick red line showing the best linear fit. As the Figure 2.5 shows, the level of similarity has increased over time, with a large increase occurring in the 1990s. Early in the time period, the *Topic Similarity* score was around 0.20. By the end of the period, the similarity score was close to 0.35. That 0.15 increase represents about a 70 percent increase in the similarity of SOTS addresses over time. This increase is also statistically significant. State agendas have become increasingly similar over time.

2.6.1 Regional Differences in Similarity Over Time

It is common knowledge that due to regional similarities in geography, culture, and ideology, the political agenda of one region may tend to look different from that of another region (Key 1950; Hopkins 2018, *e.g.*). For instance, the 1960s political agenda

Figure 2.5: Trend in Interstate SOTS Similarity



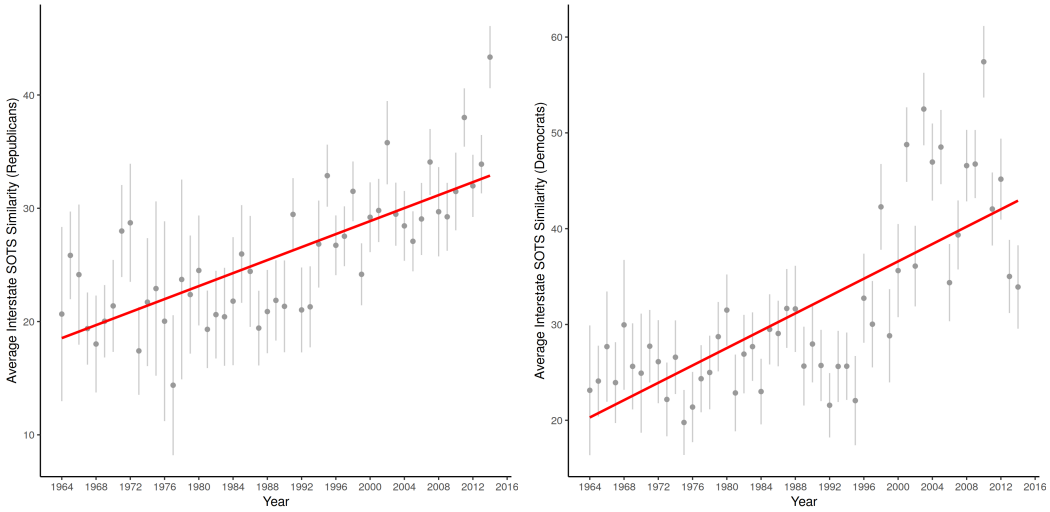
Note: Figure reports the relationship between time and the interstate SOTS similarity. Values are coefficients from a regression of similarities between the State of the State addresses given each year on a set of dummy variables for year. Each point is the coefficient estimate for that year. Similarity is a distance metric, which increases as the distance between two samples of text decreases. An ordinary least squares line of best fit is shown in red. The general trend has been for the State of the State addresses to become more similar to each other over the last six decades.

of the American Northeast, which largely supported major civil rights legislation like the Civil Rights Act of 1964, looked very different from the political agenda of the South, which was vehemently against integration of African Americans in schools, housing, and public facilities. This regional difference cut brutally within the Democratic party, and by some accounts is responsible for the transformation of the Democrats and their ideology in the South (Carmines and Stimson 1989).

In the preceding section, I show how the state agendas have become more similar to each other over time. But what of the role of region and party? In this section, I condition my analysis further to examine whether region and party play a distinct role in the evolution of the nationalized agenda. My analysis analyzes similarity over time, but subsets the data by party (Democrat, Republican) and region (West, Midwest, South, and Northeast).

Figure 2.6 plots the average *Topic Similarity* scores over time, with the thick red line showing the best linear fit. The data are subsetted by the party of the Governor delivering the SOTS address. The figure suggests that in general, in both parties, the level of agenda similarity has increased over time. However, it adds further color to the large increase occurring in the 1990s. The data suggest that the Democrats are largely responsible for the large increase in similarity in the early 1990s, with great variability coming into play throughout the later period of the time series. Republicans, on the other hand, have seen steadier agenda similarity increases over time. This result may owe to the Southern realignment during House Speaker Newt Gingrich's *Contract with America*, whereby the final "nail in the coffin" of the Democrats in the South was struck, and what state offices remained in the South were taken by Republicans.

Figure 2.6: Trend in Interstate SOTS Similarity by Party



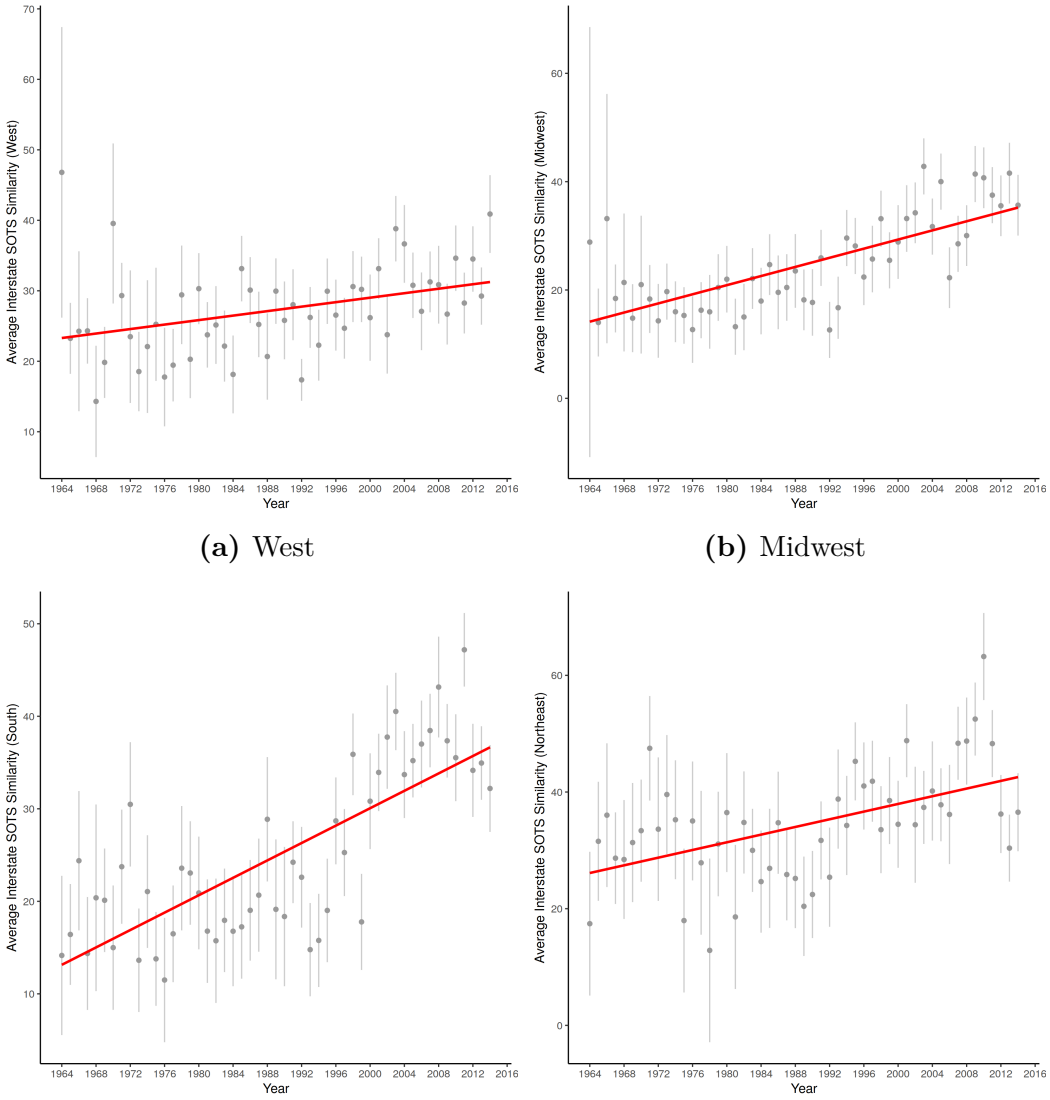
(a) Republicans **(b)** Democrats

Note: Panels report the relationship between time and the interstate SOTS similarity, by party. Values are coefficients from a regression of similarities between the State of the State addresses given each year on a set of dummy variables for year. Each point is the coefficient estimate for that year. Similarity is a distance metric, which increases as the distance between two samples of text decreases. An ordinary least squares line of best fit is shown in red. The general trend has been for the State of the State addresses to become more similar to each other over the last six decades.

While state agendas have become increasingly similar over time, the burst in the early 1990s seems to be driven by the Democrats.

Figure 2.7 is another plot of the the average *Topic Similarity* scores over time, with the thick red line showing the best linear fit. The data are subsetted in this case by the region of the Governor delivering the SOTS address. The figure suggests that in general, in all regions, the level of agenda similarity has increased over time. However, it again adds further color to the large increase occurring in the 1990s. The data suggest that the South is largely responsible for the large increase in similarity in the early 1990s, with great variability coming into play throughout the later period of the time series. The other regions, on the other hand, have seen steadier agenda similarity increases over time. This result may also owe to the Southern realignment

Figure 2.7: Trend in Interstate SOTS Similarity by Region



(a) West **(b)** Midwest

(c) South **(d)** Northeast

Note: Panels report the relationship between time and the interstate SOTS similarity, by region. Values are coefficients from a regression of similarities between the State of the State addresses given each year on a set of dummy variables for year. Each point is the coefficient estimate for that year. Similarity is a distance metric, which increases as the distance between two samples of text decreases. An ordinary least squares line of best fit is shown in red. The general trend has been for the State of the State addresses to become more similar to each other over the last six decades.

during House Speaker Newt Gingrich's *Contract with America*. While state agendas have become increasingly similar over time, the burst in the early 1990s seems to be driven by the South.

2.7 Similarity Between the SOTS and the SOTU Over Time

An even more direct way of looking for the nationalization of state policy agendas is to compare the state agendas (as laid out in the SOTS) to the the agenda laid out in the State of the Union (SOTU). If state policy agendas have nationalized, then I would expect to see the *Topic Similarity* between the SOTS and the SOTU addresses to increase over time.

For this analysis, my dataset includes one observation for each SOTS address I have. For each of SOTS address, I calculate the *Topic Similarity* between it and the SOTU address given in the same year. I tested the significance by running a linear regression model, where I include a time trend as a predictor. The results are reported in table 2.3.

Table 2.3 gives the regression results corresponding to Figure 2.8. The dependent variable in both models is the *Topic Similarity* score between the SOTS address and the SOTU address given that save year. The main independent variable is the linear time trend which is set equal to 1 in the year 1960. Column 2 also includes a dummy variable for whether the governor belongs to the same party as the president and fixed

Table 2.3: Topic Similarity in SOTS and the SOTU Addresses Over Time

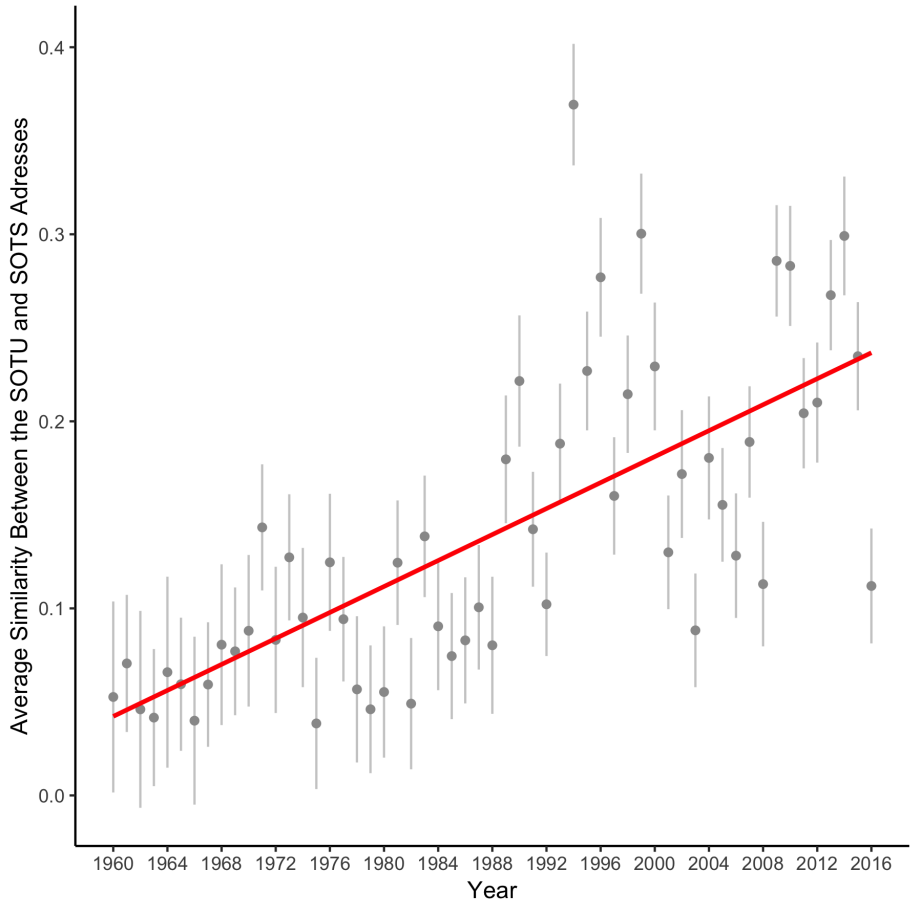
	Agenda Similarity	
	(3)	(4)
Time Trend (1960=1)	0.003***	0.003***
	(0.0001)	(0.0001)
Copartisans		−0.004
		(0.005)
Constant	0.043***	
	(0.004)	
N	2,218	2,218
State FE	No	Yes

Note: Entries are ordinary least squares regression coefficient estimates and standard errors, with clustering by year. The dependent variable is the Topic Similarity between the State of the Union and the State of the State addresses in the same year. *p<0.1; **p<0.05; ***p<0.01.

effects for states. Finally, I account for the fact that I have multiple observations in each year by clustering the standard errors by year. In both models the coefficient on the time trend is statistically significant and substantively large.

Figure 2.8 gives the average *Topic Similarity* scores over time, with the thick red line showing the best linear fit. The data shows a clear increase in the similarity over time. During the early period, the *Topic Similarity* was below 0.05. By the end of the period it had increased by over 20 percentage points (to close to 0.25). This four-fold increase is also statistically significant. Over time, the topics covered in the governors' State of the State addresses have become increasingly similar to what the president is covering in his State of the Union address.

Figure 2.8: Similarity of *State of the State* and the *State of the Union* Addresses Has Increased Over Time



Note: Figure reports the relationship between time and similarity. Values are coefficients from a regression of similarities between State of the State addresses, given by governors, and State of the Union addresses, given by the President, on a set of dummy variables for year. Each point is the coefficient estimate for that year. Similarity is a distance metric, which increases as the distance between two samples of text decreases. An ordinary least squares line of best fit is shown in red. The general trend has been for the average State of the State addresses to become more similar to the State of the Union over the last six decades.

2.7.1 Does the National Agenda Lead the Local Agenda?

While the SOTU and SOTS have in the aggregate become similar, it is still unclear if one leads the other, or if they have changed concurrently. The study of whether one leads the other is important because it reveals who the strategic policy agenda "setter" is—the national party, or the local parties. At one extreme, we might imagine a national party that, in the Schattschnederian sense, has come together to quash local "bossism" and allow local parties to set responsive agendas (Schattschneider 1942) on local issues. At the other extreme, we might imagine the combination of voter behavior and organizational largess to provide incentives to rally around a single, nationalized party agenda. The voter-based theory I have introduced in this chapter would lead us to expect that the national agenda leads the local agenda, because nationalizing factors have created strong incentives to present a clear and unified agenda within the party.

Governors also have incentives to be strategic about the agendas they lay out. Kousser and Phillips (2012), for example, demonstrate how governors interested in the success of their policy proposals may chose to adjust whether they pursue their preferences through a budgetary proposal rather than the "policy game," or; how governors with presidential (or other national) aspirations may choose to propose unpassable policies to their adversarially controlled statehouses, just to signal nationally their fitness for higher office.[12] These incentives could, at times, align to promote alignment

[12]For example, Governor Mitt Romney in his 2006 State of the State address signaled his support for socially conservative policies, such as abstinence education (Kousser and Phillips 2012, page 51). The state house was controlled by a majority of socially liberal Democrats in that biennium, making the proposal useless for the purposes of state legislation. It did, however, help to solidify his conservative base for his race for the Republican nomination in 2008, and his ultimate selection as

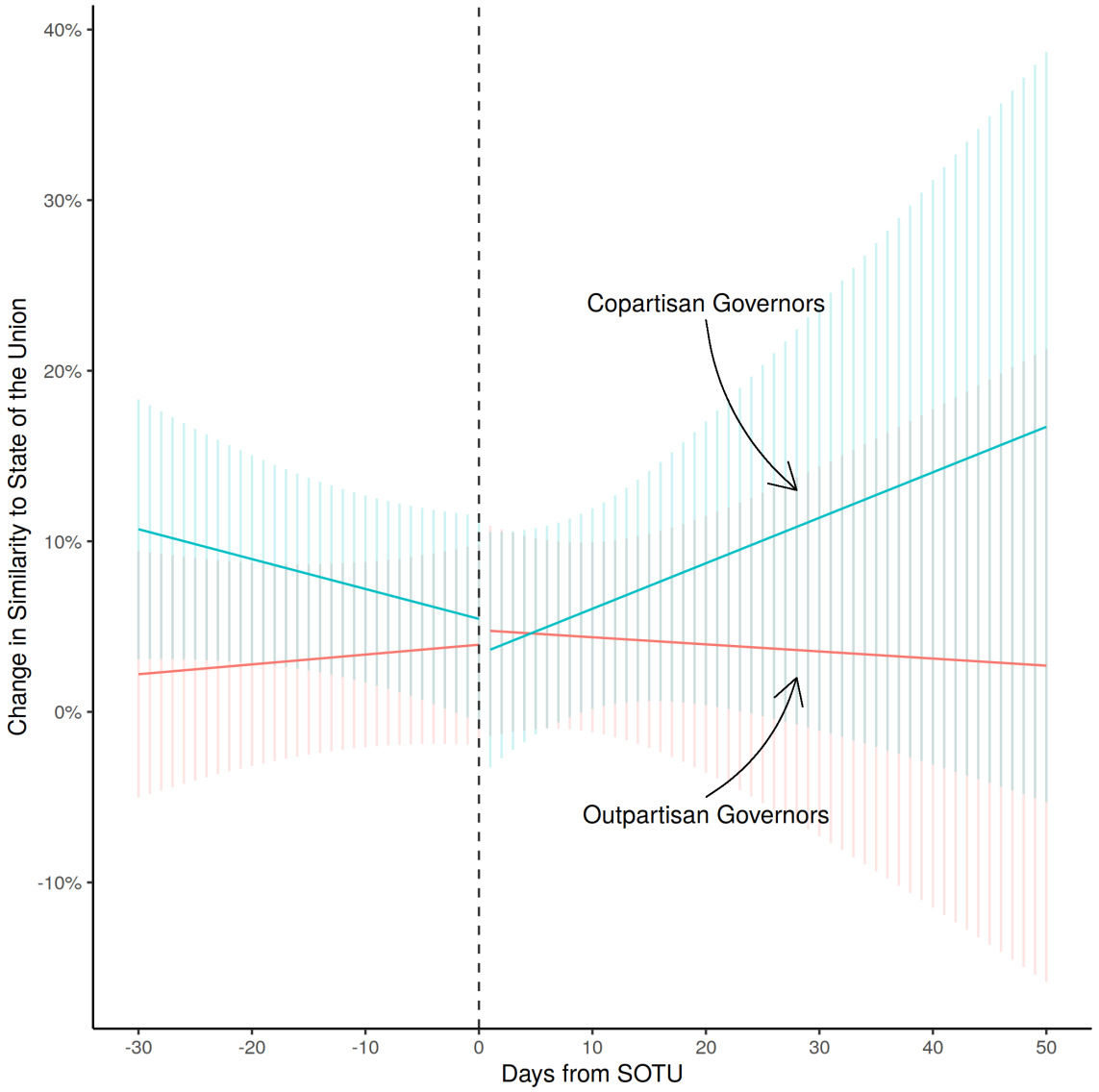
with the national agenda of the governor's party.

I leverage a unique feature of the SOTS addresses to study if the national agenda leads the local agenda. As I discuss in greater detail earlier, the SOTS address is a scheduled presentation, with dates typically set by distant, or even contrarian, state legislatures. Because this date is set in advance, it reduces the incentive and ability of governors to coordinate with the President (and the national party) in advance of the presentation of the national agenda. While the President may bring some initiatives from the Gubernatorial agendas into the SOTU – such as the inclusion of the idea behind the HOPE scholarship program from Georgia into Clinton's 1996 SOTU – there is a high cost of coordinating simultaneously with Governors and Congress, especially in the high stakes environment preceding the SOTU. These high costs serve to dissuade the White House from integrating bits and pieces from more than a handful of gubernatorial platforms.

My approach is to use a pre-post analysis to examine how similar the average SOTS is to the SOTU, before and after the SOTU is given. I construct an indicator, *Days Until SOTU*, which is difference of the date of the SOTS and the date of that year's SOTU, in days. The indicator is positive when the SOTS happened after the SOTU, and negative when the SOTS happened before the SOTU. I regress *Topic Similarity* on the *Days Until SOTU* to estimate the effect of the formal presentation of the national agenda on state agendas. I include in the regression a variable switch to allow for different slopes pre-post the SOTU, conditional on whether the Governor and the President are of the same party. I also include intercepts for region and year

candidate for the office of the President in 2012.

Figure 2.9: Local Politicians Use the National Agenda as a Guide



Note: Figure reports the relationship between time and the interstate SOTS similarity. Values are coefficients from a regression of similarities between the State of the State addresses given each year on a set of dummy variables for year. Each point is the coefficient estimate for that year. Similarity is a distance metric, which increases as the distance between two samples of text decreases. An ordinary least squares line of best fit is shown in red. The general trend has been for the State of the State addresses to become more similar to each other over the last six decades.

to capture unobserved heterogeneity.

Figure 2.9 reports the model's results for the Western region over the entire 1960–2012 period, and includes the states of Arizona, California, Colorado, Idaho, Montana,

New Mexico, Oregon, Utah, Washington,and Wyoming. Lines are fitted values. 95% confidence intervals are plotted in lighter color behind the fitted values. Figure 2.9 shows a clear break in the behavior of the President's copartisan Governors *after* the SOTU is given ($\beta_{\text{post}} = 0.18, p = 0.22$). This suggests that for States in the West, the SOTU led the topical composition of the SOTS, though the marginal coefficient is not significant.

Moreover, it is evident that the slope for out-partisans is steady over the entire pre-post period. This suggests that when the SOTU leads the SOTS, agenda change happens primarily among copartisans. Outpartisans, it seems, do not react to the nationalized agenda of the President. This is not necessarily to be expected. The "clearer choices" portion of the theory promulgated above would suggest that once the nationalized agenda of one party is made clear, the other party would begin to differentiate itself to provide clearer policy alternatives to voters. Across all the regions, it seems that at least among the SOTS and the SOTU, the only agendas changing are the copartisan ones.

Table 2.4 reports ordinary least squares coefficient estimates for the regression, broken out by region for interpretability. The coefficients are generally signed as we would expect. In fact, the effect of the SOTU leading the SOTS in the American South shows up as marginally significant at the $\alpha = 0.1$ level, with $\beta_{\text{post}} = 0.14$ ($p = 0.08$).[13] The fact that the agenda in the South is led significantly by the SOTU suggests that copartisans in the area are responsive to the nationalized agenda. An exception to

[13]The beta coefficient may be interpreted as "an increase in similarity of 14 points every 30 days," on average and all else equal.

the rule for the main effect show in the Northeast, which shows up as negative (not statistically significant).

Table 2.4: When the National Agenda Leads the State Agenda

	Dependent variable:			
	Topic Similarity			
	(5)	(6)	(7)	(8)
Time Difference (30 Day Increments)	−0.054	0.024	−0.059	0.037
	(0.040)	(0.051)	(0.037)	(0.031)
Gov. & Pres. Same Party	0.025	0.022	0.008	−0.034*
	(0.021)	(0.028)	(0.024)	(0.019)
SOTS After SOTU	0.004	−0.014	0.018	−0.002
	(0.023)	(0.032)	(0.033)	(0.022)
Time Difference∗Same Party	−0.038	0.050	0.053	−0.091**
	(0.051)	(0.065)	(0.048)	(0.043)
Time Difference∗SOTS After SOTU	0.003	−0.031	0.079	−0.006
	(0.075)	(0.094)	(0.080)	(0.063)
Same Party∗SOTS After SOTU	−0.026	−0.045	−0.010	−0.004
	(0.037)	(0.047)	(0.047)	(0.031)
Post–SOTU Copartisan Effect	0.114	0.024	−0.113	0.143*
	(0.112)	(0.133)	(0.127)	(0.079)
Constant	0.074	0.068	0.063	0.043
	(0.062)	(0.137)	(0.042)	(0.037)
Year FEs	Yes	Yes	Yes	Yes
Region	West	Midwest	Northeast	South
Observations	595	568	420	635
Adjusted R^2	0.313	0.220	0.436	0.472
F Statistic	5.299***	3.545***	6.150***	10.005***

Note: Entries are ordinary least squares regression coefficient estimates and standard errors, with clustering on year. The unit of analysis is the state–year. Values for each unit are constructed using election returns and document similarities from SOTS-SOTU dyads. The dependent variable is *Topic Similarity*. *p<0.1; **p<0.05; ***p<0.01.

2.8 The Nationalized Agenda and Election Outcomes

I have shown that there has been an increase in the nationalization of state policy agendas. I now test whether this increase varies with the increased nationalization of elections. In other words, I test whether the correlation between presidential and gubernatorial election results in a state is stronger when the agenda is more similar (*i.e.*, the *Topic Similarity* between the SOTS and SOTU address is higher that year).

In addition to the data on the similarity between the SOTS and SOTU addresses, I assembled the election returns from Congressional Quarterly. I used the election return data to create variables for the *Democratic Gubernatorial Margin of Victory* and the *Democratic Presidential Margin of Victory*. Both of these variables are calculated as follows:

$$\frac{\text{Democrat's Votes}}{\text{Democrat's Votes} + \text{Republican's Votes}} - \frac{\text{Republican's Votes}}{\text{Democrat's Votes} - \text{Republican's Votes}} \tag{2.3}$$

If the Democratic candidate got sixty percent of the vote and the Republican candidate got forty percent of the vote, then the democratic margin of victory would have been equal to 0.2. If the two candidates had received the exact same number of votes in the election then the democratic margin of victory would have been 0. If the Republican candidate got sixty percent of the vote and the Democratic candidate got forty percent of the vote, then the democratic margin of victory would have been equal to -0.2. The

margin of victory variables can range from -1 (a blowout win for the Republican to 1 (a blowout win for the Democrat). A 0.01 increase in this variable corresponds to a 1 percentage point change in the margin of victory.

The *Democratic Gubernatorial Margin of Victory* is the dependent variable for my analysis. I test whether the presidential election results predict gubernatorial elections conditional on the similarity between the state policy agenda and the national agenda (which I measure by comparing the the SOTS and SOTU addresses). To conduct that test I include variables for the *Democratic Presidential Margin of Victory*, the *SOTU-SOTS Topic Similarity Score*, and an interaction between the two variables. I also include dummies *Republican Incumbent in the Election* and *Democratic Incumbent in the Election* to control for the effect of incumbency on outcomes.

Note that my measure of *Topic Similarity* comes from before the election results are held. The SOTS speeches are usually given in the first quarter of the year, 8–10 months before the election. I am not measuring using the speeches of the winners after the election, which would raise major concerns about endogeneity (per the concerns of Montgomery, Nyhan, and Torres 2018, e.g.); instead, this variable is measured pre-treatment. Further, the decision to give a speech is not something the governor can control. All 50 states have constitutional mandates that the governor gives a State of the State address. This again helps minimize concerns about endogeneity because governors cannot choose to simply not give a speech.

The interaction term – *SOTU-SOTS Topic Similarity Score* $*$ *Democratic Presidential Margin of Victory* – is the key variable for my test. A positive coefficient on the variable would indicate that the presidential elections are more predictive of the

gubernatorial elections when the state policy agenda is closer to the national policy agenda. My full specification is as follows (where s indexes the state and t indexes the year):

$$\begin{aligned}&\text{Democratic Gubernatorial Margin of Victory}_{st} = \\ &\alpha_s + \gamma_t + \beta_1 \text{Democratic Presidential Margin of Victory}_{st} + \beta_2 \text{SOTU-SOTS Topic Similarity}_{st} + \\ &\beta_3 \text{Democratic Presidential Margin of Victory}_{st} * \text{SOTU-SOTS Topic Similarity}_{st} + \\ &\beta_4 \text{Republican Incumbent in Election}_{st} + \beta_5 \text{Democratic Incumbent in Election}_{st} + \eta_{st}\end{aligned} \tag{2.4}$$

As Equation 2.4 shows, I include indicator variables for the partisanship of the incumbent, with open seats serving as the baseline. I also include fixed effects for states (i.e., α_s) and years (γ_t). One concern is that there are secular trends in both the nationalization of elections (Hopkins 2018) and state policy agendas (see Figure 2.8), which might lead those two things to be artificially correlated. I include the year fixed effects to guard against this. I use ordinary least squares regression to estimate the model in Equation 2.4. I estimate the model for the full sample and also for the subsets of the data based on whether the gubernatorial election is held in the presidential election year. Table 2.5 displays these regression results.

Table 2.5 shows that gubernatorial and presidential elections are correlated. The positive coefficient on *Democratic Presidential Margin of Victory* shows that there is a correlation between the presidential and gubernatorial election results even when the policy agenda expressed in the SOTU and SOTS are completely different (*i.e.*,

Table 2.5: Agenda Similarity Moderates Gubernatorial Vote Share

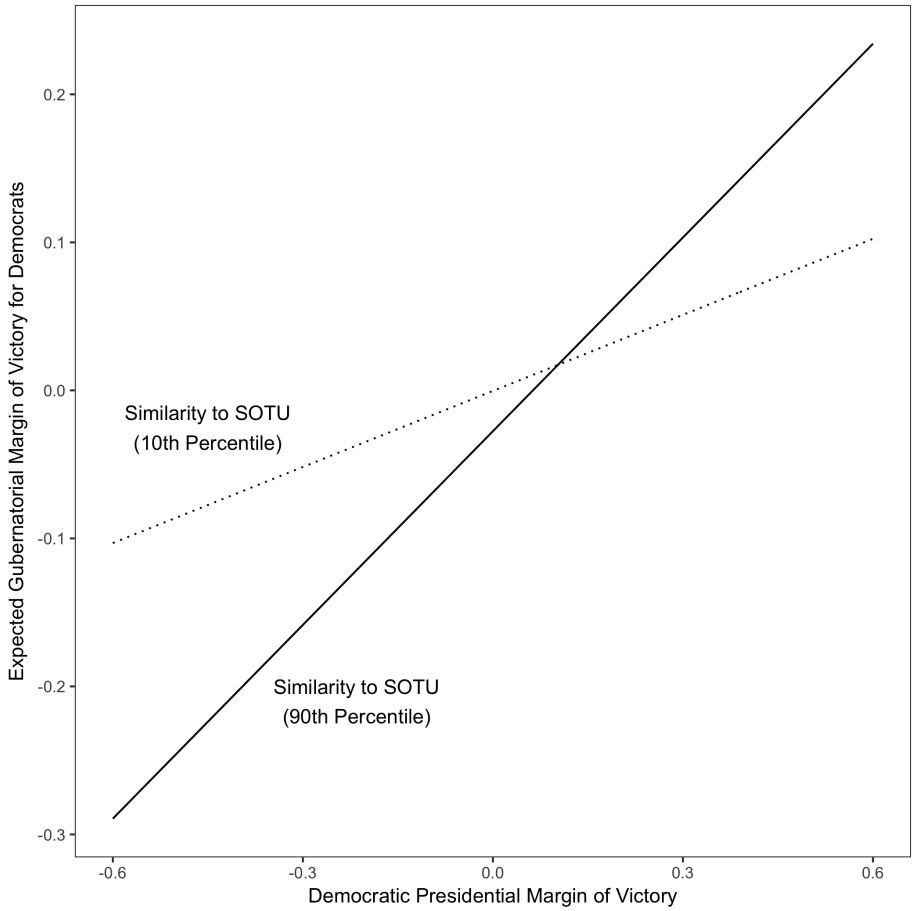
	Democratic Gubernatorial Margin of Victory		
	(9)	(10)	(11)
Democratic Presidential Margin of Victory	0.162	0.313*	0.092
	(0.103)	(0.166)	(0.132)
SOTU–SOTS Topic Similarity Score	−0.077	−0.002	−0.080
	(0.080)	(0.129)	(0.099)
Dem. Pres. Margin of Victory ∗ Sim. Score	0.751**	1.100	0.759*
	(0.357)	(0.664)	(0.430)
Republican Incumbent in the Election	−0.160***	−0.154***	−0.169***
	(0.019)	(0.037)	(0.023)
Democratic Incumbent in the Election	0.122***	0.191***	0.095***
	(0.019)	(0.032)	(0.025)
N	714	188	526
State FEs	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes
Adj. R-squared	0.32	0.40	0.29
Election Type	All Years	Presidential	Off-Cycle

Note: Entries are ordinary least squares regression coefficient estimates and standard errors, with clustering on year. The unit of analysis is the state–year. Values for each unit are constructed using election returns and document similarities from SOTS-SOTU dyads. The dependent variable is the Democratic Candidate's Margin of Victory. *p<0.1; **p<0.05; ***p<0.01.

the SOTU-SOTS Topic Similarity Score equals 0). And, as I would expect, this correlation is stronger for gubernatorial elections occurring in presidential election years (column 2) than for those occurring in non-presidential election years (column 3). The key result, of course, is the interaction term. The large positive coefficient on the interaction term shows that the relationship becomes much stronger when the State of the Union and the State of the State addresses are more similar. In other words, when the state policy agendas are more nationalized (*i.e.*, the president and governor are talking about similar topics), gubernatorial elections are more nationalized (*i.e.*, the presidential election results more strongly predict the outcome of these elections).

The coefficients on the control variables point in the expected direction and provide reassurance that the regression is reasonable. These variables show that the party with an incumbent in the race does about 10 to 15 percentage points better than they would if it were an open race.

Figure 2.10: Greater Agenda Similarity Drives Greater Nationalization



Note: The dotted line gives the relationship when the Topic Similarity score equals 0.013 (which is the 10th percentile of what we observe in the dataset). The solid line gives the relationship when the Topic Similarity score equals 0.366 (which is the 90th percentile of what we observe in the dataset). The figure demonstrates how similarity to the national agenda moderates the expected margin of victory.

Figure 2.10 illustrates the substantive significance of these results. The x-axis is different values for the democratic margin of victory in the presidential election in

the given state. The y-axis then gives the Democrat's expected margin of victory in the gubernatorial election. The two lines report the expected gubernatorial return as a function of the presidential return, conditional on high and low values of the SOTU-SOTS Topic Similarity scores. The dotted line gives the relationship when the Topic Similarity score equals 0.013 (which is the 10th percentile of what I observe in the dataset). And the solid line gives the relationship when the Topic Similarity score equals 0.366 (which is the 90th percentile of what I observe in the dataset). When elections are nationalized, I would expect a positive slope in any regression, and that is what I see for both lines. The slope, however, is much steeper for the solid line, indicating that such "high-similarity" elections are more nationalized. In other words, when the topics in the State of the State are more similar to those covered in the State of the Union, the presidential election more strongly predicts gubernatorial election results.

2.8.1 Disentangling Electoral Expectation from Agenda Moderation

Because political elites may adjust their strategies to account for trends in political behavior and electoral expectations (Butler and Nickerson 2011; Bergan 2009), one might be concerned that these results are driven by partisan anticipation of the next cycle's policy issues—a governor's partisan electoral demand effect. For instance, presidents and governors may coordinate on their agendas in anticipation of an expected election result in a swing state, in order to drive vote outcomes in their

favor. Such a phenomenon would, in the reductive context of a regression, produce an endogenous result for the interaction between agenda similarity and presidential vote share. I might observe endogeneity even though agenda similarity is measured eight months prior to any electoral outcome, because of electoral expectations.

It is possible to employ regional agenda similarity as an instrument which accounts for any agendas that are driven by expected vote outcomes, in order to recover a better estimate of the effect of the agenda on nationalization. The logic which enables the instrument is as follows. If a governor coordinates with the president (or the party) on their agenda speeches for partisan reasons, I would also expect to see coordination between other copartisan governors and the president. Therefore, I would expect to see coordination shocks among copartisans, and these shocks may be used to indicate when agenda similarity is driven by partisan influence. Meanwhile, regional agenda trends should not affect gubernatorial vote shares except through the gubernatorial agenda of that state.

I can get greater precision out of these shocks by looking for regional shocks. Governors must deal with local issues, in addition to national issues, in agenda speeches. Local issues often reach across state borders, creating regional issues which drive the idiosyncratic political processes I observe in the South, North, Midwest, and so on (Key 1950). The agenda similarities of other states in the region, therefore, will produce a more precise instrument than would a national approach. The measure is based on the average SOTS–SOTU similarity of copartisan speeches, excluding the speech itself, within any region (North, South, etc.) for the year. Let $\tilde{\theta}_s$ denote the

regional agenda similarity among copartisans for any speech s's region:

$$\tilde{\theta}_s = 1 - \frac{\sum_{i \in \mathcal{R}} \hat{\theta}_i}{\mid \mathcal{R} \mid}, \ s \notin \mathcal{R}, \tag{2.5}$$

where $\mathcal{R}$ is the set of speech similarities in s's region and year. The variable $\tilde{\theta}_s$, then, is an instrument for the degree to which a SOTS speech shares "exogenous" agenda similarity with the SOTU. A score of 1 suggests that the estimated agenda similarity $\hat{\theta}_s$ incorporates no partisan electoral demand effect, while a score of 0 suggests that the estimated agenda similarity is entirely driven by an electoral demand effect. Scores in the middle suggest varying electoral demand effects. I use two-stage least squares regression to recover the local average effect of the interaction term (Angrist and Pischke 2008).

Table 2.6 reports coefficient estimates from the instrumental variables approach.[14] The coefficients for the incumbency dummies remain signed and sized appropriately; Gubernatorial incumbents do well when running for reelection.

Coefficient estimates for the interaction term of interest in models 4, 5 and 6 are large and positive; the moderating effect of agenda similarity is present and strong; though it only achieves statistical significance in models 4 and 5. In those models, the coefficient on the democratic presidential margin of victory is now small and not statistically different from 0.[15]

[14]In order to test the assumption of exogeneity of the instruments, a test of the over-identifying restrictions was performed (Hausman 1983). The test failed to reject the hypothesis that the instruments were exogenous for models 5 and 6, but not for model 4 ($p = 0.04$). This failure to reject for model 4 may suggest inconsistent coefficient estimates.

[15]Even in model 6, where the coefficient on the democratic presidential margin of victory is statistically significant, it is much smaller than the effect on the interaction term. For most levels

Table 2.6: Instrumental Analysis of the Effect of Agenda Similarity Elections

	Democratic Gubernatorial Margin of Victory		
	(12)	(13)	(14)
Democratic Presidential Margin of Victory	−0.141	−0.108	−0.198*
	(0.094)	(0.162)	(0.116)
SOTU–SOTS Topic Similarity Score	−0.016	0.056	0.015
	(0.071)	(0.096)	(0.093)
Republican Incumbent in the Election	−0.149***	−0.133***	−0.156***
	(0.017)	(0.031)	(0.021)
Democratic Incumbent in the Election	0.147***	0.183***	0.124***
	(0.021)	(0.028)	(0.027)
Dem. Pres. Margin of Victory ∗ Sim. Score	1.470**	3.470**	1.210
	(0.736)	(1.660)	(0.836)
N	690	188	502
Adj. R-squared	0.18	0.33	0.15
Election Type	All Years	Presidential	Off-Cycle

Note: Entries are two-stage least squares regression coefficient estimates and heteroskedasticity-consistent standard errors (2SLS-adjusted). The unit of analysis is the state–year. Values for each unit are constructed using election returns and document similarities from SOTS–SOTU dyads. The dependent variable is the Democratic Candidate's Margin of Victory. The instrument is $\tilde{\theta}_s$, the average regional SOTS–SOTU similarity among copartisans in the year. *p<0.1; **p<0.05; ***p<0.01.

2.9 Discussion

In recent years, presidential election results have become increasingly predictive of the outcomes for elections to other offices. There is evidence that both congressional and state level offices have nationalized (Chubb 1988; Rogers 2016). I explored in this paper whether the changes in the policy agenda at the state level can help explain, in part, the nationalization of gubernatorial elections. Though a perfect analysis would rely on individual-level voter data, it is not possible to conduct such an analysis at the historical and data scale of the State of the States corpus.

of similarity observed in the dataset, there is a positive relationship between how Governors and Presidents do in the election.

The agenda is important because voters consider agenda issues when deciding how to vote (Lupia 1992). If the same issues are important at both the state level and the national level, and party is a strong predictor of a politician's positions (Poole and Rosenthal 2007), then voters should be likely to choose candidates of the same party at both levels. If state policy agendas are becoming more nationalized (*i.e.*, the policies are more in line with the issues at the national level), then I would expect the election results to become more nationalized too.

To test whether changes to the state policy agenda correspond with the nationalization of gubernatorial elections, I gathered the State of the State (SOTS) addresses. My data cover the period from 1960 to 2016. I combined the State of the State addresses with a corpus of State of the Union addresses. I applied topic modeling to the SOTS and SOTU addresses and used the information to carry out several tests. I analyzed the similarity between these SOTS speeches over time, and compared them to the presidential State of the Union (SOTU) addresses from the same period.

My results show two major findings. First, the policy agenda at the state level has become nationalized. I observe that SOTS addresses are increasingly covering the same topics. Moreover, the SOTS addresses are increasingly covering the same topics that are covered in the SOTU addresses.

Second, the nationalization of the state policy agendas corresponds to a nationalization of gubernatorial election results. My regression results show that the correspondence between the presidential and gubernatorial election results in a given state increases in strength as the SOTS and SOTU addresses focus on the same topics.

Some of my secondary findings include a descriptive analysis of party and regional

trends. These secondary findings suggest while there generally has been a trend in the increase in similarity between agendas over time, the Southern Democrats are largely responsible for the great increase in average similarity in the 1990s. This is interesting, as it coincided with the electoral transformation and political "blood sport" fomented by Newt Gingrich's *Contract with America*, and the backlash of the Republicans against the Clinton Democrats. This period is broadly attributed as the beginning of America's presently polarized and nationalized politics.

The findings have mixed implications for the state of American democracy. On a concerning note, the results suggest that federalism may not be functioning the way it is intended. Rather than focusing on local issues, the states are focusing their attention onto issues on the national stage. And, among other concerns, the results may mean that polarization at the national level is becoming a self-reinforcing process which state politics are being drawn into. Rather than serving as a bulwark against an increase in polarization, states may be contributing to a spiral towards increased polarization. Thus the results raise concerns about the choices that voters are facing in elections.

On the positive side, the nationalization of elections may indicate that voters are responding in a rational way to the choices that they do face. Evidence for nationalization is not evidence that voters have an irrationally motivated focus on party. Instead, it may be evidence that they are considering issues and voting accordingly. Because state policy agendas have nationalized and party is such a strong signal of a politicians' positions, voters now face similar choices at both the state and national level when they vote. Given those choices, it is not surprising that voters are

voting for candidates of the same party at both the state and national levels. This finding provides a more hopeful view of American voters (*c.f.* Achen and Bartels 2017), suggesting that voters evaluate candidates at all levels in ways that are consistent with their preferred policies. Whether this transformation has been the result of voters' direct consideration, the leadership of elites, or heuristic processes is the subject of further research.

Chapter 3

Can States Govern Effectively When Politics Are Nationalized?

In this chapter, I consider the question of whether electoral nationalization moderates the relationship between divided government and legislative productivity in the states. It takes as its unit of analysis the State over time. Conventional wisdom holds that divided control of the government hinders the ability of our elected representatives to govern, but research on whether this is the case has been a mixed bag. In this paper, I introduce new evidence by testing whether divided government affected lawmaking in the States from 1960–2012. I then test whether nationalization of electoral behavior is a part of this problem, arguing that nationalization is related to the ability to govern because of its close relationship with polarization and the incentives of elected representatives.

I use gubernatorial agenda speeches to identify salient state political issues within each year, and then with the resulting issue salience codes, I use automated content analysis to identify salient laws passed in the states. I apply these codes to study whether nationalization and divided government affected lawmaking in the States from 1960–2012, using, defining the ability to govern as the percentage of laws passed in that year that were salient. I find a null effect of divided government on lawmaking ability.

The approach shows evidence consistent with the "Mayhewvian" null findings on divided government, while adopting a similar, salience-based measurement approach of studies that do find an effect.

I also find that while nationalization is *not* related to the ability of our state governments to take action on salient issues during times of divided government, nationalization of state legislatures has generally *decreased* the production of salient laws. This finding is a somewhat troubling. It suggests that our nationalized political environment has influenced the ability of state lawmakers to govern effectively, but not through the institutional arrangements we usually consider to be the problem. In fact, the findings suggest that behavioral factors driving lawmaker decisions may be more to blame for lawmaking defects than institutional ones. A secondary finding concerns partisan effects. It appears that productivity in nationalized contexts is partially contingent on the party in control of the governor's office; this effect perhaps reinforces the institutional power of the governor, but it also underscores the incentives for a party to entirely sidestep compromise when it has the ability to do so.

3.1 Introduction

Over the past 20 years, Congress has been heavily criticized for being unable to do its job. An oft-referenced landmark critique of Congressional dysfunction is Mann and Ornstein's *It's Worse Than It Looks* (2016), which chronicled gratuitous levels of ineptitude. Congressional dysfunction was perhaps best exemplified when Majority Leader Mitch McConnell voted against his very own bill to establish the Gregg–

Conrad Commission. The commission's purpose? To establish a bipartisan process for overcoming dysfunction. This has produced an environment where "half-measures, second bests, and just-in-time legislating are the new norm, as electoral, partisan, and institutional barriers limit Congress's capacity for more than lowest-common-denominator deals." (Binder 2017, page 239). The fiscal health of the country has been affected, and so has trust in the government.[1]

We see today, however, that even in divided times Congress is still able to get things done. In the wake of the Coronavirus pandemic, Congress negotiated and passed a landmark stimulus bill of more than 2.2 trillion dollars, which was swiftly signed into law by President Donald Trump. The bill was more than twice the size the 787 billion signed into law as part of the American Recovery and Reinvestment Act of 2009—and the ARRA bill's amount was famously noted for being a "non-planetary" amount by Obama adviser Larry Summers (Matthews 2020). The bill was signed into law in a record 11 days, and within 8 days the money it promised was being distributed. these are extraordinary times, and a short burst of Congressional productivity may not suggest that on average Congress is generally productive. The strongest evidence that Congress is dysfunctional to the level that we require an amendment to our constitution to remedy it, would have perhaps been a complete inability to get anything done on this issue. But, if these recent events are to suggest anything, it's that Congress can do its job when it is important.

David Mayhew argues *Partisan Balance* that dysfunctions that can seem constitution-oclastic are actually "nonexistent, short-term, limited, tolerable, or correctable" (2011).

[1]Congressional approval ratings were for a time at 8 percent (Riffkin 2014).

Indeed, unorthodox lawmaking has perhaps *empowered* Congressional leaders to continue to do their job—at least for now. As Sinclair (2016, page 261) notes, "changes in the legislative process can be seen as the responses of members to the problems and opportunities that the institutional structure and the political environment present as members individually or collectively pursue their goals of reelection, influence in the chamber, and good public policy." That seems pretty far away from the observation of V.O. Key, who in 1964 noted that "common partisan control of executive and legislature does not assure energetic government, but division of party control precludes it" (Key 1964, page 688). Scholars have contested Mayhew's findings, and the debate over how polarization has influenced our ability to govern is far from settled. The evidence can suggest that we are either doomed or able to sustain, depending on the interpretation.

In this chapter, I test whether divided government affected lawmaking, or the ability to govern, in the States from 1960–2012. Scholarly approaches have reached mixed results on whether divided government in the federal government affects lawmaking. A major reason for these mixed results is that analyses use inconsistent definitions of what a good indicator for "ability to govern" is. As Mayhew (2005, page 201) points out, the choice of whether to benchmark legislation passed against how much we would expect them to pass can affect the cardinality and substance of results. Binder (2017) suggests that a disagreement on the definition of lawmaking ability is the major reason for these mixed results. Binder's suggestion is that we should measure lawmaking ability as a function of productivity on popularly *salient laws*, a methodology which stands in opposition to Mayhew's *important laws* method, which

uses productivity on laws experts see as impactful. Under Binder's methodology, we see that divided government *does* affect lawmaking ability, while under Mayhew's methodology, divided government *does not* affect lawmaking ability.

I take Binder's *salient laws* approach and show that in the States, divided government does not affect lawmaking ability. I use gubernatorial agenda speeches to identify salient policy issues within each year, and then I use those issue salience codes to identify salient laws passed in the states. I define the ability to govern as the percentage of laws passed in that year that were salient. It is important to point out that the representation of the agenda I use is compositional, not ideological—I am interested in identifying *salient* issues, not the framing or directionality of them. My coding picks up on the *presence and prevalence* of a given topic in agenda speech and legislation; it does *not* account for the goal direction of the agenda speech or the latent direction in which policy shifts.[2] Future work may integrate a richer computational understanding of goal direction; for now, the inferences involving ideology preclude the scope of this study.[3]

The study of divided government in the States is important because the States are the domain of important social issues that have constitutionally been proscribed from the purview of the federal government. States have a broad mandate to govern in places the federal government can't on social issues like education, abortion, and

[2]The *Comparative Agendas Project* similarly prioritizes an understanding of the topics on the agenda before the consideration of ideological direction.

[3]There is reason to believe the method picks up on goal direction spuriously. Policy topics are subject to ideological constraint, and therefore, the inclusion of a topic in the agenda might betray its ideological condition. For instance, if a Republican governor is more likely to include Republican issues in the agenda than she is to include non-Republican issues, then the mention of a topic is tantamount to declaring goal direction. Initial analyses suggest that this is indeed the case.

transportation. For instance, States can transformatively fund their educational systems through the creation of a guard-railed lottery; Governor Zell Miller of Georgia did just that in the 1990's, and by doing so created a public scholarship to support students attending state schools (college education increased in subsequent years by 8 percent; Dynarski 2000). States are laboratories of democracy that can tailor the shape of policy in a way the one-size-fits-all federal structure can't.

Consistent with Mayhew (2005), I find a null effect of divided government on lawmaking ability. This suggests that divided government is generally unrelated to legislative performance. The evidence I present is consistent with the Mayhewvian view of divided government, while adopting the salience-based approach of the Binder studies, which shows the opposite result—that divided government *does* reduce lawmaking ability by promoting gridlock. The result is also more broadly applicable in its scope, since it concerns a population of many American states, rather than the singular unit of the American federal government.

I take the analysis further by examining if another finding in Binder (2004) still holds in this new population of States. Binder proposed that a primary mechanism for gridlock is polarization. Increased levels of polarization are associated with an inability to govern because there is reduced incentive for lawmakers to compromise (Binder 2004), and the lawmakers elected share fewer ideological facets in common (*i.e.*, there is less centrist overlap between parties). The lack in state level data since the 1960s thwarts our ability to test for a relationship between polarization and legislative productivity in state houses over the past 60 years. The electoral process of nationalization, however, presents an opportunity to conduct such a test.

The relationship between nationalization and lawmaking ability is twofold. First, nationalization is closely related to, and intertwined with, polarization (Hopkins 2018). Constituents have become more attached to their parties (Green, Palmquist, and Schickler 2002), they have come to mirror their elected representatives ideologically (Barber and Mccarty 2013), and these attachments have provided increased incentives to their elected representatives to provide more polarized representation (Rogowski and Sutherland 2016; Webster and Abramowitz 2017). Second, the link between nationalization of elections and the agenda (see chapter 2) has created stronger incentives for elected representatives to eschew compromise. Agenda nationalization has created stronger partisan cues that elected officials may use in the course of lawmaking to signal their representative virtue to constituents (Levendusky 2009; Binder 2016; Theriault 2008), and voters are likely to hold them and their party accountable for their decisions (Butler and Powell 2014).

I therefore examine whether nationalization conditions the effect of divided government on legislative performance. I find that while nationalization is *not* related to the ability of our state governments to take action on salient issues during times of divided government, nationalization of state legislatures *has generally decreased* the production of salient laws. This finding is consistent with those of Binder (2017), who finds that polarization affects productivity through gridlock. In fact, the results provide evidence that congressional dysfunction may have more to do with electoral ramifications and ideology within the Congress than with the interplay between separate branches of government.

While the primary findings are somewhat encouraging, this secondary finding

is somewhat troubling. It suggests that our nationalized political environment has affected the ability of state lawmakers to govern effectively, but not through the institutional arrangements we usually consider to be the problem. In fact, it makes more prescient the prediction of Former Senate Majority Leader Tom Daschle that "If this continues, they're going to evolve, or devolve, into irrelevancy very quickly."

3.2 Why Divided Government May (Not) Matter

There are several reasons we would expect divided government to hinder lawmaking ability. Legislators in the executive's party are in the best position to pass their agenda when there is unified government. Under unified government, the legislators who have the electoral incentives to pass the executive's agenda have the institutional power to do so. Further, it is generally more likely that legislation is passed during periods of unified government. Divided government makes it more difficult to reach compromise and may lead to lower legislative output at both the national and state level (Binder 2004; Bowling and Ferguson 2001; Chiou and Rothenberg 2003; Krehbiel 1998; Rogers 2005).[4]

The stronger clarity of responsibility under unified government gives parties even stronger incentives to pass the executive's agenda (Powell and Whitten 1993). Under divided government, politicians can try to place blame on the other party. However, if the majority party fails to pass the agenda when they have unified control of government, voters will blame them. The influence executives exert on the bargaining

[4]But see Mayhew (2005) and Fiorina (1996).

process (Kousser and Phillips 2012) as both a party leader and a constitutionally empowered actor is strong, and the "pathological enactment logic" (Mayhew 2005, page 181) that drives parties to pass as many bills as they can under unified government suggests that the influence of the party will be strongest under unified government.[5] Resource constraints require that elected officials strategically choose which issues they deal with and which they ignore. Kousser and Phillips (2012), for example, observe that divided government interacts with an executive's political capital to affect the content and size of the agenda an executive will offer. Thus, strategic interactions causally entangle the language of gubernatorial agenda speech, the text of legislation, and the situational context.

The study of Congressional lawmaking is a perennial topic in the field of political science, captivating scholars who worked perhaps even before Woodrow Wilson's 1885 cornerstone work *Congressional Government*.[6] David Mayhew kicked off a new generation of interest in the topic by studying empirically the relationship between lawmaking and divided government in the post-war period. In *Divided We Govern* (1991), he executed a methodology that allowed him to determine not just how many

[5]Influence is, in this context, the ability of a party to enact its preferences. If unified government bring coordination, then it also improves the party's ability to enact its preferences. Thus, the narrow definition of influence is helpful to avoid the quagmire of influence broadly defined (Bachrach and Baratz 1962).

[6]Binder (2015, page 86) notes in an excellent review of the literature on legislative productivity that "before the American Political Science Association was launched in 1906, Frank Goodnow and other Columbia University professors contemplated creating an American Society for the Study of Comparative Legislation" to examine whether the lawmaking and regulatory bodies of the United States, and other nations, were adequately able to produce representative and timely policy. Sundquist (1968) and Chamberlain (1946) made significant contributions to empirical research on the topic, examining the relationship between divided government and the intensity of the power imbalance between the branches of government. Even in light of these contributions, Congressional institutionalism still cried out for a theory of coalitional government that could explain how the government could remain productive in times of party conflict (Sundquist 1988); the responsible party government theorists were prescriptively convincing, but their analyses lacked positivity.

bills were passed, but how many *important* bills were passed each session, by using a two-stage process to identify landmark bills. The process combined "contemporary" evaluations of importance with "retrospective" evaluations of legislative importance by policy experts to cull a set of important laws over the period from 1946 to 1990. He then regressed the number of laws passed in each session on whether the government was divided, and found that divided government was unrelated to productivity. In later updates for the period 1991-2012, Mayhew continued to find that the overall verdict on differences between conditions of party control seems to be the following: no effect on legislative volume, some effect on legislative content, a decisive effect on levels of conflict, and a strong effect on the number of congressional investigations.

From his work grew renewed scholarly debate to explain legislative performance (Binder 2004; Edwards III, Barrett, and Peake 1997, *e.g.*), culminating in Krehbiel's (1998) formal explanation for why separation of powers matters a whole lot less than the interaction between agent policy preferences and institutional rules. A key implication of the "pivotal politics" model is that divided government should produce gridlock (and thus, decay of legislative productivity) because parties have actors of opposite ideologies, and those ideologies interact with institutional pivots to prevent lawmaking.

Binder famously finds in *Stalemate* (2004) that the gridlock in fact does occur under divided government, and that in addition to polarization, it is conditional on *issue salience*; gridlock on salient issues has increased roughly 20 points since 1947 (this trend does not relate to the period of decreased party organization in the 60s and 70s). Generally, Binder finds that unified government decreases frequency of gridlock,

while narrower ideological centers and larger bicameral differences increase gridlock.

The effect of these studies has been to muddy the effect of divided government, and in an updated analysis of *Stalemate*, Binder (2017) throws kindling on the fire by pointing out the renewed role of polarization in gridlock, polarization that had greatly increased since she penned the original version. Electoral incentives for minority party organizations to play a more confrontational role have increased, while the costs of filibustering have declined (Smith 2014). Electoral competitiveness makes a difference has brought control of pivots within reach for both parties, who have incentives to invest heavily and tip the hand in their favor; they may make the risk-reward decision to "dig in now," and reap the rewards later (Fiorina, Abrams, and Pope 2005; Lee 2013). Binder (2017) concludes that "Mayhew may well be correct," but that "even so, we are left in the meantime with a national legislature plagued by low legislative capacity." Some scholars, such as Nolan McCarty, conceive the dysfunction of the Congress to be so bad, owing to such issues, that he suggests they will soon demand a change of our constitutional framework (McCarty 2016).

3.2.1 Federalized Governance, Parties, and Nationalization

A saving grace of the present situation in the national legislature is the system of federal government, which has the power to govern in a particular way each state, when the national government is unable or constitutional disempowered from doing so. The federal system as envisioned by Publius of the *Federalist Papers* preserved liberty, justice, and effective government through dual patriotism. Citizens would owe

loyalty to the national and local government simultaneously, protecting them from populist overreach, but also ensuring relevant governance.[7]

In fact, the ability of the States to govern themselves is thought to produce, in addition to fundamental republican protections, innovations that will over time more effectively solve collective action problems. Justice Louis Brandeis, in the wake of the Great Depression, noted that "one of the happy incidents of the federal system is that a single courageous state may, if its citizens choose, serve as a *laboratory*; and try novel social and economic experiments without risk to the rest of the country." (Brandeis 1932, italics mine). The configuration of our federal system of government was perhaps a blessing that would allow states to experiment with public devices that might help the nation overcome the crippling despair of the challenges it faced, which at the time was crippling economic despair.

---

[7]As Levy (2007) points out,

> With respect to the threat of military subversion, Publius maintains that a standing army at the center would be less dangerous to republican liberty than the alternative, standing armies in the several states, because "in any contest between the foederal head and one of its members, the people will be most apt to unite with their local government." "[T]he liberty of the people would be less safe in this state of things [with the states maintaining standing armies], than in that which left the national forces in the hands of the national government. As far as an army be considered a dangerous weapon of power, it had better be in those hands, of which the people are most likely to be jealous, than in those of whom they are least likely to be jealous. For it is a truth which the experience of all ages has attested, that the people are always in the most danger, when the means of injuring their rights are in the possession of those of whom they entertain the least suspicion" ([1788]2003:116; emphasis added). The reasons for this suspicion and jealousy, this natural likelihood that "first and most natural attachment of the people will be to the governments of their respective States," are so plentiful as to place the prediction "beyond doubt." People are more likely to have neighbors, friends, and family in state than in federal offices or employment. They are more likely to have reasonable hopes of such offices or employment themselves. State governments will tend to immediately felt [sic] local needs, whereas the federal government's primary business will seem far off and relatively unimportant. State politics will simply be more familiar and comprehensible. For these reasons, "the popular bias may well be expected most strongly to incline" toward the states ([1788]2003, 231).

It seems that today, however, our states no longer serve as laboratories of democracy. Instead, "the debates in states and even some localities have taken on a national hue, as state political conflicts become an extension of national conflicts, albeit with a different balance of forces [...] American federalism is no longer facilitating the expression of various issues and conflicts" (Hopkins 2018). Our federal system of elections, in which the same parties compete for office at both the national and local (state) levels, has produced a nationalized political system that hasn't been seen since before the post-war period. American federalism today looks a lot more like a set of proxy wars for nationalized policy interests than the Jeffersonian ideal of cooperative but locally focused dynamos (Grumbach 2018; Hertel-Fernandez 2016, e.g.). Our laboratories of democracy have become a giant political machine.

A myriad of factors have coalesced to make this the case. Voters today are more vote likely to vote on the basis of their partisanship (Shor and Rogowski 2018; Woon and Pope 2008), and they are much more likely to do so in the absence of information about the candidates (Jessee 2012; Lupia 1994). State legislative elections are famously deprived of adequate information about candidate, including even their names and actions (Kurtz, Rosenthal, and Zukin 2003; Rogers 2017), and as such, voters use party cues to adjudicate between candidates (Alvarez 1998). Meanwhile, national party positions get heavy coverage under conditions of high national polarization, because national positions are salient voters are more likely to be aware of national differences (Hopkins 2018).

The implications of highly nationalized power and policymaking are mixed. On one hand, nationalization may serve to renew the power of local self-government.

[8] On the other hand, nationalization may scale the ability of organized interests and ambitious actors to promote their own interests. Grossmann (2014) and Hertel-Fernandez (2016), for example, studies cross-state policy networks such as ALEC and finds that conservative organizations are much more able to achieve their policy objectives due to more capable leadership, enhanced incentives (they make state policy making "sexy"), and deeper investments over the last 60 years. This, the author argues, could explain why the Right has enjoyed a strong and increasing capacity for action across the states in recent decades whereas the Left has not. Ideologically affiliated, nationalized policy networks are the interest-group driven equivalent of the Uniform Commercial Code.[9] Even Schattschneider (1942, page 186) admitted that there are "doubtless some perils and problems implicit in party centralization [...] Centralization can probably be overdone."

The empirical evidence is cause for alarm. Boehmke et al. (2018) code 588 state policies from 1935–2014 and find that since 1960, the rate of policy sharing between states has steadily increased. The authors also divided policies into conservative and liberal categories, and found that the diffusion pattern of conservative policies

[8]E.E. Schattschneider (1942, page 182) believed, for instance, that "not a less active and influential national party organization, but a more powerful national party [would be] able to deny to the local boss all access to national patronage." The strong national interest may serve to reduce tyrannic local fiefdoms by winnowing the distribution of party resources and exerting the weight of the party in the interest of the voter, thereby restoring genuine local self-government. The majority of Responsible Party Government political scientists in Schattschneider's era agreed with him in spirit, though perhaps in varied flavors with respect to the mechanism (APSA Committee on Political Parties 1950). Stronger parties also provided voters with clearer alternatives and accountable longevity, which would serve to promote greater discernment among the everyday citizens.

[9]The significance of this cannot be undersold, as states still have significant influence on social welfare policy and have the right to enact legislation that would otherwise be rejected as unconstitutional. Consider, for example, the national debate over the Right to Die in recent decades. Such policy falls to the states to enact, but nationalized party interests took action to spin the issue in a way that might get them votes.

is different from the pattern of liberal policies. This suggests that policies may not be shared solely on the basis of how "innovative" they are. Rather, policies may be shared to support or promote specific nationalized interests. This is consistent with the findings from Chapter 2, which in summary suggest that as elections have nationalized, so have the policy agendas in states. The results of Boehmke et al. (2018) provoke the question of who wins, and where, when it comes to local lawmaking.

Whether these changes are due to broader electoral changes (Erikson, Mackuen, and Stimson 2002; Baumgartner and Jones 2010; Wlezien and Soroka 2012) or more nefarious trends (Grossmann 2014), it is evident that the tenor of lawmaking in the states has changed. The question this chapter answers is, "so what?" Does divided government change the ability of state legislatures to govern? Has the ability of state legislatures to govern changed as nationalization has surged? The preceding discussion altogether suggests the following testable hypotheses. The first is that there is no unconditioned relationship between divided government and the passage of important laws by state legislatures on average. The second is that nationalization does not condition the relationship between divided government and the passage of important laws, and it does not condition generally the ability of States to pass salient laws. My *a priori* expectations in the first case are a null effect due to Mayhew (2005). My expectations in the second case are less certain. On one hand, we may expect results consistent with the findings of Binder (2017), with the caveat that in this case, nationalized policy environments interact with institutional pivots to produce gridlock. On the other hand, we may expect nationalization not to affect legislator's abilities to get important things done—perhaps the stakes that might matter most. The results

will be informative either way.

3.2.2 Why Salient Issues Matter in Lawmaking

How to estimate legislative productivity has been a debate of diverging perspectives. On one hand, some approaches like Mayhew's count important laws (in the case of *Divided We Govern*), or relevant laws (in the case of *Partisan Balance*). On the other hand, approaches like Binder's (2004) posit the need for a denominator, or something against which to benchmark legislative performance. Binder argues that we should estimate Congressional performance as a measure of *quality*, and therefore we should differentiate between salient and non-salient laws.[10]

This chapter takes an approach that is in line with Binder's argument: in the States, we must also measure legislative performance relative to issues that are important to each State. Salience is important for the study of legislative productivity because it is the mechanism by which the relevance and legitimacy of lawmaking is established. Salient issues set the stage for partisan interests to demonstrate to the electorate how they are fighting for them, and how they are (or are not) getting results.[11]

---

[10]It is important to note the difference between retrospective and contemporaneous salience (Mayhew 2005), which has been the subject of some debate. Salience is retrospective when analysts today view an action or particular issue as salient, regardless of whether political actors at the time thought it was so. Salience is contemporaneous when analysts thought an action or issue was salient at the time it was being resolved, regardless of whether political actors later on thought it wasn't. This chapter focuses entirely on contemporaneous judgments of salience, as it uses gubernatorial agenda speeches to code for salience.

[11]It is worth noting that the way I measure legislative productivity purposefully emulates the studies of Binder and Mayhew. It is different from the approach of Kirkland and Phillips (2018), who utilize the time it takes to pass the Governor's budget as an indicator for productivity. The authors suggest that the meaning of divided government changes when the stakes are higher—when there are incentives for politicians to take action and meet deadlines, they will do so. Though Kirkland and Phillips (2018) find that divided government *does* affect productivity, the finding is complimentary to the Binder and Mayhew debate this paper comments on, because it takes a different measurement

I use State of the State addresses (SOTS) to code for salient issue in States over time. State of the State addresses are appropriate for the study of issue salience for several reasons. First, governors represent the public face of their party at the state level. Evidence shows that even in the United States, with its candidate-centered system, voters use party labels to infer information about politicians' policy priorities (Petrocik 1996; Walgrave, Lefevere, and Nuytemans 2009; Grynaviski 2010) and positions (Woon and Pope 2008). Governors are in a position to shape voter perceptions of their party's priorities because they hold a prominent political post to which state news media give ample attention. Voters are very likely to know who their governor is (Jennings 1996) and they are much more likely to hear the messages they share (Bennett and Iyengar 2008).[12] The governor is the state-level politician who holds the metaphorical megaphone.

Second, the gubernatorial State of the State address – akin to the presidential State of the Union address – is a procedural tool that allows governors to lay out their agenda without having to pass through any informational middlemen. The address, which is typically given to a full session of the state legislature, allows observers to set expectations in a high-information, low-bias environment. The governor's preferences are communicated to legislators, who, while they may impose their own valence on it, are reliably exposed to an unfiltered representation of what the governor wants.

---

approach.

[12]To the former point on gubernatorial recognition, Jennings (1996) finds that roughly 90% of the sample collected for the study are able to correctly name their governor. In contrast with the latter point on popular attentiveness, Bennett and Iyengar (2008) also find that voters today appear more likely to engage in biased information seeking, perhaps limiting the size of the network to which governors can speak.

It is also communicated to the press, which use the agenda to prioritize discussion and criticism of issue.[13] Governors credibly commit to their promises and expressed priorities because the press report on the speech immediately after it is given. These qualities suggest that gubernatorial agenda speech is an informative signal for the formation of beliefs and preferences by legislators.

Third, because the State of the State address is a legally mandated speech, the decision to give the speech is perhaps less subject to concerns about endogeneity. If the speech were not constitutionally mandated, one might worry that governor might only choose to express salient agenda issues only if they expect legislative success. Such a circumstance would produce artificially high levels of agreement between the expressed agendas of governors and policy outcomes. Constitutionally mandatory, regularly-scheduled State of the State addresses some selection bias because they *must* be given at a time that is determined well in advance of any exigent political circumstance. Thus, the State of the States are useful tools with which to computationally identify political success rates.[14]

A party's ability to get things done is an important part of how the individual politicians in the party are judged. Stokes (1963) argued that voters care both about parties' ideological positions and their valence. If a party has a poor valence, voters

[13]In fact, because the address feeds directly into press coverage, the address may also be used as an ancestral proxy for the measurement of salience through the news, as Epstein and Segal (2000) propose and apply.

[14]This method differs from other methods of issue coding, which have historically been used to determine political success. This approach instead ignores success, framing, or directionality, and simply aims to determine if an issue should be considered salient. Kousser and Phillips (2012), for example, derive a gubernatorial "batting average" using a method similar to that of Rosenthal (1990). The authors hand-code proposals in two years of State of the State addresses and use legislative session "wrap-ups," published by legislative watchdog organizations, to see if those proposals were successful.

will punish it at the ballot box. An important part of a party's valence is its ability to pass legislation. Cox and McCubbins (2005) use the damage to the Republican brand from the 1995 government shutdown to illustrate how legislative accomplishment (or the absence of it) influences the party's reputation (and voters' attitude towards the party). More recently, Butler and Powell (2014) use a series of survey experiments to show that voters reward politicians when their party passes the budget on time and do other things to maintain a reputation for getting work done.

There is well established precedent that it is salient issues in particular that inform voter evaluations of performance. Indeed, RePass (1971, page 400) argues that "by and large the electorate as at least one or two substantive issues in mind" when they vote, and voters are more likely to hold their representatives accountable to those issues (Page and Shapiro 1983). On the other hand, when salience is low, officials may not be aware of the preferences of their electoral bases, and therefore, they may follow ideological cues instead (Druckman and Jacobs 2006). It may even be salience that ensures democratic representation (Bawn et al. 2012).

Politicians cannot fully separate themselves from their party's reputation. Even if an individual politician has a reputation for being upstanding, they lose votes if their party is viewed as being unethical (Butler and Powell 2014). Voters rely on party brand name because party brands provide informational cues about how politicians are likely to act in office (Grynaviski 2010; Bartels 2000; Levendusky 2010). In other words, individual politicians are held accountable for their party's reputation above and beyond whatever else they might do. As a result, politicians have strong incentives to contribute to the building of their party's reputation. This incentive is strong

enough that Cox and McCubbins (1993, 2005) argue that legislators are even willing to give leaders the tools to apply pressure on them to vote against bills that they personally oppose. Butler and Powell (2014) find evidence in further support of this position, showing that state legislators believe that leaders are more likely to put pressure on them to vote for passage when a bill is perceived to affect the party's reputation. Co-partisan legislators thus have strong incentives to help pass items that the governor identifies in his state of the state address.

Further, parties have incentives to throw their support behind gubernatorial candidates who will support the party. Governors can exercise substantial executive control when given the power to do so Kousser and Phillips (2012). The party's strategic incentive is to support and elect someone who is at the party's median and will espouse their positions anyway. While there may be times the party is unsuccessful in electing their chosen candidate, it is reasonable to assume, given the empirical and theoretical study of parties (Karol 2009, *e.g.*), that parties are usually successful in doing so. Thus, the incentives of both governors and parties are aligned such that we would expect gubernatorial agenda speech to be harmonized with party preferences.

The agenda speech corpus allows salient issues to vary within States over time. This provides for richer variation, but it also is a truer representation of how issues are considered at the State level. For example, the creation of state lotteries to fund educational initiatives in the Southern region in the 1990's was a topic that became highly salient in election in that region, whereas issues elsewhere did not focus on the lottery. Another example is the importance of elderly care standards in Georgia in the present time. High profile nursing home issues have driven this issue to great

salience as public outcries and focus from the media have make it important.

By using gubernatorial agenda speech to measure salience, I am *certainly not* attempting to measure specific policy proposals, issue framing, priming, or persuasion.[15] Finding proposals is challenging because it can be unclear from a "distant reading" whether a governor is simply talking about an issue or proposing that the legislature take action on it. Coding the "liberalness" or "conservativeness" of detected proposals is also problematic. Every state is different, and the same policy proposal to legalize gay marriage, for example, may be extremely liberal in Alabama, whereas in New York we might consider the same policy to be status quo (Lax and Phillips 2012, *e.g.*). To demonstrate persuasiveness, one would need to show how governors change opinions of certain legislators, and a different unit of analysis would be best.

Some issues immediately come to mind for using the gubernatorial agenda speech to code for salience, and ultimately, to determine legislative performance. First, there is evidence that divided government can actually cause agenda expansion (Shipan 2006). This would induce a negative relationship between divided government and the ratio performance measure I employ, assuming some number of issues are not determined to be salient. Second, why should we even care about the governor's proposals when it comes to salience? Everyone is proposing things, and the governor

[15]Generally, this is the approach Mayhew (2005, page 220) takes. He states, "The content of legislation is of course a feature all by itself. In this book I have no enactments by,for example, whether they lean to the left or the right. 1991 through 2002, no one would be astonished to find a liberal drift under the Democrats in 1993-94 (UNI), a conservative drift under the Republicans in early 2001 (UNI), and a middling record otherwise (DIV). There is evidence for that case. In particular, the size of the Democratic tax hike of 1993 as well as the Republican tax cut of 2001 almost certainly owed to unified party control. Those were major legislative moves. The Family Leave and Motor Voter acts of 1993 (UNI) were quick achieve- ments of unified party control (Bush 41 had vetoed both during the previous Congress)."

is no different. Why not just use policy experts' evaluations? I direct the reader to the preceding discussion on why salience should be measurable through the speeches.

Gubernatorial strategy and personal rents are also a concern. They are problematic because they introduces the governor as a player with incentives that are (potentially) different from the collective party preferences—governors may want to extract personal rents, and they may hold ideal points that make the study of their influence untenable under the current data and empirical strategy. Instead, I assume that state of the state speeches, given by governors, are at least to some degree an instrument for party preferences. While the personalistic intentions of governors are almost a certainty, to parse these personalistic components from the platform of the party requires more thought and work. One continuation this work entails the inclusion of additional data, such as other personalistic and institutional indicators. Governors may introduce ideas in their speeches conditional on the strength of their electoral mandate (Kousser and Phillips 2012), which suggests that the speeches they make are a function of both their personal ambitions and the party line. Thus, the effect identified in this paper may include a personalistic component (which may not be monotonically related to the party platform).

3.3 Automated Analysis of Issue Salience in State Legislation

The State of the States corpus I use is the same one used in chapter 4, and includes 2,236 speeches from the period 1960—2008. See section section 2.5 and chapter 2 generally for detail on the development of the database, validation of the database, and descriptive statistics. I use these speeches to capture gubernatorial agendas, and I define an issue in a year for a State to be salient if it appears on the gubernatorial agenda. Any topic, if it appears in the speech, is coded as salient. This turns out to work quite well, as many governors have unique speaking styles and may apportion the lion's share of their speech to a particular issue (as Zell Miller of Georgia did for the State Flag in his 1990 speech). The average number of codes per speech using this method is 6.987 (of a possible 30). This is consistent with qualitative accounts of the State of the State speeches and other agenda setting activities, which generally do not run the gamut of all possible issues within a single year.

3.3.1 State Legislation Corpus

The second source of data is a corpus of legislation from the U.S. states, drawn from the data collected by Ash (2015). These session laws are the collection of statutes enacted by a legislature during a single session of that legislature, often published following the end of the session as a bound volume. The data on legislation consists of the full text of U.S. state session laws through 2008. The data go back to inception for most states, though the subset of the data relevant for this project is drawn from all

50 states, from 1960 to 2008. The "session laws" consist of the collection of statutes enacted by a legislature during a legislative session—published every year or every two years. These statutes may amend or repeal previous statutory provisions, or create new provisions. To paraphrase Ash (2015), these documents give the "flow," rather than the "stock," of legislation. Sometimes the laws include bills that failed or were vetoed. A team of research assistants reviewed samples and found that these practices do not change significantly within state over the time period.

The state session laws are in their raw form stored entirely as scanned images, which must be processed by OCR before being parsed. The same set of issues seen in the processing of the State of the States data apply to the session laws corpus as well. One additional challenge introduced by the session laws corpus is the need to reintroduce structure to the extracted text; to properly clean the corpus, it is necessary to segment the text into individual bills, acts, and resolutions. I augment the code used by Ash (2015), which produced 2.4 million clean sections, to yield an additional 730,411 sections, for a total of 3.2 million statute sections. The process uses a battery of regular expressions to detect common text markers for the start of a new statute.[16] Research assistants performed quality checks on samples of the statute segmenter for each state–year.

Although legislative language is arcane and abnormal, there is reason to expect the language used in statutes to be reflective of the language in agenda speech given by

[16]The prototypical procedure used to extract the text is detailed in Ash (2015). For example, indicators for new Chapters, Articles, or Titles include the line "CHAPTER" followed by a Roman numeral. Some states have their own standard indicators, such as "P.A." followed by a number to reflect a new "Public Act." The battery also searches for statute preambles (*e.g.*, "An act to...") and enacting clauses (*e.g.*, "Be it enacted that..").

the governors. There is a large literature in political science examining the process of drafting and enacting legislation (Tollison 1988; Jansa, Hansen, and Gray 2015, *e.g.*). State legislators can draft their own statutes, and most of them are trained to do so from attorney experience. They also delegate the task of drafting legislation to aides. Given the difficulty of crafting bills from scratch, legislators often borrow language from other legislatures or from interest groups. For example, Hertel-Fernandez and Kashin (2015) use text analysis to measure the influence of the conservative lobbying group ALEC on state legislatures. There are also non-partisan professional organizations such as the National Council of State Legislators, and the American Law Institute, which provide model legislation. These organizations provide information about which states have adopted particular provisions. Legislators pay attention to what other states are doing to make their state appear more competitive (Berry and Baybeck 2005). As such, whether through coordination or mimicry, it is reasonable to expect those who draft and enact legislative language to use text features that are shared with party agenda speech, present in the State of the State addresses.

The corpus I employ has 1,284,990 laws from all 50 states in the period 1960–2008. On average in each biennium, a state produces 1,102 laws (median= 749), with an inter-quartile range of 507 to 1,247 laws per state per biennium. See table 3.1 for detail on the number of laws passed.

Table 3.1: Session Law Counts and Issue Codes, by State and Year

State	Num. Laws	Avg. Laws Per Binm.	Avg. Num. Salient Codes	Avg. Num. Codes
AK	6,640	266	0.79	2.86
AL	21,096	844	0.98	3.37

. . . continued

State	Num. Laws	Avg. Laws Per Binm.	Avg. Num. Salient Codes	Avg. Num. Codes
AR	46,231	1,849	0.59	2.52
AZ	14,619	585	1.15	3.86
CA	91,820	3,673	0.66	2.82
CO	18,140	726	1.24	4.00
CT	8,850	354	0.68	3.44
DE	19,200	768	0.87	2.56
FL	33,577	1,343	1.64	3.70
GA	40,423	1,617	1.27	3.33
HI	12,233	489	1.35	3.04
IA	12,861	514	0.88	3.05
ID	17,188	688	1.38	3.22
IL	35,520	1,421	1.42	3.31
IN	16,327	653	0.86	3.10
KS	16,875	675	1.29	3.58
KY	10,073	403	1.49	3.91
LA	49,034	1,961	0.39	2.52
MA	28,686	1,147	1.02	2.69
MD	14,677	587	1.28	3.58
ME	43,141	1,726	0.76	2.47
MI	39,417	1,577	0.58	2.75
MN	10,863	435	0.91	3.37
MO	15,585	623	1.35	2.66
MS	43,092	1,724	0.90	2.66
MT	20,851	834	1.04	3.16
NC	32,805	1,312	2.06	2.76
ND	27,012	1,080	1.05	2.67
NE	28,493	1,140	2.56	3.17
NH	5,934	237	0.50	2.91
NJ	14,921	597	0.92	2.99
NM	6,929	277	1.22	3.22
NV	23,158	926	2.03	3.94
NY	66,863	2,675	1.20	2.73
OH	6,679	267	1.42	3.81
OK	17,662	706	1.03	3.48
OR	18,339	734	0.79	3.65
PA	659	26	2.46	4.35
RI	22,435	897	1.15	3.31
SC	42,994	1,720	0.70	2.33
SD	14,551	582	1.18	2.76
TN	54,386	2,175	0.73	2.57
TX	31,366	1,255	1.06	3.98
UT	12,236	489	1.13	3.82
VA	72,693	2,908	1.16	2.96
VT	6,725	269	0.95	3.46
WA	16,255	650	0.76	3.90

. . . continued

State	Num. Laws	Avg. Laws Per Binm.	Avg. Num. Salient Codes	Avg. Num. Codes
WI	20,557	822	0.47	2.29
WV	19,241	770	1.16	3.04
WY	10,343	414	1.01	2.47

Note: Values are reported counts and averages of salient law codes, produced by an automated content analysis approach to coding for salient laws. Salient topics are estimated on a corpus of state session laws, using a topic model trained on State of the States speeches by governors.

3.3.2 Automated Content Analysis of Salience

To code for salient issues, I use the same topic model used in chapter 4. See section 2.5.1 for detail on estimation and validation procedures. I first produce codes for each State and year by applying the topic model to the State of the State address in that year. When more than one address was present in a year, I coded pooled the text of them both and coded them as one document. This procedure produced 8,714 codes for the speeches, which I use as reference for salient issue codes in each State and year.

I then use the topic model to code the laws from each state. A law contains a topic if the predicted topic consumes more than 10 percent of the law. Any law may receive multiple codes, but the average number of codes per law is 1.05. I then code each law as 1. Salient if it has a code that shows up in the salient codes for that State and year (otherwise, I code it as 0. Not Salient). This procedure produced 1,318,622 codes for the laws, which I use as the numerator in my legislative performance measure.[17]

[17]Since I here compare one corpus of documents to another, separately generated corpus of documents – the SOTU, I use the delta statistic, developed in Section 1.5, to test the alternative hypothesis that the corpora were drawn from incomparable data generating processes. The statistic compares the observed distance between matched topics (probability distributions over their tokens) in each corpus to the expected distance under the null hypothesis that they were drawn from comparable data generating processes. The test fails to reject the null, which suggests the corpora

Ultimately the data are resampled to the biennium to account for legislative sessions, with the sum of laws (salient and non-salient) being the main summary quantity. See table 3.1 for detail on the average number of codes per law.

3.3.3 Divided Government, Nationalization, and Institutional Controls

The third source of data are codes for whether the state government was divided in that year, drawn from Klarner's (2003) partisan balance dataset. The data cover 1937 to 2011, and the subset of the data relevant for this project is drawn from all 50 states, from 1960 to 2008. The variables for divided government are dummies, which indicate for each unit partisan control of the governor's office and the legislature. For example, one dummy may indicate cases in which the legislature was controlled by the Democrats, but the Governor's office was controlled by the Republicans. The base category for the divided government dummy set is the case where Republicans controlled both the legislature and the governor's office. I include for the purposes of control several additional factors developed as part of the State of the States database, including the governor's legislative experience, whether the governor is in their lame duck year, whether the governor is in their last term, whether the governor is an incumbent, whether there is a gubernatorial election in the year, whether the governor is running, the governor's party, and the composition of the state legislature. These factors help to control for omitted variables bias that may affect the coefficient for

are generated by similar data generating processes. I proceed by comparing the two corpora on the basis of the SOTS topic model.

divided government (Kousser and Phillips 2012, for further work leveraging these controls).

The fourth and final source of data is the nationalization series from Chapter 2, which includes vote returns at the Presidential and Gubernatorial levels. In this chapter, I augment this database with further information on State Legislative elections (Ansolabehere, Snyder, Jr., and Stewart 2001, updated version) to provide a measure to account for the behavior of the legislature. Nebraska is omitted because of its unicameral government structure. Data are resampled to the biennial level by filtering to the first available year of data within any biennium.

3.3.4 Empirical Specification

To test the hypotheses developed in the preceding discussion, I estimate two equations, subscripting i for each State and t for each biennium:

$$Y_{it} = \alpha_i + \beta_{it} D_{it} + \gamma t + X_{it} + \delta Y_{i(t-1)} \epsilon, \text{ and} \tag{3.1}$$

$$Y_{it} = \alpha_i + \left(\sum_{k \in K} \beta_{itk} N_{itk} D_{it} + \beta'_{itk} N_{itk} \right) + \gamma t + X_{it} + \delta Y_{i(t-1)} + \epsilon, \tag{3.2}$$

where equation (3.1) estimates the effect of divided government on legislative performance, and equation (3.2) estimates the effect of nationalization on legislative performance. I define the dependent variable *Legislative Performance*, as the ratio of

salient laws passed to the total number of laws passed:

$$Y_{it} = \frac{\text{Salient Laws Passed in State–Year}}{\text{All Laws Passed in State–Year}}. \tag{3.3}$$

The term D_{it} is for divided government. I include a time trend captured by γ and state intercepts α_i. X_{it} is the matrix of aforementioned controls. Equation (3.2) includes coefficients for nationalization at each of three levels k, which are the *Nationalization* of the most recent gubernatorial, state senate (upper chamber), and state house (lower chamber) elections:

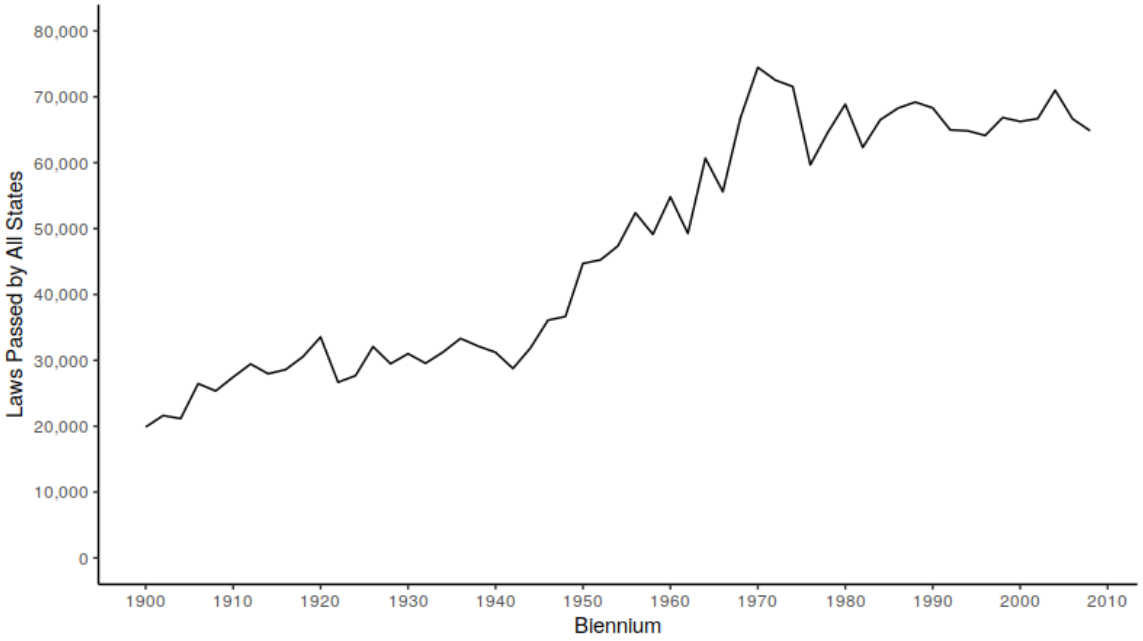
$$N_{itk} = 100 - \text{abs}\left(\text{Post}_k \text{ Democratic Margin}_{it} - \text{Presidential Democratic Margin}_{it}\right), \tag{3.4}$$

where the Post_k is governor, median state senate, or median state house. The coefficients of interest allow for positive tests the hypotheses. Specifically, we are interested if $\beta_{it} < 0$ in equation (3.1), and $\beta_{itk} \neq 0$, $\beta'_{itk} \neq 0$ in equation (3.2).

3.4 How Divided Government Fails to Affect Lawmaking

Figure 3.1 reports the number of laws passed by all states in each biennium. Even thought the data are subsetted to the period 1960–2008 for this analysis, the plot demonstrates how legislative productivity over the past 100 years looks, for context. The plot helps give us a sense of what the overall flow of "raw" productivity looks like.

Figure 3.1: Total Number of Laws Passed Over Time in the States



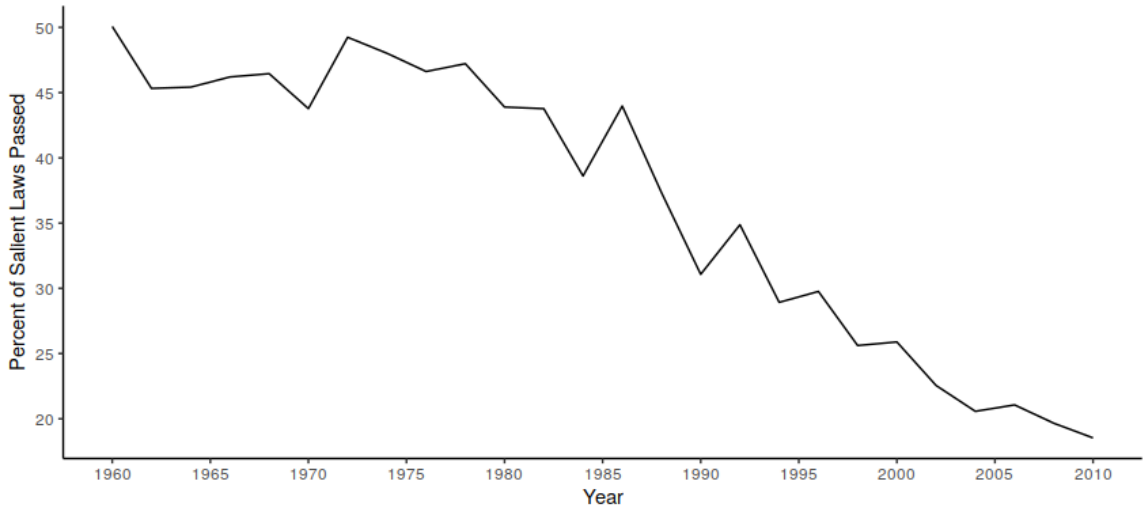
Note: Figure plots the total number of laws passed by the states in each biennium, from 1900 to 2008. The figure suggests that over time, the number of laws passed – legislative productivity – has increased steadily in the States, peaking in the post-Civil Rights Act era.

We can see, for example, that our nations most productive lawmaking years, from a topline perspective, were during and immediately after President Lyndon B. Johnson's Great Society initiative (1960s–1970s). This is consistent with the expectations of productivity scholars, who hold the post-Civil Rights Act period to be the most productive in American history since reconstruction.

We may also conclude from figure 3.1 that the number of laws passed overall has increased steadily over time, ending at around 75,000 laws produced in 2008. The ending observation is roughly 50,000 more than were produced in 1900, at the beginning of the time series.

Figure 3.2 reports the percentage of laws passed that are salient over the period 1960–2008. One major conclusion that may be drawn from the plot is that even

Figure 3.2: Salient Laws Passed Over Time in the States



Note: Figure plots the number of salient laws passed as a percentage of the total laws passed by the States in each biennium, from 1900 to 2008. The figure suggests that over time, the percent of salient law passed – legislative performance – has decreased in the States.

though the productivity of State legislatures has increased substantially over time, the salience of the laws produced – the performance – has decreased steadily over that same period.

The evident, steady decrease in the production of salient laws suggests that even though more laws are being passed today than there were sixty years ago, fewer and fewer of those laws are salient, per the codes from the gubernatorial SOTS addresses. The general implication of this trend, notwithstanding critiques of the measurement approach and endogeneity, is that officials are passing more laws, but more and more of them are simply "noise."

We might expect officals to pass more laws, even if they are not salient, to signal their value to constituents and claim credit for their re-election campaigns (Mayhew 1974), or; the party organizations in state legislatures may allow members to pass more peripheral "chum" in order to appease them and support the caucus's primary

interests (Cox and McCubbins 2005). As such, this trend may simply imply that legislative organization is becoming more amenable to the distributive needs of its members. In other words, "so what?" Such an argument would suggest disequilibrium in the system, however. If legislators have time to focus on unimportant issues, then they also have time to focus on service to the people they represent.

Generally, policymaking is part of a dynamic and multifaceted relationship between what the people want and how elected officials represent them. Short-term perturbations in the long-run equilibrium of productivity are to be expected, but the downward trend holds.The steadiness of this decrease, however, stands out, because we might perhaps expect more punctuated changes through the inclusion of new issue platforms from third parties (Carmines and Stimson 1989), or through rapid changes in the focus of the agenda (Baumgartner and Jones 2010). The general steadiness of this trend seems to suggest that productivity changes are more part of a more deliberate macro-process.

Another reason we might observe this trend is a disconnect between the salient issues proposed by the governor of one party, and the state house controlled by the other party. Simply put, the issues the governor cares about might not be the issues the house cares about, and the two just talk past each other. In fact one, the other, or neither may represent the actual set of issues that are salient to the people; voters may even encourage such a process, participating in partisan balancing to achieve moderated political outcomes, and thereby producing the illusion that salient lawmaking has decrased (Alesina and Rosenthal 1995, e.g.). These are a significant limitation of the approach I take, and for the purposes of this analysis, I assume

generally that elected officials have real and urgent incentives to take up the issues that matter most to their constituents. As such, while it is perhaps the case that demand for certain ideologies in policymaking can ebb and flow over time (Erikson, Mackuen, and Stimson 2002), we may conclude that the general downwards momentum implies a less representatively effective class of elected officials.

Table 3.2 reports the results of the ordinary least squares regressions of productivity and performance on divided government. Models 1 and 2 should be interpreted directly on the scale of dependent variable, in whole percentage points. For example, the coefficient for *Dem. Leg. Dem. Gov.* should be interpreted as the all-else equal average increase of 1.778 percentage points on salient laws passed. Models 2 and 3 should be interpreted as log-linear models. For example, the coefficient for that same dummy in those models should be interpreted as a 3 percent increase in productivity, on average and all else equal. The coefficients appear to be sized and directed appropriately, as we would expect. In fact, it is interesting that the coefficients capturing when there is divided government and a republican governor suggest a negative relationship; however, this relationship does not reach the threshold of statistical significance so I do not read into it.

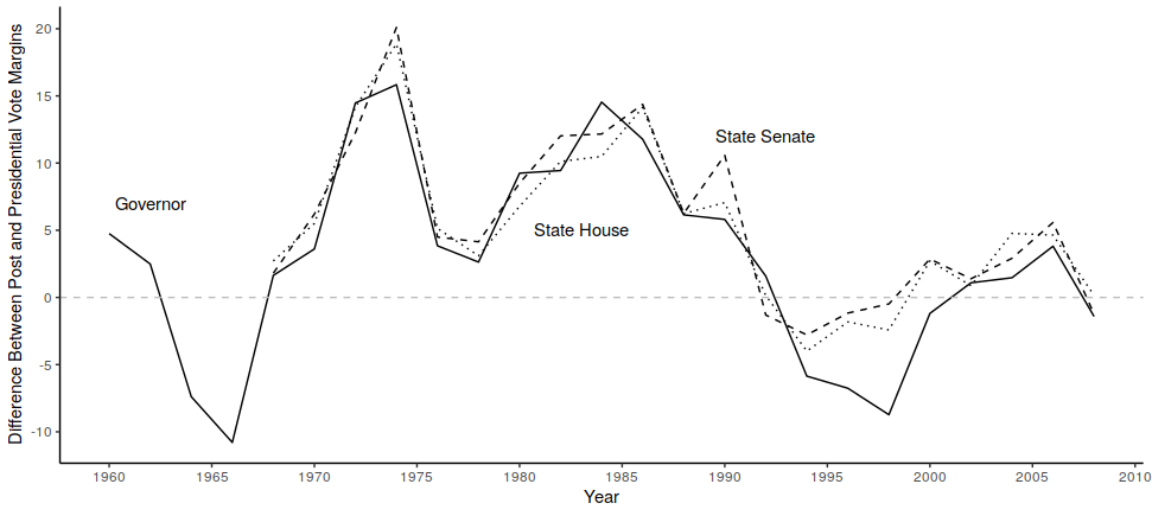
The major conclusion to draw from these models is that there is a null relationship between divided government, legislative productivity, and legislative performance in the States over this time period. This suggests that divided government in the States *does not* affect the ability of lawmakers to get the important things done; it also suggests that divided government does not affect the ability of lawmakers to generally produce laws. These results are consistent with the findings in Mayhew (2005).

Table 3.2: Effect of Divided Government on Legislative Performance

	Legislative Performance		Legislative Productivity	
	(1)	(2)	(3)	(4)
Dem. Leg. Dem. Gov.	1.778	1.199	0.031	0.061
	(1.775)	(5.750)	(0.046)	(0.142)
Split Leg. Dem. Gov.	1.262	4.369	−0.068	−0.137
	(1.882)	(7.693)	(0.063)	(0.200)
Split Leg. Rep. Gov.	0.032	−2.542	−0.039	0.026
	(1.967)	(4.126)	(0.053)	(0.089)
Rep. Leg. Dem. Gov.	−1.930	5.567	0.013	−0.135
	(1.590)	(10.466)	(0.039)	(0.256)
Dem. Leg. Rep. Gov.	2.725	−5.079	−0.007	0.137
	(1.794)	(7.153)	(0.048)	(0.162)
Gov. Election Year		−0.293		−0.052*
		(1.059)		(0.031)
Gov. & Pres. Same Party		1.109		−0.032
		(1.035)		(0.029)
Democratic Gov.		−11.565		0.188
		(13.806)		(0.330)
Democratic Leg.		14.207		−0.187
		(10.670)		(0.236)
Gov. Lame Duck Term		−1.258		0.038
		(1.376)		(0.038)
Gov. Lame Duck Year		5.165		0.011
		(4.219)		(0.059)
Gov. Leg. Exper.		−0.913		0.006
		(1.143)		(0.031)
Num. Budgets Gov. Mngd.		−0.301		−0.005
		(0.245)		(0.006)
Time Trend	−0.430***	−0.494***	−0.004***	−0.004***
	(0.036)	(0.045)	(0.001)	(0.001)
Lagged Salience	0.155***	0.139***		
	(0.028)	(0.030)		
Lagged Productivity			0.0004***	0.0004***
			(0.00005)	(0.0001)
Constant	35.524***	37.360***	6.775***	6.877***
	(5.503)	(6.060)	(0.108)	(0.113)
N	1,026	1,026	1,026	1,026
State FE	Yes	Yes	Yes	Yes
Adj. R^2	0.38	0.41	0.87	0.89

Note: Entries are ordinary least squares regression coefficient estimates and standard errors, with clustering. The dependent variables are legislative performance, which is the number of laws coded as salient divided by the total number of laws passed, defined in equation (3.3), and; legislative productivity, defined as the total number of laws passed (logged in the estimation). Models 3 and 4 should use log-linear interpretation. Models 1 and 3 are estimated without controls, and models 2 and 4 include controls. The base category is a Republican legislature and Republican Governor. *p<0.1; **p<0.05; ***p<0.01.

Figure 3.3: Nationalized Electoral Margins in Median State Offices



Note: Figure plots the median differences between the post and Presidential Democratic vote margins. A value at zero means that there was no difference between the Democratic vote margin for the post and the Democratic vote margin for the President (perfect nationalization). Values further away from zero suggest less nationalization for the median office. Median office results are determined by ordering within each year the elections occurring by the difference between the local and national vote margins and then selecting the median.

3.4.1 How Nationalized Environments Produce Gridlock

I now turn to the question of whether nationalization has affected the ability of our lawmakers to govern. I begin by engineering our notion of nationalization, specified in equation (3.4). I compute for all available gubernatorial, state senate, and state house elections the democratic vote margin. I then take the difference between the vote margin for the post and the vote margin for the President in that State and year. The result is a measure capturing the level of nationalization, where values closer to zero suggest higher levels of nationalization.

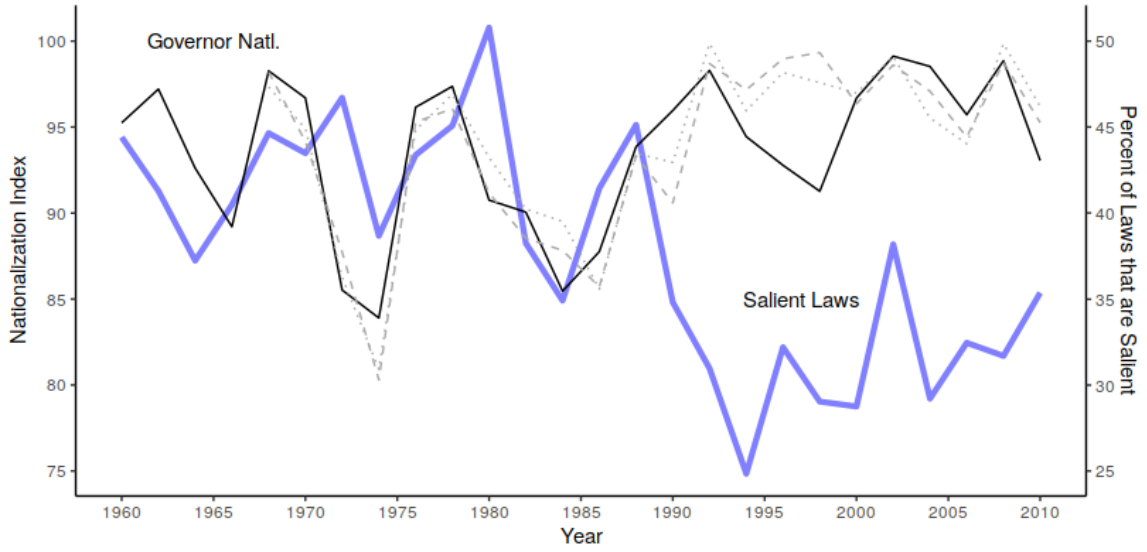
Figure 3.3 plots a nationwide roll-up of the result, using the median differences between the post and Presidential Democratic vote margins. Median office results are determined by ordering within each year the elections occurring by the difference

between the local and national vote margins and then selecting the median.

Two interesting conclusions jump out from the plot. The first is that in the time period proximal to Newt Gingrich's "Contract with America," which is often seen to be the root of today's political blood sport, state senate and state house elections became significantly more nationalized. This change is particularly remarkable when juxtaposed with the gubernatorial median over time, which varies much more broadly. The second conclusion is that the medians analysis for governors reveals perhaps less nationalization than a regression analysis would suggest. A regression of gubernatorial vote margin on presidential vote margin will still yield a positive and statistically significant coefficient (see table 2.3), but indeed, seeing the data laid out per the nation-wide median is contrastive.

Figure 3.4 plots the nationwide roll-up of the median differences against the percentage of salient laws displayed in figure 3.2. At the national level, there is no clearly discernible relationship, but the two do seem to have some common shocks. For example, from 1990–1994, both series appear to covary. The analysis of these data at the state–year level will greatly improve our ability to draw conclusions.

Figure 3.4: Nationalized Electoral Margins and Salient Legislation



Note: Figure plots the median differences between the post and Presidential Democratic vote margins against the percent of laws passed that were salient. A value at zero for the median differences series means that there was no difference between the Democratic vote margin for the post and the Democratic vote margin for the President (perfect nationalization). Values further away from zero suggest less nationalization for the median office. Median office results are determined by ordering within each year the elections occurring by the difference between the local and national vote margins and then selecting the median.

Table 3.3: Effect of Nationalization on Legislative Performance

	Legislative Performance *(cont.)*					
	(5)	(6)	(7)	(8)	(9)	(10)
Dem. Gov. Dem. Leg.	2.196	−12.094	2.612	−111.717*	−3.232	63.464
	(1.978)	(28.961)	(2.973)	(59.003)	(3.945)	(55.348)
Split Leg. Dem. Gov.	0.917	−7.777	−1.016	−127.634*	−0.700	40.868
	(2.076)	(41.264)	(2.952)	(71.805)	(3.577)	(69.346)
Split Leg. Rep. Gov.	1.378	−3.202	2.623	−119.388	−3.909	93.657
	(2.157)	(42.857)	(2.539)	(72.870)	(4.063)	(74.216)
Rep. Leg. Dem. Gov.	−2.608	−19.844	−4.393	−97.992	−0.829	31.539
	(1.764)	(36.868)	(2.828)	(78.759)	(3.084)	(62.557)
Dem. Leg. Rep. Gov.	2.353	−32.623	1.137	−104.580*	−0.071	18.596
	(1.996)	(33.944)	(3.060)	(57.922)	(3.681)	(61.348)
Gov. Nat. Score	0.023	0.019	0.083	−0.169	−0.014	0.258
	(0.062)	(0.156)	(0.099)	(0.196)	(0.104)	(0.467)
Sen. Nat. Score	0.118	0.355	0.121	0.190	0.153	0.696
	(0.115)	(0.231)	(0.180)	(0.306)	(0.168)	(0.468)
House Nat. Score	−0.342**	−0.738**	−0.200	−0.902**	−0.397**	−0.685
	(0.135)	(0.301)	(0.252)	(0.420)	(0.184)	(0.486)
Gov. Elec. Year		−0.544		−1.696		−0.368
		(1.116)		(1.625)		(1.701)
Gov. & Pres. Sm. Pty.		0.618		−0.638		4.004
		(1.205)		(1.497)		(2.478)
Democratic Gov.		−13.788		−22.750		3.233
		(14.916)		(17.679)		(36.230)
Democratic Leg.		12.835		12.963		−3.661
		(11.866)		(15.318)		(29.619)
Gov. Lame Duck Term		−0.910		−0.889		−1.910
		(1.458)		(2.362)		(2.471)
Gov. Lame Duck Year		3.593		8.700*		4.019
		(4.458)		(4.789)		(5.902)
Gov. Leg. Exper.		−0.894		−1.691		1.374
		(1.201)		(1.779)		(2.018)

	Legislative Performance					
	(5)	(6)	(7)	(8)	(9)	(10)
Num. Budgets Gov. Mngd.		−0.420		−0.099		−0.814**
		(0.269)		(0.448)		(0.390)
Time Trend	−0.514***	−0.516***	−0.547***	−0.498***	−0.441***	−0.471***
	(0.049)	(0.052)	(0.127)	(0.140)	(0.109)	(0.120)
Lagged Salience	0.129***	0.122***	0.026	0.022	0.102**	0.088
	(0.031)	(0.032)	(0.041)	(0.043)	(0.048)	(0.054)
Dem. Leg. Dem. Gov.∗Gov. Nat. Score		−0.041		0.375		−0.377
		(0.190)		(0.328)		(0.508)
Split Leg. Dem. Gov.∗Gov. Nat. Score		−0.285		−0.224		−0.475
		(0.242)		(0.395)		(0.599)
Split Leg. Rep. Gov.∗Gov. Nat. Score		−0.131		0.061		−0.104
		(0.316)		(0.403)		(0.624)
Rep. Leg. Dem. Gov.∗Gov. Nat. Score		−0.168		0.568		−0.678
		(0.236)		(0.376)		(0.548)
Dem. Leg. Rep. Gov.∗Gov. Nat. Score		0.101		0.281		−0.238
		(0.212)		(0.280)		(0.530)
Dem. Leg. Dem. Gov.∗Sen. Nat. Score		−0.217		−0.308		−0.625
		(0.318)		(0.535)		(0.529)
Split Leg. Dem. Gov.∗Sen. Nat. Score		−0.138		0.087		−0.353
		(0.466)		(0.658)		(0.803)
Split Leg. Rep. Gov.∗Sen. Nat. Score		−0.262		0.227		−0.872
		(0.397)		(0.739)		(0.620)
Rep. Leg. Dem. Gov.∗Sen. Nat. Score		−0.540		0.069		−1.156*
		(0.405)		(0.628)		(0.670)
Dem. Leg. Rep. Gov.∗Sen. Nat. Score		−0.542		−0.283		−0.506
		(0.379)		(0.584)		(0.642)
Dem. Leg. Dem. Gov.∗House. Nat. Score		0.426		1.250*		0.311
		(0.389)		(0.720)		(0.548)
Split Leg. Dem. Gov.∗House. Nat. Score		0.567		1.620**		0.378
		(0.438)		(0.708)		(0.760)
Split Leg. Rep. Gov.∗House. Nat. Score		0.398		0.960		−0.052

	Legislative Performance					
	(5)	(6)	(7)	(8)	(9)	(10)
		(0.502)		(0.843)		(0.720)
Rep. Leg. Dem. Gov.∗House. Nat. Score		0.983*		0.572		1.445**
		(0.561)		(1.038)		(0.719)
Dem. Leg. Rep. Gov.∗House. Nat. Score		0.734*		1.035		0.593
		(0.425)		(0.718)		(0.644)
Constant	59.964***	75.666***	45.983**	130.415***	70.745***	18.344
	(9.198)	(28.636)	(19.528)	(36.948)	(12.336)	(53.976)
Period	1960–2012	1960–2012	1960–1990	1960–1990	1992–2012	1992–2012
N	1,026	1,026	1,026	1,026	1,026	1,026
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.37	0.37	0.29	0.28	0.23	0.22

Note: Entries are ordinary least squares regression coefficient estimates and standard errors, with clustering. The dependent variable is legislative performance, which is the number of laws coded as salient divided by the total number of laws passed, defined in equation (3.3). Models 5, 7, and 9 are estimated without controls and interactions. Models 6, 8, and 10 include controls and interactions. *p<0.1; **p<0.05; ***p<0.01.

Table 3.3 reports the results of the ordinary least squares regressions of legislative performance on divided government and nationalization. They are estimated at the State–year unit of analysis, and as such use only the median margin difference for the state senate and state house nationalization variables. Models 7–10 break the data out into pre-1990 and post-1990 periods, to cautiously allow for regime effects owing to the "Contract with America" period. The models may be used to examine our two coefficients of interest.

The first coefficient set of interest reveals whether nationalization conditions the effect of divided government. The results are mixed, and largely depend on two factors: the party the governor, and time period. The coefficients suggest that nationalization in the state senate and the governor's office do not generally condition the effect of divided government, but nationalization in the state house does. In fact, given a Democratic governor, a 1 point increase in nationalization score results on average and all else equal in nearly a 1.3 point increase in the production of salient laws (ranging from 0.9 to 1.6, $p < 0.05$) across time periods. In the full time series cross sectional dataset, the dummy for a Republican governor with a split house also shows up as positive and significant, yet this effect seems to wash out in the broken out regressions, suggesting perhaps a confounded effect. This is a double edged result. On one hand, it suggests that when state legislators have the means to do so, they will still produce salient legislation, even when the environment is nationalized. On the other hand, however, it may also suggest that when parties have full control of the government, they may also overcome obstructive institutions meant to preserve supermajoritarianism.

The second coefficient of interest is the relationship between nationalization and legislative performance, separately from divided government. The models offer in this case a much more troubling result. Nationalization of state houses is significantly and negatively related to legislative performance. The coefficient suggests that the average effect of a 1 point increase in nationalization score is about a half of a point decrease in the production of salient laws, all else equal (ranging from -0.38 to -0.90, $p < 0.05$). This means a state that is highly nationalized (in the 90th percentile) is on average predicted to be 8 to 12 points less productive on salient laws than a less nationalized state (in the 10th percentile). The effect is robust to the inclusion of controls, it is specific to the present "post-Contract with America" time period. This suggests that more recently, nationalization of the State houses has decreased the ability of lawmakers to govern.

The finding that nationalization decreases salience is perhaps surprising, as Binder (2004) argues that variations in opinions within parties – often to the chagrin of disagreement between parties – can increase the difficulty of reaching agreement, and avoiding gridlock. When elections are nationalized, we would expect to see the party, and its held offices, to coalesce around a common agenda and theme. This would suggest the production of more salient laws; instead, we see fewer produced. It is worth noting that I do not include in these regressions differences in the procedures and norms of the state houses and senates, which, Binder (2004) claims, may serve to produce varied distributions over policy preferences and therefore promote institutional disruption of productivity.

3.5 Discussion

When the late Senator Tom Coburn of Oklahoma resigned in 2014, he was asked by the press about the dysfunction in Congress owing to polarization and the nationalized environment in which Americans find themselves. "That's why I left. You couldn't do anything anymore," he retorted. Political science accounts of politics in the States like Hopkins (2018) make it clear that the State environments are no different. This is especially troubling because the States are where nearly half of all government spending in America happen, and where decisions are made on a broad number of social issues that the federal government is not permitted to govern.

The conventional wisdom holds that divided control of the government hinders the ability of our elected representatives to govern, especially in the context of a nationalized, polarized environment. In this chapter, I provide evidence that this is not the case. Divided government in the states, in the period from 1960–2008, is unrelated to legislative performance. My findings on nationalization, however, are reason for caution. The results suggest that the nationalization of state houses has decreased significantly State legislative performance. This suggests that nationalization is related to a hindrance of the ability of our state officials to govern. A secondary finding concerns partisan effects. It appears that productivity in nationalized contexts is partially contingent on the party in control of the governor's office; this effect perhaps reinforces the institutional power of the governor, but it also underscores the incentives for a party to entirely sidestep compromise when it has the ability to do so. Together, these results suggest that our nationalized political environment has affected the ability

of state lawmakers to govern effectively, but not through the institutional arrangements we usually consider to be the problem. The finding is especially troubling, since States govern on so many of the social issues on which the federal government cannot.

It is important to note that these results use the salience-based measurement approach advanced by Binder (2004), yet we still see a null effect for divided government; we usually see a negative and significant effect when adopting the salience based approach for legislative performance. . The difference between the findings of Mayhew (2005) and Binder (2004) have at times been chocked up to a difference in methodologies. These results demonstrate that it is possible to achieve a "Mayhewvian" result while employing the salience-based strategy. Indeed, while the approach I take is not the exact same as the one used by Binder (2004) – the author's is formulated as the ratio of the number of failed measures to salient measures, and mine is the ratio of salient measures to total measures – it still employs the logic of salience in the denominator that critics of Mayhew promulgate.

The way I measure salience – through agenda setting language in SOTS addresses – may be subject to criticism: it is possible While there are numerous ways to measure the similarity between speeches, my theoretical interest lies in understanding how changes in the agenda affect election outcomes. One of the most important ways that the agenda can affect how individuals vote is by affecting what issues they vote on. Many voters make decisions based on a single issue that they think is most important in the given election (McCombs and Shaw 1972). Elections are thus often determined by which issue wins the conflict of conflicts to become the dominant issue in the eyes of voters (Schattschneider 1960). This gives politicians incentives to think about what

issues they should emphasize in order to win elections (Riker et al. 1996; Aldrich and Griffin 2003; Druckman, Jacobs, and Ostermeier 2004; Dragu and Fan 2016) Because of my theoretical interest in election outcomes and agenda-setting, I measure similarity by looking at the topics covered in the SOTS speeches. Topic modeling is perfectly suited for our purposes because it allows for a consistent measure of how much each speech focuses on each topic.

These results, it should be noted, are correlational. Future research should focus on the application of causal identification strategies, like the one employed in Kirkland and Phillips (2018), to recover the true effect of divided government and nationalization on legislative performance. A major limitation of the study I report in this chapter is that measuring legislative productivity with salient laws may not be the best way of measuring legislative productivity. First, there may be better ways to measure it. Indeed, Kirkland and Phillips (2018) argue that a better way to measure legislative productivity is with delays to proposed budgets. With budgets, governors must "put their money where their mouth is," while risking blame for any government shutdown—a more informative, higher stakes signal. Second, the agenda topics I use to code for salient laws may be endogenous to the presence of divided government. As Kousser and Phillips (2012) point out, the context may entice a governor to change their agenda strategy, putting fewer (or more) items on the agenda. This could affect both the numerator and denominator of the ratio I use for legislative productivity. This is a difficulty which may show up in a small but endogenous effect of divided government on legislative productivity. Lags on productivity might reduce this risk by controlling for the relative increase or decrease in productivity in any given time

period. Though not a perfect approach, I have included lags in my regression in an attempt to control for this risk.

In the 1960's, responsible party government theorists held that stronger parties would be able to more effectively govern because they would suppress local interests while providing voters with clearer, more informative accountability. The results reported in this chapter may serve to underscore the aphorism, "be careful what you wish for." Our parties have become so strong that it is hard to tell the difference between a local race and a national one, in terms of both the policies we set and the outcomes we witness; yet nationalization can decrease our ability to govern.

Conclusion

The substantive findings of this dissertation are focused on the present phenomenon of American political nationalization. In recent decades, national factors (such as the president's popularity) have become increasingly predictive of political issues and electoral engagement at the state, or local, level. The second paper, "Have State Policy Agendas Become More Nationalized?" extended the study of how national factors affect local ones by examining if the national policy agenda has become increasingly similar to local agendas. The analysis showed that State agendas have become more similar to each other over time, and that State agendas have become more similar to the national agenda. The analysis then demonstrated that the similarity between the state and national agendas predicts the nationalization of gubernatorial elections.

The results of this second paper build on the research of Hopkins (2018) to provide a more robust picture of how political nationalization writ large is related to the nationalization of political issues. In fact, it is more evident than ever that the nationalization of our politics is moderated by the nationalization of the political agenda, in addition to other factors including polarization and group cues. If we assume a Downsian lens, the findings suggest that voters are engaging with some form of rational, information-based issue voting, in contrast to hypotheses of purely affective

partisan teaming. Although the unit of analysis is not perfect for the inference, it would be perhaps conclusively terrible if we were to have observed that there *is not a relationship* between the nationalized agenda and the nationalization of elections. Thus, it is encouraging that we do not see a null relationship. These conclusions are tempered, of course, by concerns of endogeneity.

The third paper, "Can States Govern Effectively When Politics Are Nationalized?" considers the question of whether nationalization moderates the relationship between divided government and legislative productivity in the states. Conventional wisdom holds that divided control of the government hinders the ability of our elected representatives to govern, but research into whether this is the case has been a mixed bag. I introduced new evidence by testing whether divided government affected lawmaking in the States from 1960–2012, and found a null effect of divided government on lawmaking ability. This suggests that divided government is generally unrelated to legislative performance, but interestingly, it does so using a salience-based estimator that is consistent with studies that usually show a non-null effect. There are, of course, significant limitations to the estimator, including concerns of endogeneity.

The results also suggest that while nationalization is *not* related to the ability of our state governments to take action on salient issues during times of divided government, nationalization of state legislatures has generally *decreased* the production of salient laws. This finding is a somewhat troubling. It suggests that our nationalized political environment has affected the ability of state lawmakers to govern effectively, but not through the institutional arrangements we usually consider to be the problem.

Taken together, these papers support a discussion around whether the nation-

alization of elections could represent a rational response to the choices that voters face. They imply that nationalization is not evidence that voters have an irrationally motivated focus on party. Instead, it looks more-so that voters are considering political issues and then consequences of their voting decisions when making choices at the ballot box. Because state policy agendas have nationalized and party is such a strong signal of a politicians' positions, voters now face similar choices at both the state and national level when they are deciding how to vote. Given this situation, it is not surprising that voters are voting for candidates of the same party at both the state and national levels. My findings add depth to this story, suggesting that voters may not be voting for their parties simply because of their partisan attachments. On the contrary, my results suggest that voters still care about the issues on the policy agenda. However, my findings cast a shadow over the prospect of responsive governance. If nationalization has decreased the ability of our States to govern, then we may have allowed out most important social policy making apparatus to atrophy in exchange for stronger parties. Though these results are correlational, they suggest that nationalization is a more nuanced beast than it may seem.

This dissertation also introduced in "A Theory and Method for Pooling Naturally Distinct Corpora" a methodological framework for when and why to pool naturally distinct corpora in the course of automated content analysis. Perhaps the most stunning conclusion I drew in the course of this methods paper is that the issue of text comparison is not really a "text" issue at all. Instead, it is a subset of a longstanding problem that still has yet to be concisely solved—the comparison of disparate observations. The innovation of relating text to the problem is that the

sparseness of text enables the derivation of a proposed solution to the problem, in cases where sparse, high-dimensional data are available.

If I could summarize the methods paper of this dissertation up in a sentence, it would be "what you're actually doing when you're doing a text analysis." It is not so much that I actually believe we are able to recover latent traits—that is not falsifiable. It's more about how after decades of research, we can have a dialogue about what text analysis means, by taking the assumptions we're making to their natural conclusions. My hope is that scholars in political science, economics, and communications will be able to make better inferences because of it.

Data science and machine learning, in the form of supervised classification and regression algorithms, are not necessarily useful to the pursuit of scientific inference, because the algorithms and approaches prioritize goals that are altogether different from the goals of the scientist. As Hogg 2019 points out, while "the ML community has delivered great ideas and methods for building, fitting, and validating extremely flexible models, [...] if we want to exploit the good things about ML but achieve truly scientific goals, we need to do two things: we need to augment or modify the (currently trivial) causal structure of the ML methods to represent our very strong domain-specific beliefs about how the data are generated, and we need to be careful to use ML methods only in the parts of our problems for which we don't care about the latent structure or parameters (that is, use them to model nuisances, not use them to do everything)." Hogg uses examples from stellar astrophysics to demonstrate his point.

The application of machine learning to the analysis of text, with the intent of

scientific inference, is no different. The models we produce must reflect a theoretically motivated data generating process, and the content analyses we conduct must take full account of the design considerations at hand. Generally, this is part of a broader trend in empirical political science research. Scholars like Clifford Carrubba, Gary King, and Adam Glynn have said it best: we need to figure out theory and new analytics instead of sitting on the laurels of big data; there has been too much focus on big data collection. And, as I demonstrate in the opening chapters of the dissertation, ignoring design and analytics has put us at risk of drawing unsupported conclusions.

Speaking of analytic techniques: if there were an obvious next step on where to take this dissertation, it would be extending the methods to identify specific agenda proposals in political text, and then determining their individual framings, goal direction, and persuasiveness. Researchers then often make the assumption that the presence of the word in both context implies the same *directional meaning*, or weight. In other words, that the monotonic and increasing use of the word "death tax" implies increasing levels of conservative ideology, in the context of any document or corpus in which it is found. This assumption is problematic, as Chapter 2 discusses. This dissertation does not in detail consider the monotonicity of the weights assigned to particular words in the measurement instrument; rather, it simply looks to see if patterns in usage of words observed are similar in both corpora. The application of the method to the idea of weights (linear, non-linear, monotonic, and non-monotonic) is a future research direction.

In this dissertation, I have intentionally steered clear of any suggestion that the proposed methodologies are able to identify with a computer the ideological framing

and contextual goal direction of an agenda proposals. I have instead limited it to the examination of salience. The ability to code computationally for these other factors has been, for me, much harder than it first seemed. Luckily, however, the theory and method introduced in this dissertation should help lay the groundwork. It seems obvious to me now that the first step in teaching a computer to suss out framing, goal direction, context, and persuasiveness is helping it to understand how text relates to certain subsets and classes of political ideas, and if we should even expect any political ideas *of that particular subset or class* to be present in it, given what we know about the sources.

Appendix A

Code Samples

This appendix contains minimal examples that may be used to produce the delta-statistics reported in Chapter 3. The code has been reformatted to be printed as a "code memo" in the appendix of this dissertation.

Listing A.1: Minimal Example for Estimation of the Delta Statistic

```
#########################
# ENVIRONMENT
#########################

# Import our packages
import pandas as pd
import numpy as np
from scipy import spatial

# Our sklearn imports
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.decomposition import NMF, LatentDirichletAllocation

# Text Wrangling
from nltk.corpus import stopwords
from nltk.stem.snowball import PorterStemmer, SnowballStemmer
from nltk.stem import WordNetLemmatizer
stops = stopwords.words('english')
from unidecode import unidecode
import string
import regex as re

#########################
# FUNC METHODS
#########################

# Utility functions for similarity.
def compute_sims(mat1, mat2, K=K):
    sims = []
    for i in range(K):
        isims = []
```

```
        for j in range(K):
            sim = get_sim(mat1.components_[i,:], mat2.components_[j,:])
            #sim = 1 - spatial.distance.cosine(mat1.components_[i,:], mat2.components_[j,:])
            isims.append(sim)
        sims.append(isims)
    return sims

def get_sim(ivec, jvec):
    # b
#     bc = np.sqrt(ivec.T * jvec).sum()
#     dist = -np.log(bc)
#     return dist
    # cosine
    sim = 1 - spatial.distance.cosine(ivec, jvec)
    return sim

# Utility to print topic summaries.
def display_topics(model, feature_names, no_top_words):
    topic_collector = {}
    for topic_idx, topic in enumerate(model.components_):
        topwords = [ feature_names[i] for i in topic.argsort()[:-no_top_words - 1:-1] ]
        topweights = model.components_[topic_idx,:].argsort()[:-no_top_words - 1:-1]
        topic_collector[topic_idx] = {
            "text": " ".join(topwords),
            "words": { k:v for k, v in zip(topwords, topweights) }
        }
    return topic_collector

# Utility function for bhatta dist
def bhatta(x, y):
    score = np.sqrt(x * y).sum(1)
    return -np.log(score)

# Topic linking algorithm.
def compute_matches(simmat, K=K):

    matches = []
    i = list(range(K))
    j = list(range(K))

    for ip in range(K):
        amax = np.argmax(simmat)
        row = amax // (K - ip)
        col = amax % (K - ip)
    #     print(row, col)
        matches.append({
            "i": i.pop(row),
            "j": j.pop(col),
            "val": simmat[row, col]
        })
```

```
        usei = list(range(K-ip))
        usej = list(range(K-ip))
        usei.pop(row)
        usej.pop(col)
    #     print(usei)
    #     print(usej)
    #     print(simmat.shape)
        simmat = simmat[usei, :]
        simmat = simmat[:, usej]
    return matches

# Siimulation of random dirichlet.
def rdirichlet(a, K):
    y = np.random.gamma(a, 1, K)
    return y / y.sum()

# Null distance distribution simulation.
def null_dist(a, simK, nsims):
    null_tops = []
    for i in range(nsims):
        temp = rdirichlet(a, simK)
        null_tops.append(temp)
    return np.asmatrix(null_tops)

# MLE estimator utility for gamma.
def solve_gamma(vec):
    mu = vec.mean()
    sigma = vec.std()
    ra = (mu + np.sqrt( mu**2 + 4*sigma**2 )) / (2*sigma**2)
    sh = 1 + mu + ra
    return ra, sh

# Empirical critical value simulator.
def empirical_critical_value(topic_idx, cumdens=0.5, nsims=1000,
    model=cr_int_nmf, *args, **kwargs):
    qtop = model.components_[topic_idx,:]
    alpha_hat, beta_hat = solve_gamma(qtop)
    null_tops = null_dist(beta_hat, simK=qtop.shape[0], nsims=nsims)
    null_sims = np.apply_along_axis(get_sim, 1, null_tops, qtop)
    null_sims.sort()
    if kwargs.get("return_sims"):
        return null_sims, null_sims[int(nsims * cumdens)]
    return null_sims[int(nsims * cumdens)]

# Empirical null value simulator.
def empirical_null_distance(topic_idx, cumdens=0.5, nsims=1000,
    model=cr_int_nmf, *args, **kwargs):
    qtop = model.components_[topic_idx,:]
    alpha_hat, beta_hat = solve_gamma(qtop)
    null_tops = null_dist(beta_hat, simK=qtop.shape[0], nsims=nsims)
    dists = []
    for i in range(null_tops.shape[0]):
        a = np.squeeze(np.asarray(null_tops[i,:]))
        sr = np.sqrt(a * qtop).sum()
```

```
        dists.append(-np.log(sr))
    null_sims = np.asarray(dists)
    null_sims.sort()
    if kwargs.get("return_sims"):
        return null_sims, null_sims[int(nsims * cumdens)]
    return null_sims[int(nsims * cumdens)]

################################
# EXECUTE DELTA STATISTIC
################################

# Load corpus.
news_meta = pd.read_csv("26242-0001-Data.tsv", sep="\t", encoding="latin1")
news = pd.read_csv("26242-0002-Data.tsv", sep="\t", encoding="latin1")
cr_meta = pd.read_csv("26242-0003-Data.tsv", sep="\t", encoding="latin1")
cr = pd.read_csv("26242-0004-Data.tsv", sep="\t", encoding="latin1")
gs_out = pd.read_csv("26242-0005-Data.tsv", sep="\t", encoding="latin1")
ents = sorted(news_meta.newspaper_id.tolist() + cr_meta.congress_id.tolist())
ents_id = { k: i for i, k in enumerate(ents)}
int_dictionary = sorted(list(set(gs_out.phrase_stub.tolist())))
int_id = { k: i for i, k in enumerate(int_dictionary)}

# Construct sparse matrices.
i = []
j = []
v = []

for idx, row in news.iterrows():
    if row.count == 0:
        continue
    i.append(ents_id.get(row.newspaper_id))
    j.append(int_id.get(row.phrase_stub))
    v.append(row["count"])

for idx, row in cr.iterrows():
    if row.count == 0:
        continue
    i.append(ents_id.get(row.congress_id))
    j.append(int_id.get(row.phrase_stub))
    v.append(row["count"])

dtm = coo_matrix((np.asarray(v, dtype=int), (np.asarray(i), np.asarray(j)))).tocsc()
news_dtm_int = dtm[:434,:]
cr_dtm_int = dtm[434:,:]

# Estimate topic models.
K = 30
```

```
news_int_model = LatentDirichletAllocation(n_components=K,
    random_state=120938675).fit(news_dtm_int)
cr_int_model = LatentDirichletAllocation(n_components=K,random_state
    =99863455).fit(cr_dtm_int)

# Alternatively, fit for tl-NMF.
# news_int_tfidf = TfidfTransformer()
# news_int_tfidf = news_int_tfidf.fit_transform(news_dtm_int)
# cr_int_tfidf = TfidfTransformer()
# cr_int_tfidf = cr_int_tfidf.fit_transform(cr_dtm_int)
# news_int_model = NMF(n_components=K).fit(news_int_tfidf)
# cr_int_model = NMF(n_components=K).fit(cr_int_tfidf)

# Matches between two NMF models, CR-NEWS.
sims = compute_sims(cr_int_nmf, news_int_nmf)
simmat = np.matrix(sims)
matches = compute_matches(simmat)
matches = pd.DataFrame(matches)

# Topic descriptions.
cr_int_tops = display_topics(cr_int_model, int_dictionary, 30)
news_int_tops = display_topics(news_int_, int_dictionary, 30)
matches["cr_toptext"] = matches.i.map(lambda x: cr_int_tops.get(x).
    get("text"))
matches["news_toptext"] = matches.j.map(lambda x: news_int_tops.get(
    x).get("text"))

# Prevalences.
cr_int_doctops = cr_int_model.transform(cr_dtm_int)
news_int_doctops = news_int_model.transform(news_dtm_int)
matches["prev_i"] = cr_int_doctops.sum(0) / cr_int_doctops.sum(0).
    sum()
matches["prev_j"] = news_int_doctops.sum(0)[matches.j.values] /
    news_int_doctops.sum(0).sum()

# Test stats. Using BD, can use others from Chap 2.
matches["bd"] = bhatta(cr_int_model.components_, news_int_model.
    components_)
matches["null_bc"] = matches.i.map(empirical_null_distance)
matches["bc_resid"] = matches.bd - matches.null_bc
matches.val.hist()

# Delta stat.
delta_stat = (matches.bc_resid**2 / matches.null_bc).sum()
deg_free = K-1
chisq_cv = scipy.stats.chi2.ppf(0.95, deg_free)
print(delta_stat, "|", chisq_cv)
```

Listing A.2: Minimal Example for Winsor Lower Bound Optimization

```
#########################
# ENVIRONMENT
#########################

# Import our packages
import pandas as pd
import numpy as np
from scipy import spatial
from scipy.stats import entropy
from gensim import corpora, models, matutils, similarities
import pickle

# Our sklearn imports
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.decomposition import NMF, LatentDirichletAllocation

# Text Wrangling
from nltk.corpus import stopwords
from nltk.stem.snowball import PorterStemmer, SnowballStemmer
from nltk.stem import WordNetLemmatizer
stops = stopwords.words('english')
from unidecode import unidecode
import string
import regex as re

#########################
# FUNC METHODS
#########################

# Count winsoration tool.
def winsorise(dtm, limits=[0.001, 0.95]):
    featsums = dtm.sum(0).A.ravel()
    ordered_feats = featsums.argsort()
    prop_feats = featsums[ordered_feats] / featsums.sum()
    feat_cumsum = prop_feats.cumsum()
    cut_beg = np.argmax(feat_cumsum > limits[0])
    cut_end = np.argmax(feat_cumsum > limits[1])
    newfeats = ordered_feats[cut_beg:cut_end]
    newfeats.sort()
    return dtm[:,newfeats], newfeats

# Rank winsoration tool.
def winsorise_rank(dtm, limits=[0.05, 0.95]):
    featsums = dtm.sum(0).A.ravel()
    ordered_feats = featsums.argsort()
    prop_feats = ordered_feats / ordered_feats.shape[0]
    newfeats = ordered_feats[(prop_feats > limits[0]) & prop_feats <
    limits[1]]
    newfeats.sort()
    return dtm[:,newfeats], newfeats

```

```
# Map winsoration over several parameters.
def winsor_tops(wlb, K):

    for gramlen, dat in dtms.items():

        print("Winsor Lower Bound={}, K={}, Ngram={}".format(wlb, K, gramlen))

        dtm = dat['dtm']
        featnames = dat['featnames']

        if gramlen == 1:
            wlb_use = wlb - 0.2
        else:
            wlb_use = wlb

        wc_dtm, keeps = winsorise(dtm, limits=[wlb_use, 0.95])
        wc_featnames = featnames[keeps]
        dictionary = { i: k for i, k in enumerate(wc_featnames) }

        corpus = matutils.Sparse2Corpus(wc_dtm.T)
        lda = models.LdaModel(corpus, id2word=dictionary, num_topics=K)

        cm = models.CoherenceModel(model=lda, corpus=corpus, coherence="u_mass")
        coherence = cm.get_coherence()
        top_coh = cm.get_coherence_per_topic()

        return {
            "dictionary": dictionary,
            "f_k": (wc_dtm.sum(0) / wc_dtm.sum()).A.ravel(),
            "lda_model": lda,
            "model_coh": coherence,
            "topic_coh": top_coh,
            "winsor_lb": wlb,
            "K": K,
            "ngram": gramlen
        }

# Helpers for topic summarization methods and RA valiation.
def _jensen_shannon(_P, _Q):
    _M = 0.5 * (_P + _Q)
    return 0.5 * (entropy(_P, _M) + entropy(_Q, _M))

def relevance_w(phi_k, p_w, lam=0.33):
    logbeta = np.log(phi_k)
    return lam * logbeta + (1 - lam) * (logbeta - np.log(p_w))

def frex_w(phi_k, p_w, w=0.5):
    logbeta = np.log(phi_w)
    excl = logbeta - np.logaddexp.reduce(logbeta)
    freqscore = logbeta.argsort() / logbeta.shape[1]
    exclscore = excl.argsort() / excl.shape[1]
```

```
    frex = 1 / ( w / freqscore + (1-w) / exclscore )
    return frex

def lift_w(phi_k, p_w):
    return phi_k / p_w

def phi(phi_k, p_w):
    return phi_k

def topic_summary(phi_w, p_w, wc_featnames, nwords=10, *args, **
    kwargs):

    methods = {
        "phi": phi,
        "rel": relevance_w,
        "lift": lift_w,
        "frex": frex_w
    }

    res = []

    use_methods = []
    eligible_methods = list(methods.keys())

    method_compute = kwargs.get("all", True)
    if method_compute:
        use_methods = eligible_methods
    else:
        for i in kwargs:
            if i in eligible_methods:
                use_methods.append(i)

    for method in use_methods:
        f = methods[method]
        rankmat = f(phi_w, p_w)
        for k in range(rankmat.shape[0]):
            topwords = rankmat[k,:].argsort()[:-nwords-1:-1]
            for w in topwords:
                res.append({
                    "topic": k,
                    "word": wc_featnames[w],
                    "rank": w,
                    "weight":rankmat[k,w],
                    "method": method
                })

    return pd.DataFrame(res)

def get_top_topics(vec, n=3):
    vec = sorted(vec, key=lambda x: x[1], reverse=True)
    top = vec[:n]
    out = {}
    for i, (tidx, tphi) in enumerate(top):
        out["topic_{}".format(i)] = tidx
```

```
        out["topic_{}_prop".format(i)] = tphi
    return out


################################
# EXECUTE WLB FITTING
################################

# Load corpus
df = pd.read_csv("../data/sots.csv")

# Paragraph downsampling
downsamples = []

for i, r in df.iterrows():
    sample = []
    cleaned = r.clean2.split(" ")
    for idx, w in enumerate(cleaned):
        sample.append(w)
        if (idx % 200) == 0 and idx > 1:
            newsamp = r.copy()
            newsamp["clean3"] = " ".join(sample)
            sample = []
            downsamples.append(newsamp)
    if len(sample) > 0:
        newsamp = r.copy()
        newsamp["clean3"] = " ".join(sample)
        sample = []
        downsamples.append(newsamp)

df = pd.DataFrame(downsamples)

# Create simple or phrased features.
vec = CountVectorizer(ngram_range=(1, 2))
dtm = vec.fit_transform(df.clean3)
vec_uni = CountVectorizer()
dtm_uni = vec_uni.fit_transform(df.clean3)
binder = np.vectorize(lambda x: "_".join(x.split()))
featnames = binder(np.asarray(vec.get_feature_names()))
featnames_uni = binder(np.asarray(vec_uni.get_feature_names()))
dtms = {
    1: {
        "dtm": dtm_uni,
        "vec": vec_uni,
        'featnames': featnames_uni
    },
    2: {
        "dtm": dtm,
        "vec": vec,
        "featnames": featnames
    }
}

# Create the grid optimization matrix. Method saves output
```

```
# as it is created, because this can take a long time.
collector = []
K_start = 30
K = 120
kstep = 10
WLB_start = 250
WLB = 600
wlstep = 25
for k in range(K_start, K+kstep, kstep):
    for wlb in range(WLB_start, WLB+wlstep, wlstep):
        tout = winsor_tops(wlb / 1000, k)
        collector.append(tout)
        with open("./coherence_curve_fine_gram.pkl", "wb") as f:
            pickle.dump(collector, f)

with open("./coherence_curve_fine_gram.pkl", "rb") as f:
    collector = pickle.load(f)

# Create coherence matrix.
coherence_params = []
for d in collector:
    coherence_params.append({
        "K" : d.get("K"),
        "wlb": d.get("winsor_lb"),
        "coh": d.get("model_coh"),
        "ngram": d.get("ngram")
    })
coherence_params = pd.DataFrame(coherence_params)
coherence_params.sort_values("coh")

# Find argmax that optimizes coherence.
coherence = pd.DataFrame(collector)
coherence = coherence.sort_values("model_coh", ascending=True)
use = coherence.iloc[0]
phi_w = use.lda_model.get_topics()
wc_featnames = use.dictionary
p_w = use.f_k

# Create automated topic summaries.
tops = topic_summary(phi_w, p_w, wc_featnames)
tops.groupby(["topic", "method"]).apply(
    lambda x: ", ".join(x['word'])).to_frame().to_csv("./
   topic_summaries.csv")

# Create validation sets for research assistants.
topsums = tops.groupby(["topic", "method"]).apply(
    lambda x: ", ".join(x['word'])).to_frame().reset_index().pivot("
   topic", "method", 0)
topsums.to_csv("./topic_summaries_pivot.csv")
top_topics_chunks = [get_top_topics(lda_vecs[i]) for i in range(
   df_codes.shape[0])]
top_topics_chunks = pd.DataFrame(top_topics_chunks)
df_codes = df[["clean3"]]
df_codes = pd.concat([df_codes.reset_index(drop=True),
```

```
    top_topics_chunks], 1)
df_codes.sample(500).to_csv("./topic_chunks_toptops_coded.csv")
```

Appendix B

Appendix to Chapter 1

To overcome this problem, we choose to build a model to estimate $\hat{y}_i$ for each Newspaper. But what may we use on the right hand side of the equation? The answer, employed by Gentzkow and Shapiro (2010), is that we may use text—speeches from Members, and articles from Newspapers. We create a model of text and ideology $y_i = f(W_i, \epsilon)$, where W_i is the $1 \times (M-1)$ row vector of words used in the Member's speech or the Newspaper's articles. We regress known ideology on W^L among Members (decorated with an L to indicate that they are labelled with observed data), and then we estimate $\hat{y}_i = f(W_i^U)$ for Newspapers (decorated with a U to indicated that they are unlabelled). This lets us estimate $\bar{Y} = \frac{1}{2}\mu_Y + \frac{1}{2}\mu_{\hat{Y}}$. The question this chapter asks is, what does this estimand $\bar{Y}$ mean? When is it suitable as a population inference? There is also a slippery slope. If we can estimate $\bar{Y}$, then can't we also estimate $\rho(Y, \hat{Y})$? Can't we draw meaning from $\mathbb{E}[y_i - \hat{y}_{i'}]$?

The answer is that these estimands and others only mean something if you can prove that the model of text you posit works for both populations. Most analyses assume that the model of text posited works for both populations but do not test this assumption. The delta-statistic is a method that allows the researcher to test this assumption. I argue you can prove that the model of text posited works for both

populations if you can prove that language patterns in both corpora are significantly more similar than patterns generated by chance.

B.1 The Systematic Categorization of Texts (cont.)

The systematic categorization of text documents is an age-old problem. Enclaves of monks, employed by the Catholic church in the late 1600s, worked around the clock to track anti-church sentiment by tracking the ratio of combatant texts to all relevant texts in several European districts (Krippendorff 2012). Harold Lasswell led a generation of early communications scholars by applying the same techniques used by the monks to wartime communications during the Second World War to predict the probability of an axis attack on allied forces; Bernard Berelson (1952), building on this early social scientific work, proposed the first theoretical approach to his neologistic practice of "content analysis."

Understanding how the latent structures measured by text influence or are influenced is a natural avenue of inquiry for social scientists, who have since the beginning of modern sociology incorporated text data into their research designs (Lippmann 1922; Lasswell 1938). The fundamental problem of text analysis is that of representing the latent phenomena which generated the text at hand, through the process of *content analysis*.

Content analysis has since spread to an impressive number of disciplines, and

changed the way we interact with the world. The ability to automate such analyses with machines has played a significant role in the digitization of the world we see today. One need only consider the search website Google to understand how transformative content analysis has been: every day, 700 million people visit Google to search the web; Google then yields results by comparing the content analyses of their queries and against a database of trillions of content analyses it has performed on web pages.

The ubiquity of readily accessible "big data" is perhaps the defining characteristic of the present day, and this phenomenon has changed the nature of social scientific research on text. One day of textual *exhaust data* from user traces on the internet – captured in search keywords, Facebook posts, and "tweets" – dwarfs the estimated amount of data we as a species produced through the year 2003 (Wiener and Bronson 2014; Siegler 2010). Data processing of Catalist's database of 200 million U.S. voters, and linkage of it (Monroe 2013, "vinculation") to online text data feeds, would involve the processing of at least tens of trillions of individual records (before considering the memory implications of storing and processing sparse data structures for the representation of token frequencies). Numerous research projects are digitizing large quantities of material which until now existed exclusively in print (*e.g.* ongoing research from Windett, Harden, and Hall 2015; Clark and Lauderdale 2012).[1]

[1]The value of text data to the study of politics is overwhelming. As this massive store of data increases in size, so do the multitudes of newly observable social, economic, and political interactions that generated them. Indeed, one need only consider a brief survey of the topics reviewed in an introductory political science regimen to see that anything studiable in politics is studiable because it entails text. Legislation is debated and passed in text; elected representatives use press releases to communicate with their constituencies in text; political platforms are drafted and ratified in text; the news is published for voter consumption in text; trade deals, contracts, and court decisions are considered and agreed to in text; international treaties are negotiated in text; bureaucrats collect comments and register new federal regulations in text. Though there exist many more corpora than those which have been delineated, the sheer number of topics studiable through text is astounding.

Any political discussion, whether it be in Congress or between neighbors, may be captured and analyzed in the form of text. There are, however, major difficulties entailed in the analysis of political text. Conducting content analysis by hand-coding text is slow and requires domain-specific expertise, which makes manual text analysis expensive. In many cases, these costs are insurmountable for researchers, who must do research on a timeline and budget. Furthermore, manual content analysis carries the bias of the researcher, who brings with her preconceived notions of what she is looking for, the rates at which she should observe it, and the ennui of repeated coding. This bias produces ample fodder for reviewers, who may critique the research on the basis of irreplicability. While this explosion of data provides great opportunity for new applications and research, it has made population inference from hand-coded content analyses infeasible, due to cost and error owing to the attrition of the analysts (King and Lowe 2003, page 618).

B.2 The Problems of Content Analysis and Joint Scaling are the Same

As discussed *supra*, *content analysis* is the method used to estimate latent traits in text data. Similarly, *scaling* is the method used estimate latent traits in opinion data, or sparse voting data. The data structures used in the practice of content analysis, and in the practice of scaling, are identical. It is therefore evident by way of analogy that any problems one might encounter in the joint scaling of ideal points would also

be encountered in the pooled comparison of texts. It is clear by analogy, too, that any solutions one might employ in the comparison of texts to one another might also be employed in the joint scaling of ideal points. See figure I1 for detail.

Going even further, why might we expect text data to be comparable to responses in a survey, or roll call votes? If one accepts the idea that text is the realization of a latent attitude, communicated to the world and measured by language, then the two are actually very similar. When asked on a survey, "do you support gun control, and how much," the respondent may answer "yes, and very much (a five on the Likert scale)." When asked to respond to the question, "do you support gun control, and how much," the respondent may answer yes, and the use language to indicate the intensity of her preference. The two methods both elicit data from the respondent, with which the researcher may measure her latent attitudes.[2] Indeed, this is the same process of opinion measurement which has been proposed by Zaller (1991).

We may go further than analogy, however. Item Response Theory is the theoretical basis which allows for the measurement of latent preferences in text data, survey data, and roll call data. using text. The framework may easily be applied. Whereas on exams individuals are usually asked a number of questions, to which they produce answers, free-text responses are presented as unstructured data. Whereas on surveys

[2]The measurement instrument for the attitude, however, differs, as does its proximity to the source. The survey item directly asks a respondent to generated a realization of her opinion. Except for the presence of self-censorship, dishonesty, perceptions, or other sorts of error, the recorded data are quite close to the respondent's latent opinion. Text data, on the other hand, must be secondarily processed using statistical methods to approximate what a survey response would have looked like—and as such, is subject to increased noise. Worthy of note, however, is the fact that a realization generated by way of survey may not include the same degree of error as a realization generated by way of language. For instance, through language, a respondent may express a measurable opinion that is more true to her latent opinion (for instance, due to a lack of a demand effect).

individuals are usually asked a number of questions, to which they provide responses, roll call votes are presented as sparse binary data.

B.3 Common Issues in the Analysis of Text

The complexity of language presents numerous challenges for the analysis of text. How should we define the unit of analysis, which we usually refer to as the *document*? For example, consider a set of agenda speeches from an incumbent politician. Should we consider each speech to be a document? Or should we consider the paragraphs of each speech to be a document? This process, in which we define what constitutes a document, can cause significant variability in results we see. Topic models, for instance, are better able to recover political agenda items at the paragraph level, since political agenda speeches naturally discuss the same agenda item within each paragraph (for further detail, see section 1.5).[3] In this dissertation, I define a document to be a cohesive collection of political statements, communicated by a unitary political actor at a discrete point in time. Let d be a raw, unprocessed document.

In addition to deciding what the unit of analysis is, we must also specify the population, which we usually refer to as the *corpus*. This process of *corpus specification* has significant implications for the extensibility of the analysis. What is the ultimate population we wish to make inferences about, and how are the documents we observe related to that population? How were the documents we observe selected, or sampled?

[3]If the document unit of analysis is more appropriate population-level inference than the paragraph, how should we aggregate model predictions to allow for such research? There are trade-offs between the theoretically-driven unit of analysis and the one that works best for modeling. The mathematics of such models produce more reliably produce discrete topics.

Design considerations such as selection bias can easily defeat a text analysis, without regard for the hundreds of millions of documents entailed in the analysis.[4] Strong *a priori* theory for how documents measure the latent traits of the processes, individuals, or groups that generated them serves as armor against the inevitable false positives the researcher will encounter.[5] In this dissertation, I generally define a corpus to be a set of documents generated by a population of unitary political actors, such that any political actor might generate one or more documents. Let $\mathcal{D}$ be a corpus of documents, and i any individual unitary political actor, such that $d_i \in \mathcal{D}\ \forall d, i$.

The form and magnitude of text data make it difficult to analyze. An analyst must interpret any given document in its *unstructured* form through the process of reading. Depending on the viewpoint of that analyst, or the time period in which the analyst reads the document, the conclusions drawn by the analyst for that document can vary significantly.[6] For example, simply consider any one of George Wallace's speeches delivered prior to the Mississippi Freedom Summer, or when Reverend Dr. Joseph Lowery spoke directly to him about his Methodist transgressions in awakening racial animus. Readers today are likely to have a very different impression of what they mean than readers several decades ago.

The magnitude of text data can also overwhelm the analyst, and promote biases

[4]Corpus specification is especially important for a particular type of text analysis, which uses models trained on labeled data to produce labels for unlabeled data. The communication of the same agenda topic by two different politicians may differ significantly, and the set of words used to do so by the two politicians may overlap very little. Should we include documents from both politicians as part of the same population?

[5]Text data entail many features, and variable distributional profiles within those features, that classical hypothesis testing becomes inappropriate (for more detail, see Hastie, Tibshirani, and Friedman 2009).

[6]For further discussion, see Krippendorff (2012).

such as conceptual drift and satisficing (Simon 1997). There are three reasons why the sheer magnitude of text data can make text analysis hard. First, any document can, in its *structured* form, reach unmanageable size. While we cannot outside of human-produced scores conduct empirical analysis on a document in unstructured form, we can do so on a document in structured form. In its most basic structured form, a political agenda speech could be considered as a spreadsheet with a row for each paragraph, a column for each type of word, and a cell value for the number of times we observe each word in each paragraph. The speech's dimension, which we may derive from the number of rows and columns in the spreadsheet, can quickly reach considerable size.[7] How we usually produce the structured form is discussed in the next section.

Second, the number of documents can be overwhelming. Consider a corpus of one billion microblogged political agenda speeches (a small fraction of the data presently available online). It is untenable for the researcher to read every speech and generate an analysis of it, because the exercise would involve significant time and cost. It is perhaps even futile, considering the effect mental exhaustion and ennui would have on the ability of the researcher to produce an unbiased and replicable result. At a certain point, the researcher and any number of her staff employed to do the same will begin looking for the answer they are satisfied with—not the answer which is universally true.

[7]The simulation of a null document – the document we would expect to see by chance presents even greater challenge. The dimension of the null distribution is even more challenging to represent. In a world of pure entropy, where there is no rhyme or reason to the order or domain of words which may be strung together, the number of potential combinations that might emerge from a 30-word paragraph (such as the one you are presently reading) is equal to 10^{30}—more than the number of atoms in the universe!

Third, and most importantly, such considerable size at both the document and corpus level is useless when it comes to the practice of science, which requires us to describe, test, and explain simple truths about the world. To say anything of meaning with text analysis, we must reduce the dimensionality of the text we observe, and map it to a simpler, usually unidimensional, metric. For example, when we use the text of a politician's speech to estimate her ideology, we usually attempt to distill the hundreds of thousands – if not millions – of words the politician uses down to a single number.[8]

B.4 Issues in Measurement with Content

The concept of content in representation of a latent trait is problematic. A scholar generally might hold one of three conceptions of the relationship between a latent phenomenon – which is captured by particular content of interest – and observed text. In the first conception, content is an inherent property of the text; one need look no further than the text itself to distill the message within (Berelson 1952; Gerbner 1985; Shapiro and Markoff 1997). This conception is irrelevant for the purposes of source inferences. Content analysis for the purpose of inference must assume that content is a property of the underlying phenomena that generated the text. If content is a property of observed text, then there can be no theory that links it to the fundamental process that generated it; doing so would immediately allow for the presence of a source whose traits are predictable using text. Therefore, inference must assume that

[8]In other cases, a set of fewer than three-hundred numbers might be the goal.

content is a property of the underlying phenomena that generated the text. Most text analyses published today take this view (Laver et al. 2003; Grimmer 2010; Quinn et al. 2010).

In the second conception, content is a property of the source that generated the text; an underlying process possesses qualities that its communications can be used to measure (Krippendorff and Stone 1969; Osgood 1959; Holsti 1969). It is *prima facie* the case that text may reliably predict the characteristics of a source. For instance, I have demonstrated in a short study for Penn State that even short survey free responses that concern a general evaluation of the performance of the President may predict characteristics of the respondents, such as age, employment status, education, and political attitudes.[9]

In the third conception, the properties of content and the source that generated it are recoverable only by understanding the relationship of the reader to the text and the source (Krippendorff 2012). In other words, there are several perspectives for which one must account to make inferences about the relationship between a source, a text, and a reader (or receiver). This conception implies a ensemble or boosted method for the prediction of latent traits, *given the traits of the reader*. Though this

[9]In the example I have just provided, I explicitly call out the use of observable characteristics as variables, such as age, employment status, and other demographics. The common theoretical relationship between demographics and what we observed is perhaps taken as a given, but in this case the theory is that the language respondents use to evaluate the president is likely related to these traits, because the evaluation of the president is often also related to these things (Bond and Fleisher 2001; Hehman, Gaertner, and Dovidio 2011). Secondarily, one problem with the latent traits theory is that we use text to measure traits, but we cannot observe what the outcome would have been given the source had a different level value for one or more traits. This is the fundamental problem of causal inference. I plan eventually to experimentally manipulate the perceived characteristics of respondents either by way of vignette or framing to see if this manipulation affects the language the source uses. If there is an effect, then we may identify the link between latent traits and observed language.

dissertation will not address this third conception directly, this is an avenue of research which will generate insight into the two-sided nature of political communication, and I very much wish to pursue this agenda in the future.

In conceptions one and two, the boundary of content is that which is common to all readers. In conception three, content is conditional on context. I contend that content analysis for the purpose of inference *must assume that content is a property of the underlying phenomena that generated the text.*

Proof (sketch). Content that is a property of the text lends itself, for instance, to the empirical study of the effect of text on behavior. A study of framing might experimentally assign textual vignettes and compare attitudinal outcomes on the basis of variations in text. The study would generate a mapping from the observed text to the average value on the attitudinal scale. Importantly, however, the researcher supervising this study would be limited in what she may claim about the nature of the experimental treatment. If content is a property of observed text, then there is no theory that links it to the fundamental attitudinal process. Therefore, inference must assume that content is a property of the underlying phenomena that generated the text.

B.4.1 Empirically Linking Text to Latent Traits

The critical assumption involved in using a topic model to estimate the underlying preference and choice spaces is that the language we observe is actually related to the latent traits of the individual, or data generating process, which produced the text.

There is ample precedent for this assumption (as has been discussed at length in earlier sections). Taking the second conception of content, described in appendix B.4, to its natural end allows for the linkage of latent traits and text through an appropriate mathemetical model: the *topic model*, which clusters words that occur together into topics, and then represents speech as probability distributions over those topics. Because language is so context-specific, variables that explain the context of the source – such as gender, ideology, party, geography, etc. – will be statistically related to the topical predictions generated by the model. This is akin to the argument of stylometry research, applied within a topic modeling context (Mosteller and Wallace 1964).[10]

It is evident in descriptive research that the text we observe in survey free responses is highly related to latent opinion. Consider, for example, a result from the "Mood of the Nation" poll, run by Pennsylvania State University's McCourtney Institute for Democracy. In the poll, subjects are asked to respond to several free responses (Berkman and Plutzer 2018). These free responses concern the condition of the nation at present. Two of the free responses entail a manipulation. In the first free response, the subject is asked to disclose what they are proud of. Then, they are primed to put themselves in the shoes of the other party: "You said earlier you identified as a [PARTY]. You have neighbors who are [OUT-PARTY]. Why do you think they voted for the President?" In the second free response, they respond to the question of pride

[10]Language is highly context specific, to the point that patterns of word usage may be used to forensically recover the identity of an author (Mosteller and Wallace 1964) or trace the transmission of an idea from one actor to another (Duranti and Goodwin 1992). Text contains lingual *signatures* suggestive of the presence of an underlying data generating process. The contention of this dissertation is the signature of one dimension will appear in any document the dimension influenced. Though the prevalence of the signature may vary conditional on its salience relative to other topics, the co-occurrence of the terms will not vary.

again, but as if they were in the out-party.

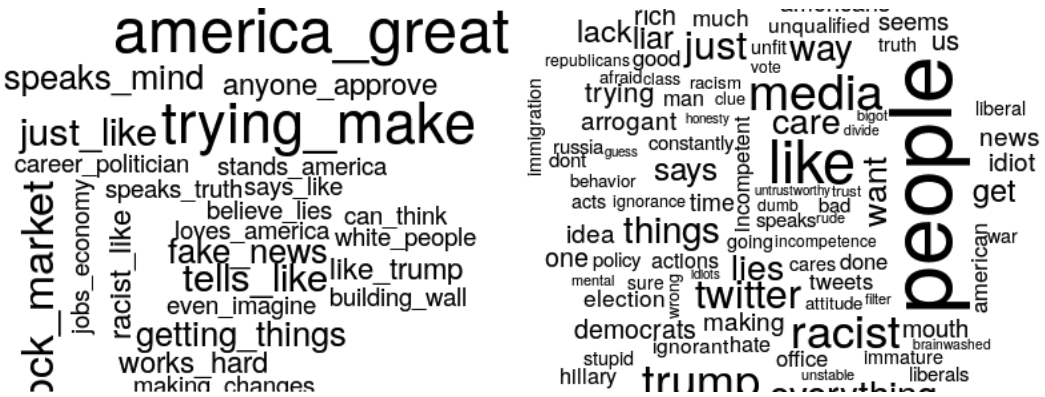
We can use the χ^2 statistic to determine which, if any terms, differ significantly between the first and second free responses. The χ^2 test is applied to test the independence of two events. In selecting the language that is most diagnostic of an event or trait, the two events are occurrence of the term and occurrence of the class. We then rank terms with respect to the following quantity:

$$\begin{aligned}\chi^2(\mathbb{D}, t, c) &= \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \\ &= \frac{(N_{11} + N_{10} + N_{01} + N_{00})(N_{11}N_{10} - N_{01}N_{00})^2}{(N_{11} + N_{01})(N_{11} + N_{10})(N_{10} + N_{00})(N_{01} + N_{00})},\end{aligned}$$

where N is the number of times a term is observed, and E is the number of times a term would be expected, t is the term in question, and c is the class in which we are interested.

Figure B.1 reports the terms that uniquely predict the first and second free responses, weighted by their tf-idf values. The purpose of this figure is to demonstrate that there is indeed a clear relationship between the language used to evaluate the president and the selected conditioning latent trait I have chosen (in this case, party). For instance, you can see that co-partisans use positive language like "making America great" and "doing away with fake news" to evaluate, while out-partisans believe the president is "dumb" and "racist." This tracks, and is not at all surprising. It is clear that conditional on the respondent's party, the language used to respond to the question changes. This quick example suggests that even in short-form free response

Figure B.1: The Language Expressed in Free Responses in 2018 Was Highly Relevant to the Respondent's Support for the President



(a) Republicans (Support President) **(b)** Democrats (Oppose President)

Note: Panels are wordclouds generated by free response data from the "Mood of the Nation" Poll, run by the McCourtney Institute at Penn State ($N = 1,000$), during the spring (pre-midterm) of 2018. Respondents were asked why they believe the opposing block of people voted (or did not vote) for President Trump. The size of each word is proportional to its tf-idf weight, a measure which is larger for words that better discriminate between the classes. Panel (a) includes key features for respondents who support the President. Panel (b) includes key features for respondents who oppose the President.

text, the observed language is related to latent traits.[11] What remains to be shown is that topic models can recover clusters of words which are also related to latent traits.[12]

B.5 A Typology of Text Analyses

To map the theory I develop to an empirical approach, I propose an analytical typology of text analyses, in order to target for the reader the types of analyses for which

[11]One drawback to this approach with respect to any inference causally linking traits to text, however, is that it does not control for demand characteristics (Orne 1962); the setup cues to participants what is expected of them, and it is obvious that the respondents have used this sequence as an opportunity to praise their own party and then bash the out-party. Instead of a manipulation on their self-perceived traits, the design gives respondents another opportunity to express the pre-existing trait.

[12]For further empirical study of the relationship between language and latent traits, see Yeoman (2017), in which the researchers manipulate "politeness" and observe changes in language as such.

satisfaction of the bridge criteria are necessary. The typology divides text analyses into two types. Type I analyses test for relationships between a corpus of text and a variable of interest. An example of Type I analysis is the research published in Mosteller and Wallace (1964), which tests for differences in certain "style words" conditional on the authorship of several Federalist papers. Type II analyses predict labels (codes) for unlabeled sets of documents, often using a model based on a pre-labeled set. An example of Type II analysis is the paper Gentzkow and Shapiro (2010), which uses a model of text and ideology trained on the Congressional Record to code for ideology in a corpus of newspaper articles. It is the second type which invokes the bridge criteria; it is the second type which necessitates testing for the criteria.

To introduce the notation used herein, consider two sets of textual documents. Let $\mathcal{D}^L$ include N^L labeled documents, and let $\mathcal{D}^U$ include N^U unlabeled documents. The generic subscript i indexes individual documents within each set, and the total number of documents is $N = N^L + N^U$. Each document may be assigned a value y_{ik}, which indicates a document's membership in each category label k, for an exhaustive but not mutually exclusive set of label categories, $k \in \{1, \ldots, K\}$. Labels y_{ik} are not observed in the unlabeled document set.

I now describe the basics of a routine text analysis, for the benefit of the reader who is not familiar with text analysis. In every text analysis, the first step after corpus specification[13] is to map the unstructured corpus space $\mathcal{D}$ into a space of constructed

[13]Corpus specification (see appendix B.3) is an exercise in which the researcher defines what the unit of analysis for the study will be, from what population the units used for analysis were sampled, and what the sampling procedure was.

Figure B.2: Excerpt from the President Trump's State of the Union Address in 2019

Sentence 1. Victory is not winning for our party.

Sentence 2. Victory is winning for our country.

Sentence 3. Gentlemen, we salute you.

Table B.1: Exemplar Document–Term Matrix for President Trump's SOTU Address

	victory (w)	winning	party	country	gentlemen	salute
Sentence 1 (i)	1 (c_{iw})	1	1	0	0	0
Sentence 2	1	1	0	1	0	0
Sentence 3	0	0	0	0	1	1

features called a document–term matrix, $\mathbf{C}$.[14] $\mathbf{C}$ may be thought of as a spreadsheet, with a row for each document i, a column for each feature type w (word, or token), and a cell value for the number of times we observe each word in each document c_{iw}. Also recall the earlier notation, in which any document i may also be represented as a row vector of words W_i, $c_{iw} \in W_i \; \forall w$.

For example, figure B.2 presents three sentences from Donald Trump's State of the Union address in 2019, and appendix B.5 presents the resulting matrix $\mathbf{C}$ generated from it. Each sentence is considered as if it were a document, and as such, each row of the matrix corresponds to each sentence. The reader will note that certain features of the text, such as punctuation, capitalization, and certain words have been removed from the example. This is a result of the mapping function, $m(\cdot)$.

There are several ways of mapping $\mathcal{D}$ to $\mathbf{C}$, a critical step of which is commonly referred to as "pre-processing" the unstructured text. For example, Laver et al. (2003) discard numbers, punctuation, and stopwords to yield a set of unstemmed, lowercased

[14] $\mathbf{C}$ is sometimes referred to as the three-letter acronym for document–term matrix, the "DTM."

words for each document, which are then counted to produce cell values.[15] We may also engineer features from these documents that capture syntactic relationships between those words, or represent detected entities. The reader should note that there is no "one-size-fits-all" solution to pre-processing, and the steps taken to prepare text for the construction of $\mathbf{C}$ should be carefully considered in light of the research question.[16] Generally, let the function $m : \mathcal{D} \rightarrow \mathbf{C}$ exist to map documents to the space of constructed features, and assume that the same mapping function is used for all documents, $m_i = m_{i'} \ \forall i \in \mathcal{D}$.[17] $\mathbf{C}$ is usually stored in a sparse storage format (Witten et al. 1999).

B.5.1 Typology of Text Analyses and the Bridge Criteria

In the process of content analysis, the researcher makes numerous assumptions about the process that generated the text. Principally, one assumes that the bridge criteria are satisfied (section 1.3.1). The bridge criteria, however, are not necessary to invoke in every text analysis. When the analysis entails the comparison of two or more political texts, the criteria are necessary. However, in simple descriptive examinations of a single corpus, the bridge criteria are unnecessary. It is only once any text, or set

[15]Grimmer and Stewart (2013) review in their paper best practices for pre-processing, should the reader be interested.

[16]For example, Yeoman (2017) demonstrate that stopwords can be indicative of politeness; they should not be removed if the purpose of the study is to examine politeness. See also Handler et al. (2016), which demonstrates the utility of features constructed from phrases, called "phrasemes", can improve the performance of estimators that estimate ideology based on speech in Congress.

[17]The choice of m is an important because it drives the distance metrics used in downstream analyses. Using an aggressive stemmer in m that pools several tokens together into a single feature will reduce the size of the support space, thereby yielding a smaller denominator. Using a weighting technique in m will also affect downstream results. For example, using tf-idf will reduce the sampling variability of similarity queries between document subsets, but it will also reduce the average similarity between them.

of texts, involved in the analysis may be considered sufficiently distinct from other texts pooled together with them that the bridge criteria become critical.[18]

The mapping of text to another variable of interest may be used in practice for two purposes. First, it may be used to test for a statistical relationship between the text and the variable of interest. Second, it may be used to predict the variable of interest for another corpus of text. Let these two analyses be analyses of the "first" (I) and "second" (II) types.

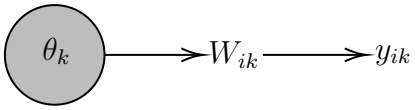
B.5.2 Characterization of Type I Analysis

Type I analyses, in which the researcher tests for a relationship between a corpus of text and a variable of interest, proceed as follows. The researcher develops a strong theory for why the text should be related to the metric of interest. Then, the researcher demonstrates a statistical relationship between the two ("testing" for a relationship). The researcher may then draw the conclusion that the two are related. For example, it makes intuitive sense that a politician's speech would be linked to their ideology, since ideology is the sum total of her policy preferences, and political speeches usually communicate policy preferences; we would expect to observe a statistical relationship between the two, and in many cases we do (Laver et al. 2003; Gentzkow, Kelly, and Taddy 2017, *e.g.*).

Testing for a statistical relationship between a corpus of text and a variable of

[18]The criteria are, of course, important to consider even in single-corpus analyses. The principal question in many cases is, *how sure are we that the set of texts being analyzed actually compose a single corpus?* Looking for natural groupings of text, perhaps based on another variable or annotation, is a natural place to start.

Figure B.3: Type I Analyses: Posited Data Generating Process



Note: Figure represents graphically the data generating process posited in Type I analyses. θ_k is the latent dimension of interest, about which the researcher wishes to say something. W_{ik} is the vector of words which the researcher believes are observed because of the latent dimension. y_{ik} is the observed covariate – the measurement – which the researcher believes may be used to represent the latent dimension θ_k.

interest is a common exercise. In many cases, there may be an interest in what language is more likely to be used within certain groups, or conditional on the levels of the variable of interest (Laver et al. 2003; Monroe, Colaresi, and Quinn 2008). In other cases, the quantity of interest may be the proportion of documents which fall within a predetermined set of categories (Hopkins and King 2010; Jerzak, King, and Strezhnev 2018). The researcher may even be interested in discovering a new way to organize texts (Grimmer and King 2011, *e.g.*), or to estimate the ideal points of actors in a spatial model (Laver et al. 2003; Monroe and Maeda 2004; Slapin and Proksch 2008). The posited data generating process for such cases in presented in figure B.3.

Simply put, the goal of a Type I analysis is to test for a relationship between documents and their labels. One approach is to test for differences in the outcome, given the content of the documents. Consider a binary variable to flag when the phrase "death tax" appears in a document $w_{\text{death tax}} > 0$. The approach is to test:

$$\mathbb{E}[y_{iw} \mid w_i = 1] = \mathbb{E}[y_{iw} \mid w_i = 0]. \tag{B.1}$$

Another (perhaps more interesting) approach is to test for differences in the content

of the documents, given the label. For example, the following approach tests if the phrase "death tax" is just as likely to occur in documents with label k as in documents with label k':

$$\mathbb{E}[w_i \mid y_{ik}] = \mathbb{E}[w_i \mid y_{ik'},\ \forall\ k \neq k']. \tag{B.2}$$

These approaches may be generalized to test for an ordinal or continuous relationship between the number of times the phrase appears in a document, and the ordered labels K

$$\mathbb{E}[\rho(y_k, \log(C_w))] = 0, \tag{B.3}$$

or the number of times any wording $w \in \{$"death tax", "second amendment", $\ldots,\ W\}$ appears in the documents:

$$y_{ik} = \log(C_{iw})\beta_w + \epsilon_i, \tag{B.4}$$

$$\mathbb{E}[\beta_w] = 0,\ \forall w_i. \tag{B.5}$$

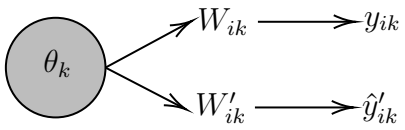
In performing these tests, the researcher usually produces a "machine-interpretable codebook," which is used to automate content analysis on the documents. The codebook is a set of key words or phrases which indicate the latent category the researcher wishes to detect. These words or phrases may contribute to the assignment of a label in complex ways. For example, the researcher may assign each word a

second-order function which weights its marginal contribution to the label outcome, or; the researcher may weight each word relative to the usage rates of other words within a given document, or the usage rates of other words throughout the entire corpus.

A computer program then counts the number of times it observes each codebook word in a set of documents. These counts are fed into a set of logical rules, which produce labels for the documents. Machine learning provides a way to select the words used in the lexicon and write the rules that label the documents. These techniques train a model on text that has already been annotated by content analysts, and then use that model to predict annotations in another dataset. The stripped-down codebook procedure allows for the analysis of documents at great scale. But the secondary effect of the machine-readable codebook is the abandonment of the researcher feedback loop.

Perhaps most important to the Type I analysis is the researcher's ability to make the case for valid measurement. It is imperative that a strong case is made for the following: that the variable of interest does indeed exist in the source of the text, that the source is able to transmit that variable of interest via text, and that the analytical technique we use to relate the text to the variable of interest reveals true markers of the variable of interest. Indeed, it would be silly to conduct an analysis of the *Oxford English Dictionary* and conclude that the author of the text was predisposed to the liberal ideology.

Figure B.4: Type II Analyses: Posited Data Generating Process



Note: Figure represents graphically the data generating process posited in Type II analyses. θ_k is the latent dimension of interest, about which the researcher wishes to say something. W_{ik} is the vector of words which the researcher believes are observed because of the latent dimension, and $W_{ik'}$ is the vector of words which the researcher believes are observed because of the latent dimension in another set of documents. y_{ik} is the observed covariate – the measurement – which the researcher believes may be used to represent the latent dimension θ_k, and $\hat{y}_{ik'}$ is the estimated value which the researcher believes may be used to represent the latent dimension θ_k.

B.5.3 Characterization of Type II Analysis

Type I analyses are limited by the number of labeled documents available to the study. Type II analyses, in which the researcher uses a regression on old text to estimate the variable of interest on new text, proceed as follows. The researcher produces a Type I analysis, which usually yields a model which maps text to the variable of interest. The researcher then applies that model to new text to produce labels for the unlabeled documents on hand. The posited data generating process for such cases in presented in figure B.4.

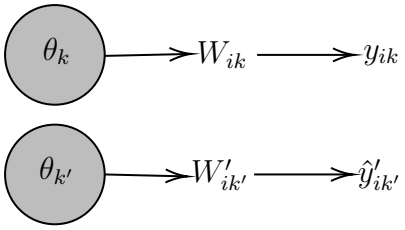
Type II analysis predicts the category label y_k for the unlabeled set of documents $\mathcal{D}^U$, using a regression or classification model trained on $\mathcal{D}^L$:

$$y_{ik} = f(W_i^L), \tag{B.6}$$

$$\hat{y}_{ik} = f(W_i^U). \tag{B.7}$$

The ability to conduct a valid Type I analysis is necessary for any Type II analysis, because without an appropriate Type I analysis, the ability of a model to predict the

Figure B.5: Type II Analyses: Alternative Data Generating Process



Note: Figure represents graphically the (alternative) data generating process in Type II analyses. θ_k is the latent dimension of interest, about which the researcher wishes to say something. W_{ik} is the vector of words which the researcher believes are observed because of the latent dimension. y_{ik} is the observed covariate – the measurement – which the researcher believes may be used to represent the latent dimension θ_k. Meanwhile, it is another dimension $\theta_{k'}$ which produces $\hat{y}_{ik'}$, and it therefore cannot be used to represent the latent dimension θ_k in the other corpus.

metric is not validated. One cannot use an invalid model to make valid predictions for the purpose of scientific inference. This principle is unwavering, regardless of the database to which the model is applied to make predictions.

Using a mapping to predict the variable of interest for a new corpus of text is a practice which has grown exponentially in recent years. In such cases, there may be an interest in using the predicted values to draw a conclusion about the nature of the new source of text. Papers tend to focus on the estimation of ideal points for the new corpus, which are then used to describe the political actors that generated the text (Groseclose and Milyo 2005), or make inferences about the relationship between the ideal points and another phenomenon (Martin and McCrain 2019; Martin and Yurukoglu 2017; Gentzkow and Shapiro 2010). Emerging approaches, focusing on the detection of certain patterns in political speech (e.g., "Could this speech be considered religious? Could this advertisement be considered negative?") consider the variable of interest to be membership in a particular pattern, or cluster, of words, and use sources with existing annotations on those variables to then annotate new

corpora (Morgan n.d.). These emerging approaches are the scholarly grandchildren of automated content analysis approaches based on keyword detection, or dictionary methods (Nacos et al. 1991); the emerging approaches conduct statistical inference to estimate the function $f(\cdot)$ that links the text to the variable of interest, whereas previous approaches assume $f(\cdot)$ as known.

Of course, as has been discussed in section 1.1, a major issue with scientific research using these methods is that the predicted labels may not be validated against actual "ground truth" labels, because in many cases, the ground truth is unobservable (or not easily measured), and in most cases, the ground truth is not readily accessible in the database as hand. An example might add refreshing detail. Consider, for example, a model trained on *Twitter* data and donation activity to estimate political ideology from text. What might happen if we apply the model to a corpus of spelling tests from kindergartners? Scoring the ideology of kindergartners on the basis of their spelling tests might produce ideology scores that appear distributionally valid, but the results might be driven by misspellings. Kindergartners likely do not possess or express political ideology as *Twitter* users do.[19] The alternative data generating process for such cases in presented in figure B.5.

B.6 Supporting Tables

[19]Though the reader may consider consulting Achen (2002) for one theory on the acquisition of party with a fully-fledged retrospective ability, at a young age.

Table B.2: Matched Topics and Match Statistics (Congressional Record and News)

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
10	16	91.1	domesticviolence violencewomen violencewomenact victimdomestic victimdomesticviolen violencesexualassaul lawenforcementagenci domesticviolencesexu justicedepartment committehomelandsecu asianpacific lookforward iraqiwomen businessowner pacificislander nucleareoption checkbal asianpacificislander houserepublican asianamerican ownparty taxcutpeople percentafricanameric titleix asianpacificcamerican martinluther lutherking millionchildren personalinformation medicliability	domesticviolence taxincrease childsupport creditcard nationalguard thirdtime budgetcut lowincome victimdomesticviolen victimdomestic businessmeeting savemoney methlab republicanparty gulfcoast drinkingwater federgencymanage educationprogram financecommitte privateproperty votingmachine blockgrant disabledamericanvete hurricanekatrina pledgeallegiance liabilityinsurance godbless tradeagreement fretrade additionalfunding	1.516273343	0.070325106

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
26	5	84.4	africanamerican tuskegeairmen blackcaucu congressionalblackca americancommunity africanamerican-commu africanameri-canwomen hazardou-material wariraq percentafricanameric votingmachine postof-fice democratrepub-lican ryanwhitecare whitecareact mi-norityownedbusines businessowner wash-ingtondc lutherking martinluther com-mittegovernmentre manufacturingjob iwojima votecounted blockgrant civilright jobcreation budget-cut nationalguard immigrationlaw	africanamerican civil-right hurricanekatrina lowincome lutherking bottomline martin-luther americancom-munity collegestudent courtappeal civilright-movement rightmove-ment middleclass africanamerican-commu budgetcut courtjudge republi-canparty gulfcoast lookforward ros-apark thirdtime littlegirl transitsys-tem asianamerican africanamericanwomen communitydevelop-ment billiondollar taxrevenue poorpeople hatecrime	0.452825813	0.093260796

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
8	2	82.9	hurricanekatrina gulf-coast naturaldisaster louisianamississippi nationalguard passbil medicmalpractice floodinsuranceprogra nationalfloodinsuran committehomeland-secu chemicplant pas-sengerrail billiondollar malpracticeinsuranc2 terrischiavo federe-mergencymanage presidentbudget taxre-lief americabloodcent legalsystem lowincome securityplan mediclia-bility millionamerican guardreserve block-grant medicmalpracti-ceinsu katrinavictim pricegouging bilcut	hurricanekatrina vfed-eremergencymanage littlerock warriraq driverlicense nation-alguard justicedepart-ment billiondollar katrinavictim gulfcoast courtappeal colleges-tudent childrenfamili americanpeople nat-uraldisaster iraqwar insurancecompani prescriptiondrug oil-compani warterrorism democratrepublican courtjudge district-judge nursinghome pro-gramhelp finalminute boyscoutamerica republicansenator circuitcourt paytax	0.044147374	0.203310469

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
9	19	81	naturalga naturalre-source climatechange foreignoil energynat-uralresourc washing-tondc oilnaturalga nu-clearpower highwaybil gasolineprice lowin-come economicgrowth financecommitte so-larenergy spendmoney dependenceforeignoil cleanenergy wildlifer-efuge nationalforest tradedeficit createjob oilcompani pricenat-uralga fueleconomy gaoil presidentan-nounce forestservice cleanairact rhodeisland nationalwildlife	naturalga wildlifer-efuge nationalwildlife nationalwildliferefu oilfield hurricaneka-trina oilcompani creditcard oilnaturalga arcticnationalwildli gasolineprice legisla-tivesession growthrate oilindustry cashflow taxrate bottomline economicgrowth financecommitte bil-liondollar courtappeal oilproduction natural-resource exxonmobil oildrilling heatingoil justicedepartment circuitcourtappeal lowincome rateincrease	0.455496829	0.348087417

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
25	0	69.6	affordablehousing lowincome communitydevelopment blockgrant terrorismriskinsuran incomefamili civilright communitydevelopmen2 developmentblockgran voterregistration billiondollar floodinsuranceprogra incomepeople millionamerican lowincomefamili nationalfloodinsuran hurricanekatrina insuranceindustry republicanparty ryanwhitecare whitecareact lowincomepeople familyvalue endangeredspeciact housingmarket lookforward railsystem programhelp drinkingwater palestinianauthority	lowincome affordablehousing hurricanekatrina educationprogram creditcard wariraq budgetcut childrenfamili businessowner bottomline illegalimmigrant incomefamili communitydevelopment iraqwar collegestudent housingmarket drinkingwater lowincomefamili domesticviolence civilright budgetdeficit programhelp billiondollar prescriptiondrug borderpatrol seniorcitizen stemcel taxrevenue illegalimmigration lookforward	0.755355615	0.381327545

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
14	11	40.8	iraqipeople saddamhussein iraqafghanistan nationalguard wariraq oilfood warterrorism forceiraq foodprogram oilfoodprogram coalitionforce iraqiwomen troophome troopiraq unreform bodyarmor boyscout godbless billiondollar iraqwar bringtroop bringtroophome militaryfamili militaryoperation freelection operationiraqifreedo suicidebomber passbil sovietunion globalwarterrorism	saddamhussein wariraq iraqwar justicedepartment americanpeople billiondollar europeanunion civilright nuclearweapon creditcard administrationoffici suicidebomber courtappeal chiefjustice worldtrade ronaldreagan middleclass gulfcoast johnrobert prescriptiondrug republicanparty warterror millionamerican bottomline karlrove troopiraq iraqafghanistan minorityleader economicgrowth terrischiavo	0.746657949	0.512742527

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
19	23	40.4	lookforward transitsystem warterrorism lawenforcementagenci nationalguard naturalresource methlab foreignoil godbless washingtondc thirdtime americancommunity foreignrelation americanvalue globalwar americanconsumer nationalsecurityissu tradecommission nationaldefense globalwarterrorism federtradecommission federtrade humanlife heritagemonth immigrationreform committehomelandsecu safetynet dependenceforeignoil millionamerican reducedependencefore	creditcard lookforward bottomline businessowner hurricanekatrina boyscout littlegirl postoffice thirdtime nationalguard memorialday worldtrade godbless spendmoney savemoney educationprogram millionamerican civilright postalservice katrinavictim economicgrowth naturaldisaster singleday boygirlclub familybusiness middleclass closeddoor professionalsport centdiseasecontrol personalinformation	0.266024952	0.11910747

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
24	3	35.6	chiefjustice circuitcourt judicialnomine courtappeal judgerobert nuclearoption circuitcourtappeal chiefjusticerehnquis courtjudge democraticleader johnrobert checkbal foreignrelation guardreserve changerule timevote pledgeallegiance justicesupremecourt nationalguard climatechange financecommitte majorityvote nationalsecuritylett rulechange unitedsupremecourt justicedepartment parentnotification washingtondc roewade creditcard	circuitcourt circuitcourtjudge hurricanekatrina naturalresource communitydevelopment taxrevenue gulfcoast taxincrease lowincome bottomline lookforward budgetcut businessowner financecommitte courtappeal taxrate seniorcitizen littlerock circuitjudge judgerobert courtjudge lawenforcementagenci educationprogram martinluther economicgrowth stpaul budgetcommitte gasolineprice naturalgaheadstart	0.560370094	0.693625416

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
21	26	33.6	personalaccount securitysystem socialsecuritysystem taxrelief taxincrease securitybenefit socialsecuritybenefi americanpeople retirementaccount economicgrowth ratereturn currentsystem raisetaxe personalretirementac budgetcommitte socialsecurityprogra socialsecuritytrust securitytrust securitytrustfund financecommitte socialsecurityreform securityreform reformsocialsecurity savingaccount millionjob seniorcitizen raisingtaxe federbudget lowincome benefitcut	taxrate taxincrease republicanparty taxrelief taxrevenue collegestudent savemoney seniorcitizen lowincome hurricanekatrina bottomline rateincrease financecommitte budgetproposal taxbreak wariraq nursinghome raisetaxe sexoffender memorialday closeddoor civilright minorityleader partyline middleclass spendmoney democratrepublican iraqwar thirdtime increasetaxe	0.137192294	0.57047713

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
3	7	23.7	rosapark civilright rightmovement civil-rightmovement lit-tlerock lutherking martinluther railser-vice passengerrail africanamerican high-waybil railsystem servingcountry na-tionalheritagecorr votingright drmartin-luther armeniangeno-cide washingtondc littlerocknine pass-bil votingrightact governmentresponse passengerrailservice centdiseasecontrol foodstamp creditcard civilrightprotection collegeeducation dis-easecontrolpreven lostjob	martinluther lutherk-ing africanamerican civilright education-program creditcard businessowner dr-martinluther low-income postoffice seniorcitizen saletax rosapark rightmove-ment civilrightmove-ment socialsecuritycard taxincrease children-famili collegeeducation boygirlclub economic-growth postalservice minimumwage votin-gright washingtondc minoritywomenowned boyscoutamerica unitedairforce colleges-tudent communityde-velopment	0.786747687	0.0916668731

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
17	25	23.1	privateproperty property endangered speciact sexoffender privatepropertyright privatepropertyowner terrischiavo taxrevenue postoffice sexualpredator medicliability medicmalpractice wildliferefuge nationalguard guardreserve nationalguardreserve urgesupport driverlicense millionamerican abusepower columbiarivergorge triallawyer protectprivateproper oilnaturalgas fretrade patientsafety federbudget illegalimmigrant timevote sexoffenderregistrat	forestservice nationalforest privateproperty lowincome naturalresource endangeredspeciact propertyright businessowner mountsthelen courtjudge indiantribe legislativesession drinkingwater communitydevelopment taxrevenue illegalimmigrant boygirlclub sexoffender postoffice borderpatrol budgetcommitte minimumwage indianaffair timbersale illegalimmigration courtappeal timberindustry litlegirl incomefamili headstart	0.195022151	0.045459475

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
13	29	21.1	deathtax businessowner familybusiness estatetax urgesupport repealdeathtax economicgrowth taxrelief taxrate taxrepeal naturaldisaster farmbureau deathtaxrepeal createjob pensionplan humanlife millionjob dependenceforeignoil foreignoil committeehomelandsecu timemove illegalimmigration jobcreated dutyfree retirementaccount saletax federspending billiondollar jobcreation highwaybil	businessowner collegestudent driverlicense savemoney programhelp hurricanekatrina prescriptiondrug courtappeal budgetcut naturalgas insurancecompani familybusiness lowincome bottomline rateincrease lostjob victimdomestic minimumwage gulfcoast incomepeople consumerprotection victimdomesticviolen economicgrowth collegeeducation appropriationbil losejob naturalresource holdhearing circuitcourtappeal paytax	1.620348731	0.173689395

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
16	21	14.4	veteranhealth veteranhealthcare vahealthcare iraqafghanistan presidentbudget healthcareveteran guardreserve lowincome healthcareservice creditcard mentheal thcare vahospital nursinghome nationalguard appropriationbil iwojima budgetcut billiondollar educationprogram fundingveteranhealth healthcarebudget millioncut budgetproposal wariraq militaryfamili adequatehealthcare additionalfunding programhelp housepublican disabledamericanvete	nationalguard legislativesession lowincome budgetcommitte childleft prescriptiondrug sexoffender boyscout iraqafghanistan littlerock wildliferefuge savemoney spendmoney taxbreak postalservice saletax federemergencymanage civilright wariraq nationalwildlife chiefjustice educationprogram centdiseasecontrol businessmeeting gulfcoast nationalwildliferefu postoffice collegeeducation houserepublican republicanleader	0.219995879	0.755674376

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
20	9	13.2	prescriptiondrug patientsafety medicaremedicaid lowincome medicareprescription medicliability seniorcitizen nursinghome millionamerican rhodeisland sexoffender safetynet judgerobert savemoney costprescription-drug businessowner iraqafghanistan tradecommission pharmaceuticcompani commonlyprescribeddr federtrade federtradecommission drugbil laborlaw oilcompani baseclosing budgetoffice centralamerican postoffice cutmedicaid	seniorcitizen boyscout businessmeeting lookforward collegestudent educationprogram postoffice memorialday headstart hurricanekatrina godbless prescriptiondrug programhelp childsupport naturalga millionamerican childrenfamili creditcard pledgeallegiance operationiraqifreedo boyscoutamerica exxonmobil nationalguard foodstamp blessamerica sadamhussein godblessamerica americanpeople medicareprescription littlegirl	0.618507997	0.309812105

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
11	20	11.6	postalservice postoffice unitedpostalservice committegovernmentre driverlicense passbil urgesupport washingtondc rateincrease iwojima foodstamp spendmoney gulfcoast oilfood millionamerican americanworker drinkingwater temporaryworker pledgegaliance minorityleader millionjob budgetoffice tencommandment passengerrail holdhearing democratrepublican incomefamili congressionalbudgeto civilright federelection	farmbureau lookforward boyscout collegestudent postoffice nationalguard businessmeeting memorialday naturalga educationprogram naturalresource farmbureaufederation headstart childrenfamili driverlicense pledgegaliance privateproperty courtjudge lowincome prescriptiondrug childleft programhelp budgetcut lawenforcementagenci iraqwar familybusiness katrinavictim budgetdeficit godbless postalservice	0.738562619	0.094740508

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
4	24	9.9	tradeagreement fre- trade centralamer- ican tradedeficit americanworker americanfretrade centralamericanfre worldtrade tradepolicy manufacturingjob dutyfre tradepolici worldtradeorgani- zati jobloss laborlaw strongopposition lostjob democratre- publican workerright propertyright joblost tradedeal economic- growth createjob privateaccount jobo- versea losejob pen- sionplan minimumtax troophome	circuitjudge circuit- court nursinghome creditcard seniorcitizen sexoffender prescrip- tiondrug thirdtime fretrade childrenfamili taxbreak insurancecom- pani tradeagreement boyscout courtappeal savemoney bankruptcy- court legislativesession rateincrease closed- door terrischiavo methlab postoffice houserepublican saletax medicmalprac- tice centralamerican classaction childleft sexualpredator	0.1574901	0.27168367

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
23	17	9.8	minimumwage millionamerican lowincome civilright pensionplan foreignoil creditcard increaseminimumwage nationalforest americanworker taxbreak nationalsecuritylett childleft dependenceforeignoil drinkingwater foodstamp billiondollar wildliferefuge unitedairline wageworker educationprogram childsupport wariraq peopledisabiliti poorpeople incomefamili programhelp raiseminimumwage financecommitte creditcardcompani	naturalresource financecommitte lookforward hurricanekatrina nationalguard millionamerican memorialday educationprogram wariraq nationalwildlife drinkingwater nuclearpower businessowner savemoney budgetcut lawenforcementagenci gulfcoast nationalwildliferefu postoffice wildliferefuge consumerprotection operationiraqifreedo republicanparty lowincome taxrevenue publictelevision circuitcourtjudge rulechange peopledisabiliti heatingcost	1.052058693	0.664490207

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
15	22	8.8	illegalalien illegalimmigration immigrationreform illegalimmigrant immigrationlaw guestworker driverlicense passbil godbless minoritywomenowned guestworkerprogram greencard lawenforcementagenci nationalsecuritylett temporaryworker enforceimmigrationla millionamerican socialsecuritycard millionillegalalien oilfood triallawyer circuitcourt temporaryworkerprogr americanworker oilfoodprogram justicedepartment unitednationreform amnestyillegalalien housingmarket nationalhomeownershi	lawenforcementagenci sexoffender lookforward lowincome seniorcitizen rateincrease communitydevelopment nursinghome domesticviolence taxrate memorialday littlegirl thirdtime bottomline republicanparty prescriptiondrug spendmoney godbless illegalimmigrant wariraq wildliferefuge borderpatrol billiondollar judgerobert programhelp nationalwildliferefu incomefamili nationalforest collegestudent nationalwildlife	0.90332509	0.157356105

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
28	28	8.1	nationaldebt studentloan foodstamp budgetdeficit childsupport billiondollar privateaccount securitytrust socialsecuritytrust securitytrustfund billiontax spendingcut loanprogram billiontaxcut bottomline cutmedicaid presidentbudget taxbreak tradedeficit billiondeficit budgetcut cuttaxe budgetoffice cutspending wariraq congressionalbudgeto costbillion asianpacific estatetax iraqafghanistan	crimelaw bottomline mountsthelen hurricanekatrina forestservice courtjudge wildliferefuge boygirlclub nursinghome domesticviolence nationalwildlife nationalwildliferefu creditcard lowincome courtappeal collegestudent nationalguard wariraq prescriptiondrug driverlicense insurancecompani educationprogram iraqwar republicanparty legislativesession postoffice circuitcourtappeal peopledisabiliti boyscout savemoney	0.863493521	0.840374952

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
22	18	7.9	middleclass taxrelief minimumtax alternativeminimumta middleclassfamili taxincrease millionamerican passbil estatetax createjob taxreliefact economicgrowth medicalpractice taxbreak americanworker highwaytrustfund billiontax billiontaxbreak socialsecuritybenefi securitybenefit middleclassamerican foreignoil childsupport reliefmiddleclass gasolineprice protectsocialsecurit workerright growingeconomy saletax raisetaxe	saletax taxrevenue taxincrease lookforward businessowner bottomline spendmoney thirdtime boyscout littlerock littlegirl savemoney privateproperty childleft collegestudent sexoffender budgetcut johnrobert lowincome courtjudge rateincrease programhelp naturalga gulfcoast legislativesession republicanparty propertyright gasolineprice drinkingwater collegeeducation	0.785896687	0.163045782

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
6	10	7.6	classaction actionlawsuit madisoncounty classactionfairness nationalsecuritylett actionfairnessact childsupport courtappeal legalsystem classactionreform judicialnomine createjob justicedepartment highwaybil californiasupremecou minorityleader economicgrowth creditcard triallawyer lookforward americanbarassociati cuttaxe passbil strongsupport safetybeltlaw civilright pharmaceuticindustry bottomline bankruptcyreform courtjudge	madisoncounty hurricanekatrina collegestudent saletax methlab thirdtime lookforward educationprogram taxrate africanamerican gulfcoast taxrevenue nursinghome nationalguard taxincrease creditcard lawenforcementagenci boygirlclub civilright rateincrease programhelp medicmalpractice lowincome memorialday savemoney postalservice circuitcourt bottomline baseclosing domesticviolence	1.106371148	0.1540213 11

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
7	27	7.1	headstart educationprogram lowincome civilright tuskegeairmen childleft forestservice americanworker childrenfamili jobtrainingprogram militaryfamili insurancecompani headstartteacher studentaid programhelp blockgrant stpaul businessowner collegestudent lowincomechildren medicmalpractice strongopposition educationhealth timemove militaryoperation financialmanagement littlegirl millionamerican professionalsport budgetproposal	communitydevelopment lookforward blockgrant businessowner development blockgran communitydevelopmen2 postoffice courtjudge financecommitte educationprogram boyscout boygirlclub memorialday datetime nationalguard civilright affordablehousing naturalga privateproperty hazardoumaterial wariraq taxrevenue budgetcut rateincrease programhelp godbless childleft additionalfunding americanpeople parenright	1.498656093	0.07827468

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
1	14	5.2	appropriationbil endingseptember forestservice de- fenseappropriation boyscout president- move urgesupport nationalforest billion- dollar budgetcommitte foodstamp rhodeisland bilcut wildliferefuge nationaldefense mi- noritytyleader presi- dentbudget timem- ove iraqafghanistan railservice washing- tondc baseclosing speakertable demo- craticleader datetime highwaybil climate- change savemoney millionbudgetrequest lowincome	rhodeisland fi- nancecommitte affordablehousing courtjudge budgetpro- posal childrenfamili heatingoil baseclos- ing minorityleader hurricanekatrina bot- tomline businessowner seniorcitizen naturalga closeddoor postoffice taxrate privateprop- erty childleft budgetcut prescriptiondrug thirdtime washing- tondc jointcommitte consumerprotection educationprogram judgerobert postalser- vice lookforward celresearch	0.45113568	0.497829473

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
18	4	3.9	nuclearweapon billiondollar taxbreak creditcard millionbudgetrequest unreformeuropeanunion bottomline defenseauthorization secretaryairforce pensionplan humanlifefretrade sovietunioneconomicgrowth armeniangenocide presidentbudget bunkerbuster tradedeficitnuclearpower saletaxoilcompani estatetaxwariraq nuclearoptionissuefacingamericantaxbreakcompani foreignrelation troopiraqinternationalatomice	districtjudge taxincrease taxreliefcreditcard legislativesession bottomlinethirdtime seniorcitizenlowincome nationalguard naturalresourceprofessionalsport nationalfa closeddooraisetaxe civilrightgasolineprice gulfcoast hurricanekatrinataxrevenue childleftjudgerobert hearingscheduled lookforward littlegirl estatetaxblockgrant minimumwage taxbreakcircuitcourtappeal	1.361980998	0.14930699

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
2	8	2.9	stemcel celresearch embryonicstem embryonicstemcel cordblood adultstem adultstemcel humanlife celline humanembryo stemcelline terrischiavo bloodstemcel umbilicord cordbloodstem millionamerican medicmalpractice humancloning driverlicense malpracticeinsuranc2 illegalimmigrant growthrate percentgrowth saddamhussein oilfood bottomline nursinghome armenianegenocide passengerrail hurricanekatrina	courtjudge seniorcitizen affordablehousing courtappeal sexoffender prescriptiondrug closeddoor domesticviolence nationalguard thirdtime wildliferefuge childsupport drinkingwater postoffice worldtrade hearingscheduled lookforward memorialday minoritytleader budgetcut nationalwildlife lawenforcementagenci personalinformation nationalwildlifer-efu iraqafghanistan childleft bidcontract celresearch savemoney taxrevenue	1.58024334	0.020217469

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
29	13	2.4	publicbroadcasting lowincome corporationpublicbro foreignoil publictelevision lowincomefamili sovietunion programhelp educationprogram incomefamili millionamerican drinkingwater billiondollar hurricanekatrina billcut millionchildren dependenceforeignoil oilfood naturaldisaster helpveteran billiontax taxbreak billiontaxbreak safetynet fuelefficiency oilfoodprogram houserepublican foodprogram cuttaxe wildliferefuge	gulfcoast federemergencymanage hurricanekatrina affordablehousing businessowner taxrate lookforward naturaldisaster katrinavictim privateproperty communitydevelopment drinkingwater legislativesession drmartinluther saletax insurancecompani civilright validdriverlicense nursinghome millionamerican taxrevenue southeasttexa louisianamississippi naturalresource nationalwildlife housingmarket saddamhussein programhelp wildliferefuge johnrobert	1.446541122	0.051598923

... continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
12	12	1.8	bordersecurity borderpatrol illegalimmigration driverlicense committehomelandsecu illegalimmigrant passbil guestworker immigrationreform washingtondc securityplan justicedepartment immigrationlaw terrischiavo chemicplant democratrepublican guestworkerprogram nationalwildliferefu nationalwildlife worldtrade presidentbudget armeniangenocide civilright secretaryairforce personalinformation wildliferefuge tuskegeairmen bottomline billiondollar strongsupport	easttexa hurricanekatrina creditcard taxrate southeasttexa districtjudge littlerock saletax legislativesession forestservice oilfield lutherking martinluther lookforward federemergencymanage boygirlclub nursinghome collegestudent naturalga taxrevenue africanamerican sexoffender gulfcoast headstart lowincome seniorcitizen boyscout borderpatrol rateincrease godbless	1.356316452	0.039424784

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
27	1	1.4	warterror globalwar globalwarterror warterrorism iraqafghanistan enemycombatant tuskegeairmen washingtondc endangeredspecact globalwarterrorism foreignrelation sovietunion guardreserve borderpatrol ronaldreagan terrischiavo winwarterror nationalsecurityissu fightingwarterror fretrade economicgrowth federbudget womenarmedforce nationaldefense majorityvote spendmoney violencewomen lawenforcementintell createjob winningwar	nursinghome collegestudent nationalguard boyscout financecommitte disabledamericanvete seniorcitizen stpaul postoffice memorialday postalservice servingcountry vahospital johnrobert republicanparty washingtondc educationprogram programhelp headstart creditcard methlab unitedairforce warterrorism summacumlaude insurancecompani taxrate nationalrifleassocia sexoffender familybusiness littlegirl	0.609708482	0.060597728

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
5	15	0.9	strongsupport urgesupport passbil strongoppositionmillionamericanmethlabcurrentsystemcommittegovernmentrerailsystemloanprogrambusinessownerasianpacifichighwaytrustfundiwojimamedicmalpracticeamericancommunitysavemoneyconsumerprotectionsexoffenderbilliondollarasianpacificamericanparentnotificationhugochaveziraqiwomencutfundingpassengerrailprogramhelpendangeredspeciactbilcutsovietunion	jeffersoncounty southeasttexa nursinghomehurricanekatrina nationalguard civilrightunitedairlinecourtappeallowincomeminorityleaderdistrictjudgenaturalgarepublicanparty legislativesessiontaxrevenuecourtjudge chiefjusticenuclearweapon saletaxcreditcard bottomlinebudgetcommitteprofessionalsport prescriptiondrug collegestudentdomesticviolencebudgetcutclassaction senaterepublicanillegalimmigrant	1.497900692	0.094178278

...continued

CR Index	News Index	Similarity Est.	CR Text	News Text	Observed Distance	Null Distance
0	6	0.1	americanpeople millionamerican wariraq houserepublican democratrepublican oilcompani republicanparty billiondollar spendmoney checkbal iraqwar americanpeopledeserv creditcard drugcompani karlrove senaterepublican taxbreak insurancecompani americanvalue republicanleader ciaagent abusepower partyline americanworker budgetdeficit middleincome abughraib naturaldisaster democraticleader privateaccount	stpaul boyscout naturalresource nursinghome saletax communitydevelopment thirdtime martinluther lowincome professionalsport seniorcitizen businessowner postoffice operationiraqifreedo hurricanekatrina boygirlclub familybusiness sexoffender postalservice childsupport creditcard memorialday childrenfamili budgetdeficit businessmeeting katrinavictim savemoney boyscoutamerica birthcontrol disabledamericanvete	0.756260912	0.260060941

Table B.3: Matched Topics and Match Statistics (Congressional Record)

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
1	1	96.5	stemcel embryonicstem embryonicstemcel celresearch cordblood adultstem adultstemcel celline stemcelline cordbloodstem bloodstemcel driverlicense umbilicord humanlife percentgrowth prescriptiondrug millionamerican saddamhussein unbornchild humanembryo pricegouging terrischiavo humancloning oilfood producestemcel nuclearpower umbilicordblood hatecrime illegalimmigrant embryostemcel	stemcel embryonicstem embryonicstemcel celresearch cordblood adultstem adultstemcel humanembryo americanpeople celline humanlife terrischiavo umbilicord stemcelline humancloning strongsupport nationaladcampaign medicmalpractice malpracticeinsuranc2 bloodstemcel methlab growthrate tencommandment cordbloodstem armeniangenocide illegalimmigrant fretrade drinkingwater nursinghome oilcompani	3.105479448	0.477979328

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
5	12	93.5	domesticviolence violencewomen violencewomenact violencesexualassaul victimdomestic domesticviolencesexu lawenforcementagenci victimdomesticviolen justicedepartment committehomeland-secu nuclearoption iraqiwomen lookforward pacificislander taxcutpeople ownparty asianpacificislander nationalendowmentart titleix climatechange houserepublican driverlicense minorityright wariraq nursinghome asianpacific checkbal cleanwateract wateract lawchange	domesticviolence victimdomestic violencewomen timdomesticviolen violencewomenact affordablehousing percentafricanameric sniperrifle personalinformation violencesexualassaul boyscout iraqiwomen justicedepartment hatecrime bilcut calibersniperrifle domesticviolencesexu lowincome timemove medicmalpractice committehomelandsecu unitedtraderepresent railsystem angelecitycouncil fretrade purchasingpower businessowner millionchildren transitsystem checkbal	1.830489893	0.112271704

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
9	9	91.7	headstart educationprogram civilright lowincome americanworker stpaul headstartteacher childleft childrenfamili medicmalpractice millionamerican insurancecompani studentaid lowincomechildren timemove strongoppposition blockgrant middleclass collegestudent loanprogram jobtrainingprogram educationhealth millionjob angelecitycouncil millionjobcreated foodprogram gundeal programhelp additionalfunding watchlist	headstart forestservice jobtrainingprogram strongsupport educationprogram passbil asianpacific militaryfamili financialmanagement hurricanekatrina iraqiwomen africanamericanwomen asianpacificamerican affordablehousing partylinevote partyline lowincome healthcarecoverage lostjob moneytaxcut budgetdeficit headstartteacher pacificislander budgetsurpluse rosapark katrinavictim paycollege studentaid asianamerican programhelp	1.764424431	0.335141629

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
10	6	91.5	postalservice postoffice unitedpostalservice committegovernmentre foodstamp rateincrease foodstampprogram oilfood driverlicense militaryfamili drinkingwater spendmoney commonlyprescribeddr foodprogram tencommandment washingtondc oilfoodprogram pledgeallegiance vprescriptiondrug urgesupport loanprogram iwojima abughraib planreform terrischiavo campendleton nationaldebt internationalpeacese actionlawsuit outingciaagent	postalservice postoffice unitedpostalservice committegovernmentre driverlicense committehomelandsecu temporaryworker timemove urgesupport custombor derprotecti passbil memorialday dutyfre federelection rateincrease americanpeople voterregistration strongsupport gulfcoast chiefjudge cordblood congressionalbudgeto budgetoffice americanworker washingtondc incomefamili temporaryworkerprogr bordersecurity provisionalballot votingmachine	0.795701246	0.056636046

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
8	3	90.7	classaction actionlawsuit californiasupremecou creditcard courtappeal madisoncounty judicialnomine ameri-canpeople highwaybil legalsystem classactionfairness triallawyer actionfairnessact pharmaceuticindustry bankruptcycourt createjob nuclearoption circuitcourt africanamerican classactionreform democraticleader privateproperty minorityleader judicialconfirmation circuitcourtappeal fifthcircuitcourt privatepropertyright jobcreation thirdtime cuttaxe	classaction actionlawsuit nationalsecuritylett madisoncounty justicedepartment classactionfairness childsupport actionfairnessact hurricanekatrina classactionreform prescriptiondrug safetybeltlaw legalsystem warterror courtappeal economicgrowth cordblood naturaldisaster judicialnomine crimelaw illegalalien americanbarassociati lowincome oilcompani timemove chiefjudge iraqipeople jointcommitte iwojima californiasupremecou	1.892628747	0.225648097

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
21	22	90	publicbroadcasting corporationpublicbro low-income patientsafety publictelevision americanpeople sovietunion programhelp lowincomefamili strongsupport billiondollar millionchildren incomefamili educationprogram millionamerican safetynet floodinsuranceprogra helpveteran africanamerican fuelefficiency centdiseasecontrol bilcut endingseptember asianpacific hundredbilliondollar drinkingwater paycollege houserepublican asianamerican medicaremedicaid	publicbroadcasting affordablehousing corporationpublicbro publictelevision hurricanekatrina lowincome foreignoil middleclassfamili reliefmiddleclass congressionalblackca blackcaucu oilfood bilcut lowincomefamili educationprogram dependenceforeignoil palestinianauthority embryonicstem embryonicstemcel naturaldisaster oilfoodprogram incomefamili voterregistration drinkingwater foodprogram celresearch cutfunding tradedeficit budgetdeficit americanpeople	1.149312193	0.556254101

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
7	10	89.5	deathtax familybusiness repealdeathtax businessowner urgesupport economicgrowth pensionplan americanpeople committehomelandsecu millionjob naturaldisaster taxrate warterror timemove housepublican postoffice dependenceforeignoil retirementaccount worldtradeorganizati palestinianauthority medicliability billiondollar foreignoil humanlife estatetax deathtaxrepeal taxrepeal hugochavez borderpatrol worldtrade	deathtax familybusiness farmbureau deathtaxrepeal taxrepeal repealdeathtax taxrate dutyfree madisoncounty strongsupport judgerobert businessowner illegalimmigration methlab urgesupport taxrelief commonsensereform centralamerican americanfarmbureau federspending economicgrowth estatetax highwaybil carefunding endangeredspeciact marketaccess classactionreform lookforward actionlawsuit jobcreation	0.939802627	0.063431371

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
2	15	88.5	rosapark civilright civilrightmovement rightmovement africanamerican blackcaucu congressionalblackca lutherking martinluther africanamericancommu americanpeople americancommunity armeniangenocide votingright centdiseasecontrol governmentresponse servingcountry hurricanekatrina civilrightprotection diseasecontrolpreven votingrightact railservice littlerock solarenergy voterregistration freelection drmartinluther middleclass assistantsecretary railpassengerservice	rosapark civilright civilrightmovement rightmovement highwaybil africanamerican votingsystem martinluther lutherking hatecrime cordblood foodstamp nationalheritagecorr budgetcut spendmoney racialethnicminoriti votingright universalhealth programcut drmartinluther lostjob votingrightact americanpeople umbiliccord postoffice poorpeople millionchildren familybusiness naturaldisaster universalhealthcare	1.015697361	0.036617458

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
3	7	87.9	tradeagreement fre-trade centralamerican americanfretrade centralamericanfre tradedeficit tradepolici tradepolicy americanworker manufacturingjob tradedeal jobloss laborlaw lostjob dutyfre strongopposition joblost illegalimmigration democratrepublican fasttrack northamericanfre purchasingpower joboversea losejob worldtradeorganizati worldtrade propertyright driverlicense buyamericanproduct buyamerican	tradeagreement fre-trade tradedeficit centralamerican americanworker tradepolicy worldtrade workerright manufacturingjob worldtradeorganizati dutyfre americanfretrade privateaccount blessamerica centralamericanfre createjob tradepolici laborlaw internationallaborst rulechange jobloss strongopposition lostjob joblost economicgrowth rightworker propertyright abughraib paycollege billiontradedeficit	1.730939611	0.107490123

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
16	11	84.9	tuskegeairmen africanamerican hazardoumaterial blackcaucu iwojima congressionalblackca ryanwhitecare whitecareact lawenforcementagenci minorityownedbusines professionalsport litlegirl safetystandard americanpeople familyvalue methlab wariraq spendmoney borderpatrol unitedairforce americancommunity alternativeminimumtaminimumtax gunsafety recorddeficit troopbringhome africanamericanwomen increasetaxe cutbil postoffice	tuskegeairmen africanamerican civilright gulfcoast committegovernmentre postoffice driverlicense communitydevelopment votingsystem communitydevelopmen2 affordablehousing developmentblockgran rightmovement blockgrant civilrightmovement hurricanekatrina unitedairforce americanpeople justicedepartment globalwarterror washingtondc courtappeal propertyright familybusiness vahealthcare vahospital justicejaniceroger louisianamississippi servingcountry	1.326862181	0.74218454

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
25	19	83	personalaccount socialsecuritysystem securitysystem personalretirementac retirementaccount socialsecuritybenefi securitybenefit socialsecurityreform securityreform currentsystem americanpeople ratereturn savingaccount socialsecuritytrust securitytrust securitytrustfund warterror lookforward benefitcut taxincrease socialsecurityprogra palestinianauthority raisingtaxe growingeconomy cashflow raisetaxe reformsocialsecurity cutbenefit healthsavingaccount spendingsocialsecuri	personalaccount socialsecuritysystem security taxincrease socialsecuritybenefi socialsecurityprogra securitybenefit ratereturn reformsocialsecurity retirementaccount personalretirementac climatechange currentsystem taxrelief boyscout alternativeminimumtax seniorcitizen nearretirementage raisetaxe socialsecurityreform securitytrustfund socialsecuritytrust securitytrust economicgrowth millionjob iraqiwomen increasetaxe	0.380331844	0.368581896

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
6	23	82.4	nuclearweapon creditcard europeanunion bunkerbuster defenseauthorization secretaryairforce nuclearoption fretrade foreignrelation cardcompani creditcardcompani gunviolence economicgrowth warirraq warterror nuclearbunkerbuster iraqiwomen assaultweapon drugcompani nationaldefense sovietunion iraqafghanistan calibersniperrifle sniperrifle nunnlugarprogram guardreserve growingeconomy defenseappropriation saddamhussein governmentspending	nuclearweapon issuefacingamerican millionbudgetrequest pensionplan taxbreakcompani taxbreakcompani lookforward lawenforcementagenci strongsupport unreform saletax armeniangenocide presidentbudget endingseptember tradedeficit billiondollar appropriationbil bottomline menthealthcare bankruptcycourt nuclearpower strongopposition americanconsumer financialaccountings joboversea taxrate oilcompani angelecitycouncil insurancecompani internationalatomice	0.616299853	0.138873589

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
12	8	81.8	naturalga solarenergy gasolineprice nuclearpower consentdecre pricenaturalga cleanenergy oilnaturalga naturalresource gaoil patientsafety fundingcut savingaccount privateaccount methlab naturalgaoil savemoney godbless socialsecurityreform federtrade federtradecommission securityreform supplinaturalga supplynaturalga tradecommission cleanairact financecommitte ganatural ganaturalga nationalguard	naturalga climatechange highwaybil naturalresource energynaturalresourc oilnaturalga washingtondc financecommitte nuclearpower economicgrowth highwaytrustfund spendmoney cleanenergy ratereturn presidentmove presidentannounce lookforward hearingscheduled fueleconomy tradedeficit lowincome oilindustry witnesstestify courtappeal rhodeisland globalwar globalwarteror appropriationbil terrorismriskinsuran pricenaturalga	0.488665464	0.055410145

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
26	16	75.8	chiefjustice judgerobert chiefjusticerehnquis circuitcourt courtappeal tencommandment prescriptiondrug nuclearoption judicialnomine unitedsupremecourt timevote boyscout nationalsecuritylett courtjudge easttexa rulechange justicesupremecourt permanentresidentsta circuitcourtappeal sexualpredator americanpeople greencard supremecourtunited checkbal chiefjusticewilliam sovietunion foreignrelation johnrobert changerule roewade	judgerobert chiefjustice judicialnomine nuclearoption circuitcourt johnrobert courtappeal democraticleader pledgeallegiance checkbal courtjudge lookforward nationalguard circuitcourtappeal parentnotification chiefjusticerehnquis foreignrelation majorityvote guardreserve datetime minorityright appropriationbil changerule justicesupremecourt judgejohnrobert roewade republicansenator financecommitte godbless stpaul	1.023557705	0.078859302

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
0	0	70.8	americanpeople privateaccount prescriptiondrug middleclass oilcompani karlrove ciaagent taxbreak budgetdeficit foodstamp nationaldebt houserepublican democratrepublican republicanparty checkbal drugcompani billiontax budgetcut billiondollar bottomline creditcard drugbil prescriptiondrugbil studentloan millionamerican warirag securitytrust socialsecuritytrust securitytrustfund billiontaxcut	americanpeople appropriationbil warterror iraqafghanistan warirag saddamhussein iraqipeople americanvalue billiondollar warterrorism republicanparty vahealthcare creditcard globalwar closeddoor veteranhealth houserepublican globalwarterror nationalguard federbudget americanpeopledeserv abughraib veteranhealthcare iraqwar partylinevote militaryoperation administrationoffici exportimport partyline senateintelligenceco	1.15859525	0.2403867

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
24	21	69.4	endangeredspeciact privateproperty privatepropertyowner propertyright railservice sexoffender nationalguardreserve guardreserve americanconsumer privatepropertyright sexualpredator nationaldebt largestexportmarket forestservice nationalguard naturalresource nationalendowmentart buyamerican sexoffenderregistrat oilnaturalga fueleconomy fueleconomystandard oilgadrilling budgetoffice congressionalbudgeto pensionplan americabloodcent washingtondc dutyfree terrischiavo	privateproperty endangeredspeciact propertyright privatepropertyowner columbiariver gorge postoffice enemycombatant wildlifererefuge terrischiavo urgesupport warterror taxrevenue armoredcavalryregime oilnaturalga warraq illegalimmigrant borderpatrol federbudget billiondollar lookforward passbil taxrelief ronaldreagan protectprivateproper pledgeallegiance saddamhussein consumerprotection createjob triallawyer	0.616525814	0.68326239

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
28	25	65.9	minimumwage nationalsecuritylett nationalforest patientsafety increaseminimumwage forestservice billiondollar boyscout timbersale voterregistration civilright governmentspending raisemininumwage humanlife abusedneglectedchild wageworker hurricanekatrina saletax poorpeople americanworker taxbreak millionamerican bilfail personalinformation parentnotification communitydevelopmen2 bushwhitehouse millionjob lowincome developmentblockgran	minimumwage pensionplan increaseminimumwage creditcard lowincome prescriptiondrug civilright peopledisabiliti unitedairline millionamerican foodstamp equalpay gundeal childleft wageworker incomefamili creditcardcompani cardcompani childsupport lowincomefamili wariraq childlabor educationprogram safetynet gulfcoast hurricanekatrina millionchildren raiseminimumwage gunviolence foodstampprogram	0.796534783	0.83222017

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
4	4	56.4	bordersecurity borderpatrol committehomelandsecu chemicplant illegalimmigrant driverlicense securityplan illegalimmigration strongsupport personalinformation millioncut cutfunding secretaryairforce chemicplantsecurity foodprogram joblost americanpeople presidentbudget nationalsecurityissu programhelp unofficial immigrationreform immigrationlaw educationprogram bottomline hurricanekatrina personaldata childleft lookforward personalaccount	illegalimmigration borderpatrol illegalien bordersecurity immigrationreform immigrationlaw lawenforcementagenci warterror illegalimmigrant driverlicense terrischiavo winwarteror washingtondc lookforward justicedepartment passbil sovietunion strongsupport appropriationbil amnestyillegalalien temporaryworker minoritywomenowned iwojima temporaryworkerprogr americabloodcent socialsecuritycard civilright greencard strongopposition millionillegalalien	0.610603212	0.257843964

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
15	18	54.3	hurricanekatrina gulf-coast naturaldisaster louisianamississippi passbil terrischiavo nationalfloodinsuran medicmalpractice floodinsurranceprogra bilcut lowincome-famili incomefamili americabloodcent chemicplant national-guard europeanunion taxrelief lookforward medicliability fed-eremergencymanage privateaccount freelec-tion lowincome taxre-liefact nationaldebt katrinavictim securi-tyreform southeasttexa presidentbudget judgealbertogonzale	strongsupport gulf-coast hurricanekatrina floodinsurranceprogra passengerrail nation-alfloodinsuran railsys-tem methlab passbil appropriationbil railser-vice naturaldisaster boyscout block-grant nationalguard lookforward com-munitydevelopment passengerrailsystem developmentblockgran billiondollar communi-tydevelopmen2 passen-gerrailservice postoffice boyscoutamerica lowincomeindivid-ual strongopposi-tion iraqafghanistan guardreserve afford-ablehousing president-budget	0.505417832	0.174250585

. . . continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
29	27	41.4	foreignoil transitsystem naturalresource dependenceforeignoil lookforward reducedependencefore climatechange energynaturalresourc forestservice circuitcourt nationalforest fuelefficiency presidentannounce courtappeal judicialnomine circuitcourtappeal indianaffair washingtondc fueleconomy federtradecommission federtrade oilcompani tradecommission indiantribe witnesstestify pensionplan tradepolici tradeagreement highwaybil limitedtime	wildliferefuge foreignoil nationalwildliferefu nationalwildlife drinkingwater oilcompani arcticnationalwildli militaryfamili rhodeisland arcticrefuge appropriationbil wildbird taxbreak wateract dependenceforeignoil fuelefficiency patientsafety prescriptiondrug pricegouging lowincome oildrilling iraqafghanistan arcticwildliferefuge securitytrust climatechange socialsecuritytrust guardreserve securitytrustfund cleanwateract heatingoil	0.793704749	0.066171138

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
13	17	40	taxrelief budgetcommitte financecommitte taxincrease minimumtax spendmoney alternativeminimumta federbudget raisetaxe economicgrowth lowincome increasetaxe americanpeople reformprogram taxrate middleclass millionjob budgetdeficit createjob medicaremedicaid governmentspending incomefamili nationaldefense federspending raisingtaxe budgetoffice highwaytrustfund congressionalbudgeto votecounted planreform	middleclass taxincrease taxrelief minimumtax middleclassfamili alternativeminimumta taxreliefact millionamerican taxbreak americanworker highwaytrustfund billiontaxbreak americanpeople raisetaxe economicgrowth protectsocialsecurit createjob billiontax socialsecuritybenefi pensionplan middleclasstax securitybenefit tradepolicy reliefmiddleclass tradepolici militaryfamili socialsecuritysurplu strongsupport buyamericanproduct childsupport	1.226818041	0.153467811

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
11	5	38.7	iraqipeople saddamhussein nationalguard forceiraq americanpeople wariraq warterror bringtroophome iraqafghanistan bringtroop troophome bodyarmor boyscout coalitionforce militaryfamili troopiraq warterrorism guardreserve policyiraq iraqwar missionaccomplished createjob methlab winningwar operationiraqifreedo businessowner iraqiwomen globalwar spendmoney billiondollar	oilfood oilfoodprogram iraqipeople foodprogram unreform humanlife saddamhussein iraqiwomen reformunitednation unitednationreform strongsupport internationalatomice atomicenergyagency godbless coalitionforce humanrightbody foodscandal oilfoodscandal warterror neareathobject unpeacekeeper warterrorism globalwar americancommunity europeanunion suicidebomber stpaul globalwarterror forceiraq sovietunion	0.58659888	0.110350605

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
20	28	32.2	asianpacific medicmalpractice asianpacificamerican heritagemonth nationaldebt malpracticeinsuranc2 medicmalpracticeinsu terrischiavo pacificamericanherit americancancommunity japaneseamerican middleclass insuranceindustry americanheritagemont loanprogram minimumtax alternativeminimumta strongsupport pacificislander asianpacificislander titleix malpracticeinsurance studentaid studentaidprogram strongopposition congressionalhispani coalitionforce middleclassfamili lostjob healthcarecrisi	littlerock littlerocknine asianpacific asianpacificamerican strongsupport pacificislander pacificamericanherit civilright asianamerican asianpacificislander martinluther lutherking columbiariver-gorge drmartinluther americancommunity americanheritagemont heritagemonth africanamerican brownboardeducation republicanparty urgeswiftpassage medicmalpractice johnrobert civilrightmovement rightmovement fretrade celresearch committegovernmentre asianamericanpacific privateaccount	0.183173605	0.297693253

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
14	26	29.2	veteranhealth veteranhealthcare vahealthcare healthcareveteran prescriptiondrug appropriationbil healthcareservice menthealthcare iraqafghanistan guardreserve lowincome vahospital adequatehealthcare healthcarebudget billiondollar fundingveteranhealth medicareprescription nursinghome houserepublican bilcut americanvalue forestservice federbudget wrongdirection americanconsumer rhodeisland additionalfunding educationhealth carefunding programhelp	nationaldebt studentloan childsupport billiontaxcut cutmedicaid spendingcut foodstamp billiontax veteranhealthcare loanprogram veteranhealth bottomline securitytrust paynationaldebt socialsecuritytrust securitytrustfund cutspending presidentbudget cuttaxe childsupportenforcem budgetdeficit paytaxcut programcut billiondeficit paytax budgetcut studentloanprogram seniorcitizen insurancecompani billiondollar	0.266866589	0.496511843

. . . continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
27	24	21.6	appropriationbil endingseptember lookforward presidentmove urgesupport cordblood prescriptiondrug foodstamp thirdtime environmentlaw timemove budgetcommitte warterror defenseappropriation foodstampprogram globalwar democraticleader speakertable baseclosing nationalguard nationaldefense godbless savemoney propertyright datetime strongsupport abughraib federbudget presidentbudget reformprogram	nationalforest forestservice programcut endingseptember congressionalhispani millionbudgetrequest speakertable passbil tongassnationalfores terrischiavo appropriationbil urgesupport boyscout naturalresource congressionalblackca blackcaucu timberindustry oilfield drinkingwater createjob nationalendowmentart easttexa boyscoutamerica energynaturalresourc defenseappropriation washingtondc timbersale limitedtime careeducation presidentannounce	0.698933949	0.077478591

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
22	13	11.1	iwojima strongsupport highwaytrustfund passbil consumerprotection urgesupport currentsystem boyscout postoffice privatepropertyowner rateretturn millionamerican neareathobject boyscoutamerica committegovermmentre godbless privateproperty retirementaccount driverlicense speakertable cashflow strongopposition budgetcommitte illegalimmigration savemoney illegalimmigrant programhelp politicparty lookforward hugochavez	sexoffender driverlicense spendmoney highwaytrustfund southeasttexa sexoffenderregistrat bilfail privatepropperty patientsafety terrischiavo strongsupport middleincome postoffice sexualpredator blessamerica millionbudgetrequest godblessamerica godbless methlab illegalalien abusepower fifthcongressionaldi undergroundstorageta democratrepublican greencard singleday boygirlclub childrenfamili worldtrade humanrightbody	1.148778058	0.150059881

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
17	2	9.6	affordablehousing floodinsuranceprogra nationalfloodinsuran terrorismriskinsuran ryanwhitecare housingmarket lowincome whitecareact voterregistration incomepeople communitydevelopment familyvalue civilright insuranceindustry nationalhomeownershi lowincomepeople littlerock africanamerican federtradecommission federtrade strongsupport hatecrime communitydevelopmen2 tradecommission developmentblockgran incomefamili blockgrant republicanparty poorpeople indiantribe	africanamerican congressionalblackca blackcaucu africanamericanicancommu americanicancommunity africanamericanwomen percentafricanamericvotingmachine hurricanekatrina martinluther lutherking democratrepublican naturaldisaster armembargo wariraq lowincome nationalguard ryanwhitecare whitecareact urgeswiftpassage forceiraq jobcreation gulfcoast centdiseasecontrol drmartinluther privateaccount votecounted medicmalpractice urgesupport washingtondc	0.964676952	0.02951549

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
18	14	5.4	illegalalien illegalimmigrant immigrationlaw illegalimmigration godbless immigrationreform enforceimmigrationla driverlicense oilfoodprogram nationalhomeownershi oilfood millionillegalalien businessowner unitednationreform patientsafety circuitcourtappeal strongsupport foodprogram heritagemonth circuitcourt housingmarket thirdtime courtappeal taxrevenue americanworker millionamerican lookforward prescriptiondrug internationalpeacese justicedepartment	medicliability medicliabilityreform medicliabilitycrisi liabilityinsurance medicliabilityinsura prescriptiondrug armembargo malpracticeinsuranc2 medicmalpractice medicmalpracticeinsu patientsafety parentnotification saddamhussein bankruptcycourt americanpeople raisetaxe largestexportmarket armembargochina medicareprescription socialsecuritysystem freenterprisesystem taxrevenue celline freenterprise securitysystem hurricanekatrina legalsystem judicialnomine insurancecompani strongsupport	1.157975408	0.136694248

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
23	20	2.8	passengerrail railsystem railservice passengerrailservice passengerrailsystem hazardoumaterial communitydevelopment nationalpassengerrai clinicsocialworker billiondollar indiantribe highwaytrustfund appropriationbil japaneseamerican railpassengerservice rosapark blockgrant gunviolence strongsupport consumerprotection gunsafety indianaffair missionaccomplished creditcard nationalwildlife livingpoverty rateretурn peopleliving-poverty gunsafetylaw naturaldisaster	estatetax taxrelief medicmalpractice malpracticeinsuranc2 armoredcavalryregime iraqipeople medicmalpracticeinsu tradedeal nationalguard hurricanekatrina malpracticeinsuranc3 saletax gulfcoast securitytrustfund socialsecuritytrust securitytrust tradedeficit operationiraqifreedo politicparty landwaterconservatio dividendtaxcut familybusiness malpracticeinsurance privateaccount servingcountry unreform safetynet taxrepeal budgetcut financialaccountings	0.537475423	0.657449032

...continued

CR1 Index	CR2 Index	Similarity Est.	CR1 Text	CR2 Text	Observed Distance	Null Distance
19	29	1.9	guestworker guestwork-erprogram illegalimmi-gration bordersecurity immigrationreform armeniangenocide forestservice passbil nationalwildliferefu nationalwildlife wildlif-erefuge americanpeople illegalimmigrant chiefjusticerehnquis democratrepublican temporaryworker millionillegalimmigr taxrate federincometax rulechange terrischiavo justicedepartment energynaturalresourc fifthcongressionaldi climatechange con-gressionalhispani presidentannounce gunindustry civilright nationalforest	businessowner loan-program createjob minorityownedbusines minoritywomenowned jobcreated education-program programhelp budgetsurplu middlein-come budgetdeficit healthcarecrisi bil-liondollar singleday rhodeisland insurance-compani american-people bottomline womenarmedforce budgetoffice con-gressionalbudgeto millionamerican healthcarecoverage littlegirl baseclosing lowincome president-business jobcreation africanamerican health-saving	0.624387838	0.133672344

Table B.4: Matched Topics and Match Statistics (News)

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
1	1	95.3	stemcel embryonicstem embryonicstemcel celresearch adultstem adultstemcel cordblood humanembryo celline illegalimmigrant stemcelline nationaladcampaign driverlicense growthrate drinkingwater pricegouging armeniangenocide humanlife oildrilling prescriptiondrug bilcut americanpeople educationprogram unitedairline strongopposition passengerrailsystem humancloning palestinianauthority umbiliccord californiasupremecou	stemcel embryonicstem embryonicstemcel celresearch cordblood adultstemcel adultstem bloodstemcel cordbloodstem celline humanembryo stemcelline humanlife umbiliccord humancloning americanpeople percentgrowth terrischiavo millionamerican unbornchild malpracticeinsuranc2 oilcompani naturalga umbiliccordblood nuclearpower childleft tencommandment producestemcel centralamerican embryostemcel	2.425386417	0.477746021

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
18	5	92.3	headstart forestservice educationprogram jobtrainingprogram lowincome financialmanagement borderpatrol lostjob strongsupport hurricanekatrina blockgrant programhelp presidentbudget lookforward lowincomechildren bringtroop troophome studentaid childleft bringtroophome paycollege strongopposition educationhealth federelection healthcarevetteran childrenfamili cutfunding iraqwar civilrightprotection centdiseasecontrol	headstart civilright educationprogram childleft stpaul businessowner americanworker insurancecompani minimumwage lowincome bordersecurity headstartteacher childrenfamili studentaid collegestudent committehomelandsecu loanprogram foodprogram millionjob millionamerican programhelp millionjobcreated strongopposition lowincomechildren affordablehousing iraqafghanistan electitycouncil victimdomestic americanpeople africanamericanwomen	1.763858225	0.136942237

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
4	0	92.2	tradeagreement fre-trade centralamer-ican tradedeficit americanfretrade centralamericanfre tradepolicy labor-law americanworker jobloss strongoppo-sition propertyright abughraib dutyfre workerright joblost paycollege manufac-turingjob unitedairline wariraq tradepolici tradedeficitchina illegalimmigration socialsecuritybenefi internationallaboror civilright security-benefit blockgrant middleclass studentaid	tradeagreement fre-trade centralamerican americanfretrade centralamerican-fre americanworker manufacturingjob tradepolicy trad-edeficit worldtrade tradepolici tradedeal worldtradeorgani-zati americanpeople democratrepublican jobloss workerright cre-atejob lostjob dutyfre joblost economic-growth blessamerica propertyright fasttrack northamericanfre lose-job joboversea taxrelief laborlaw	1.686247427	0.259041074

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
5	25	89.7	publicbroadcasting corporationpublicbro low-income americanpeople oilfood affordable-housing postalservice hurricanekatrina sovietunion publictele-vision oilfoodprogram lowincomefamili pro-gramhelp foodprogram bilcut incomefamili helpveteran chiefjustice educationprogram naturaldisaster mil-lionchildren united-postalservice unreform endingseptember centdiseasecontrol safetynet celresearch billiondollar medi-caremodernizatio hundredbilliondollar	publicbroadcasting publictelevision pa-tientsafety corpora-tionpublicbro foreignoil floodinsuranceprogra millionamerican postof-fice reliefmiddleclass middleclassfamili drink-ingwater oilcompani naturalga boyscout dependenceforeignoil fuelefficiency traded-eficit sexualpredator taxrate landwater-conservatio safedrink-ingwater heatingcost columbiarivergorge naturalresource save-money lowincome fuelefficiencystanda gasolineprice com-mittegovernmentre boyscoutamerica	1.184991258	0.112704754

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
9	9	88.6	tuskegeairmen africanamerican hazardoumaterial iwojima civilright minorityownedbusines gulfcoast unitedairforce civilrightmovement safetystandard rightmovement americanpeople affordablehousing alternativeminimumtax minimumtax borderpatrol strongsupport committegovernmentre justicedepartment courtappeal illegalalien nationaldefense hurricanekatrina increasetaxe whitecareact ryanwhitecare raisingtaxe judicialnomine warterror justicejaniceroger	tuskegeairmen lawenforcementagenci communitydevelopment africanamerican driverlicense communitydevelopmen2 developmentblockgran blockgrant professionalsport votingsystem americanpeople littlegirl methlab propertyright troopbringhome familybusiness vahealthcare nationalsecuritylett deathtax spendmoney ryanwhitecare naturalga taxcutpeople vahospital stressdisorder traumaticstressdisor whitecareact wariraq posttraumaticposttraumaticstress	2.105044502	0.335041288

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
6	8	86.4	deathtax familybusiness estatetax classaction taxrepeal deathtaxrepeal farmbureau repealdeathtax committehomeland-secu businessowner actionlawsuit humanlife judgerobert americabloodcent hugochavez taxrate economicgrowth humanembryo americanpeople houserepublican postoffice adultstem middleclass votingright adultstemcel legalsystem saletax votingrightact urgesupport lookforward	deathtax familybusiness urgesupport naturaldisaster privateproperty businessowner taxrate repealdeathtax commonsense-form federspending taxrelief createjob privatepropertyright dutyfre nationaldebt centralamerican endangeredspeciact madisoncounty propertyright borderpatrol economicgrowth illegalimmigrant medicliability appropriationbil carefunding deathtaxrepeal circuitcourt taxrepeal minoritylleader enemycombatant	0.994531465	0.138817928

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
12	21	86.3	domesticviolence violencewomen violencewomenact victimdomestic victimdomesticviolen pacificislander climatechange nuclearoption asianpacificislander violencesexualassaul domesticviolencesexu passengerrail budgetcut cutmedicaid railsystem foreignrelation childsupport indiantribe indianaffair medicaidcut strongopposition asianamerican minorityright karlrove lowincome titleix passengerrailsystem millionamerican undergroundstorageta oildrilling	domesticviolence victimdomestic victimdomesticviolen violencewomen violencewomenact iraqiwomen privateproperty violencesexualassaul sexualpredator nationalsecuritylett propertyright sniperrifle ownparty privatepropertyright domesticviolencesexu angelecitycouncil personalinformation affordablehousing houserepublican provisionalballot justicedepartment boyscout committehomelandsecu socialsecurityreform securityreform illegalimmigrant racialethnicminoriti socialsecuritysystem businessowner celresearch	0.563459783	0.055585666

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
2	2	83.1	rosapark civilright civilrightmovement rightmovement africanamerican highwaybil cordblood votingsystem lutherking martinluther foodstamp nationalheritagecorr highwaytrustfund railpassengerservice votingright civilrightprotection railservice votingrightact drmartinluther budgetcut congressionalblackca blackcaucu republicanparty americanpeople collegestudent familybusiness passengerrailservice spendmoney millionchildren domesticviolence	rosapark civilright civilrightmovement rightmovement armeniangenocide martinluther lutherking africanamerican littlerock servingcountry governmentresponse medicliability solarenergy hatecrime manufacturingjob diseasecontrolpreven centdiseasecontrol americancommunity africanamericancommu racialethnicminoriti americanpeople universalhealth railservice freelection drmartinluther naturalresource millionamerican littlerocknine jobloss illegalimmigrant	0.83423249	0.036230154

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
3	16	80.3	illegalalien illegalimmigrant illegalimmigration immigrationlaw guestworker immigrationreform godbless guestworkerprogram strongsupport temporaryworker nationalhomeownershi enforceimmigrationla affordablehousing heritagemonth housingmarket millionillegalalien temporaryworkerprogr lookforward driverlicense oilfood americanworker thirdtime unitednationreform oilfoodprogram bordersecurity nationalsecuritylett businessowner foodprogram millionamerican millionillegalimmigr	illegalalien illegalimmigration driverlicense immigrationreform bordersecurity greencard socialsecuritycard lawenforcementagenci illegalimmigrant americabloodcent passbil boyscout lookforward postoffice iwojima immigrationlaw medicliability guestworker currentsystem transitsystem federtradecommission federtrade loanprogram tradecommission americanpeople circuitcourtappeal highwaybil celresearch senatewhitehouse triallawyer	1.086328596	0.10708991

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
20	23	79	nuclearweapon pen-sionplan taxbreak taxbreakcompani secretaryairforce armeniangenocide wariraq presidentbud-get saddamhussein bankruptcycourt bunkerbuster strongop-position drugcompani billiondollar sovietu-nion appropriationbil strongsupport abuse-power menthealthcare troopiraq bringtroop bringtroophome hazardoumaterial republicanparty studentloan defenseau-thorization troophome nuclearbunkerbuster corporationpublicbro insurancecompani	nuclearweapon is-suefacingamerican millionbudgetrequest europeanunion lawen-forcementagenci bottomline unre-form appropriationbil saletax lookforward bunkerbuster cred-itcard tradedeficit endingseptember es-tatetax nuclearpower economicgrowth gov-ernmentspending atomicenergyagency internationalatomice americanpeople fre-trade billiondollar nationalwildlifer-efu nationalwildlife wildliferefuge de-fenseauthorization growingeconomy oilga-exploration oildrilling	0.985565301	0.297367949

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
13	13	74.9	africanamerican africanamerican-commu americancommunity africanamericanwomen armembargo nationalguard percentafricanamericvotingmachine lutherking martinluthervotingright jobcreation votingrightactcivilright naturaldisaster urgeswiftpassage drmartinluther littlerock economicgrowth urgesupport businessowner forceiraq committegovernmentre joblost iraqipeople hurricanekatrina levelingplayingfield legalsystem lookforward minimumwage	africanamerican blackcaucu congressionalblackca votingmachine hazardoumaterial africanamericanwomen rosapark democratrepublican votecounted lowincome familyvalue ryanwhitecare whitecareact africanamericancommu voterregistration earthday ohioelectoralvote postoffice provisionalballot washingtondc immigrationlaw politicparty poorpeople americancommunity americanworker wateract electionreform gulfcoast rhodeisland cleanwateract	0.525648983	0.153031964

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
8	6	72.6	postoffice postalservice unitedpostalservice committegovernmentre urgesupport foodstamp iwojima tencommandment passbil pledgeallegiance civilright borderpatrol foodstampprogram lookforward planreform commonlyprescribeddr urgeswiftpassage camppendleton spendmoney rightmovement strongsupport civilrightmovement educationprogram privateproperty terrischiavo sexualpredator temporaryworker nationaldebt abughraib outingciaagent	postalservice postoffice unitedpostalservice americanworker washingtondc lowincome dutyfre committegovernmentre driverlicense rateincrease voterregistration incomefamili memorialday chiefjudge millionamerican drinkingwater minimumwage lowincomefamili defenseauthorization iwojima budgetdeficit heatingoil socialsecuritybenefi temporaryworker securitybenefit committehomelandsecu spendmoney rhodeisland millionjob borderpatrol	0.554775267	0.226375061

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
23	18	70.1	sexoffender highwaytrustfund driverlicense sexoffenderregistrat sexualpredator estatetax nationalguardreserve urgesupport methlab strongsupport guardreserve fifthcongressionaldi nationalguard currentsystem boygirlclub patientsafety postoffice triallawyer savemoney madisoncounty tencommandment globalwarterrorism endangeredspeciact terrischiavo globalwar seriouconcern minorityleader presidentplan repealdeathtax speakertable	sexoffender privateproperty spendmoney millionbudgetrequest bilfail businessowner middleincome undergroundstorageta driverlicense patientsafety taxrevenue abusepower singleday postoffice strongsupport privatepropertyright personalaccount cleanairact propertyright sexualpredator economygrowing naturalga housingmarket blessamerica godblesslessamerica godbless personalsavingaccoun nationalguard seniorcitizen lookforward	1.585811368	0.657325615

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
0	4	68.6	americanpeople privateaccount oilcompani wariraq drugcompani billiondollar democraticleader creditcard houserepublican millionamerican democratrepublican republicanparty tradedeficit studentloan iraqipeople invasioniraq starttalking pricegouging iraqwar oilindustry troophome bringtroophome middleclass bringtroop pharmaceuticindustry ciaagent republicanleader socialsecuritysurplu abughraib borrowmoney	americanpeople privateaccount nationaldebt foodstamp prescriptiondrug studentloan billiontaxcut karlrove childsupport billiontax republicanparty hurricanekatrina presidentplan budgetcut cutmedicaid partyline middleclass securitytrust socialsecuritytrust taxbreak ciaagent bottomline securitytrustfund housere-publican budgetdeficit veteranhealth privatizationplan oilcompani presidentbudget wariraq	1.53882893	0.240354825

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
7	3	66.9	naturalga oilnaturalga energynaturalresourc oilindustry naturalresource foreignoil presidentannounce hearingscheduled businessowner strongsupport gaoil naturalgaoil climatechange nuclearpower terrorismriskinsuran witnesstestifyappropriationbil urgesupport heartmind dependenceforeignoil committegovernmentre privatepropertyright fuelefficiency presidentcheney vicepresidentcheney insuranceindustry limitedtime washingtondc pricenaturalgapropertyright	naturalga climatechange financecommitte highwaybil naturalresource economicgrowth ratereturn highwaytrustfund solarenergy energynaturalresourc courtappeal lowincome presidentmove spendmoney rhodeisland cleanenergy circuitcourt creditcard billiondollar oilnaturalga savemoney fueleconomy tradedeficit warterrorism budgetcommitte lookforward createjob appropriationbil enemycombatant methlab	0.102895771	0.063218165

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
17	28	66.4	personalaccount personalretirementac retirementaccount socialsecuritysystem climatechange securitysystem iwojima rateretur currentsystem socialsecurityreform securityreform nearretirementage socialsecurityprogra reformsocialsecurity socialsecuritybenefi securitybenefit humanlife terrischiavo globalwar cashflow lookforward memorialday godbless americanpeople benefitsocialsecurit raisetaxe sanctityhumanlife securitytrustfund socialsecuritytrust securitytrust	socialsecuritysystem securitysystem personalaccount socialsecuritybenefi securitybenefit taxincrease reformsocialsecurity ratereturn seniorcitizen securitytrustfund securitytrust socialsecuritytrust americanpeople socialsecuritypro- gra currentsystem taxrelief increasetaxe economicgrowth palestinianauthority benefitcut raisingtaxe savingaccount socialsecuritysreform cutbenefit securityreform spendingsocialsecuri socialsecuritypresid savemoney americanworker taxrevenue	0.349989375	0.028430369

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
25	20	55.5	nationalforest forest-service timberindustry programcut naturalresource timbersale tongassnationalfores oilgaexploration foreignoil defenseappropriation energynaturalresourc wildliferefuge lookforward judicialnomine presidentmove highwaybil minorityleader oilfield washingtondc climatechange postalservice oilproduction appropriationbil assistantsecretary transitsystem childleft indianaffair nationalwildlife presidentannounce committecommercescie	littlerock nationalforest forestservice terrischiavo passbil littlerocknine congressionalhispani nationalsecuritylett washingtondc boyscout createjob civilright foreignrelation democratrepublican drinkingwater urgesupport minimumwage nationalheritagecorr familyvalue humanlife unitedsupremecourt speakertable borders ecurity parentright congressionalblackca changerule blackcaucu climatechange privatepropertyowner humanembryo	1.350914128	0.368630625

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
29	11	53.8	taxincrease taxrelief middleclass raise-taxe budgetcommitte lowincome taxre-liefact pensionplan financecommitte createjob federbud-get americanpeople reformprogram mini-mumtax growthrate economicgrowth al-ternativeminimumta governmentspending millionjob increasetaxe billiontax sovietunion democraticleader socialsecurityprogra federspending spend-money jobcreation passbil socialsecuri-tybenefi socialsecuri-tysystem	middleclass minimum-tax alternativemini-mumta middleclass-famili taxincrease taxrelief highwaytrust-fund passbil billion-taxbreak taxreliefact protectsocialsecurit socialsecuritybenefi securitybenefit eco-nomicgrowth taxbreak marketaccess childsup-port millionamerican japaneseamerican worldtradeorganizati billiontax reliefmid-dleclass cuttaxe worldtrade civil-right raisingtaxe unitedtraderepresent growingeconomy lowin-come lowincomefamili	0.605487266	0.066003331

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
24	22	52.3	medicmalpractice terrischiavo malpracticeinsuranc2 medicmalpracticeinsu mediclibility hurricanekatrina insuranceindustry liabilityinsurance medicliabilityreform privateproperty privateaccount malpracticeinsuranc3 americanpeople strongsupport servingcountry armembargo healthcarecrisi landwaterconservatio legalsystem floodinsuranceprogra gulfcoast privatepropertyright congressionalbudgeto budgetoffice insurancecompani paytaxcut medicliabilityinsura rateincrease nationalfloodinsuran propertyright	estatetax malpractice malpracticeinsuranc2 medicmalpracticeinsu insurancecompani rhodeisland malpracticeinsurance malpracticeinsuranc3 workerright asianpacific drugcompani asianpacificamerican minimumwage naturalga pharmaceuticindustry dividendtaxcut millionamerican increaseminimumwage stpaul insuranceindustry tradedeal healthcarecrisi lostjob titleix taxrelief americanpeople economicgrowthjob hurricanekatrina studentaid patientsafety	0.1016638249	0.6833342149

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
28	27	50.8	minimumwage nucleaoption judgerobert judicialnomine classaction creditcard guardreserve courtappeal foreignoil circuitcourt civilright changerule chiefjustice nationalguard checkbal dependenceforeignoil lookforward gundeal financecommitte circuitcourtappeal justicedepartment americanworker gasolineprice bankruptcycourt pensionplan increaseminimumwage presidentprotempore defenseauthorization improvisedexplosived cardcompani	chiefjustice nationalsecuritylett judicialnomine judgerobert circuitcourt foreignrelation courtappeal roewade courtjudge lookforward thirdtime parentnotification circuitjudge democraticleader chiefjusticerehnquis circuitcourtappeal nuclearoption majorityvote johnrobert creditcard humanlife pensionplan supremecourtunited worldtrade datetime holdhearing senatetelligenceco tradeagreement timevote financecommitte	1.378479049	0.832338952

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
11	19	49.1	saddamhussein iraqipeople iraqafghanistan warterrorism warterror globalwar godbless coalitionforce globalwarterrorism fightingwarterror nationalsecurityissu oilfood guardreserve forceiraq freelection militaryoperation humanlife iraqiwomen lawenforcementintell globalwarterror oilfoodprogram troopiraq foodprogram enemycombatant medicliability milionjobcreated iraqwar appropriationbil lookforward womenarmedforce	warterror americanpeople globalwar globalwarterror appropriationbil iraqipeople cordblood saddamhussein federbudget urgesupport defenseappropriation timemove grossnationalproduct nationalguard sovietunion tradedejustmentassis minorityleader winningwar reformunitednation umbiliccordblood ronaldreagan warterrorism freelection violencewomen lutherking martinluther globalwarterrorism nationaldefense terrischiavo lookforward	0.180695718	0.110357603

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
16	29	47.9	strongsupport passbil railsystem hurricaneka-trina passengerrail gulfcoast railservice blockgrant passenger-railsystem naturald-isaster urgesupport currentsystem com-munitydevelopment passengerrailservice developmentblockgran billiondollar com-munitydevelopmen2 highwaytrustfund classaction palestinia-nauthority savemoney millionamerican strongopposition tran-sitsystem lowincomein-dividual cutfunding floodinsuranceprogra lookforward appropria-tionbil bottomline	strongsupport meth-lab europeanunion committegovernmen-tre strongopposition budgetcommitte iraqi-women iraqipeople iwojima economic-growth tradedeficit assistantsecretary fasttrack endmeet federtradecommission federtrade tradecom-mission unreform americancommu-nity memorialday consumerprotec-tion currentsystem atomicenergyagency internationalatomice tradepolicy freelection urgesupport politic-party democratrepubli-can asianpacific	0.693536278	0.741738556

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
27	12	45.7	privateproperty endangeredspeciact privatepropertyowner propertyright iwojima columbiariver gorge privatepropertyright warterror wildlifer efuge terrischiavo enemycombatant healthcareveteran illegalimmigration naturalresource oilnatural ga womenarmedforce ronaldreagan hugochavez timevote iraqiwomen strongsupport sovietunion dutyfree levelingplaying field lookforward borderpatrol majorityvote fretrade pledgeallegiance timberindustry	endangeredspeciact oilgacompani americanconsumer gacompani appropriationbil pensionplan lowincome foreignoil wildlifer efuge fueleconomy fueleconomystandard nationalwildliferefu largestexportmarket environmentlaw lookforward nationalwildlife billcut privateproperty buyamerican railservice billiondollar forestservice fuelefficiency programcut nationalendowmentart strongsupport dependenceforeignoil landwaterconservatio nationalendowmenthum strongopposition	0.177670965	0.077852927

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
19	10	29.8	appropriationbil endingseptember veteranhealth veteranhealthcare lookforward vahealthcare speakertable presidentmove hurricanekatrina forestservice millionbudgetrequest naturaldisaster budgetcommitte gulfcoast driverlicense iraqafghanistan urgesupport healthcareveteran lowincome palestinianauthority warterror environmentlaw additionalfunding healthcarebudget rhodeisland nationalguard americanpeople godbless federbudget drinkingwater	hurricanekatrina gulfcoast louisianamissisippi naturaldisaster appropriationbil taxrelief nationalguard passbil floodinsuranceprogra nationalfloodinsuran guardreserve postoffice lookforward incomefamili lowincomefamili timemove bilcut lowincome chemicplant europeanunion thirdtime freelection billiondollar taxreliefact securityplan judgealbertogonzale financecommitte americancommunity saddamhussein urgesupport	0.122807501	0.134067835

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
21	17	27.5	congressionalblackca blackcaucu affordable-housing hurricanekat-rina presidentbudget africanamerican low-income minimumwage communitydevel-opment commu-nitydevelopmen2 developmentblockgran blockgrant presi-dentpropose childleft committehomeland-secu ryanwhitecare whitecareact gulfcoast lowincomechildren educationprogram menthealthcare bud-getcut cutfunding mil-lioncut americanpeople veteranhealthcare vet-eranhealth voterregis-tration reformprogram federbudget	affordablehousing floodinsuranceprogra nationalfloodinsuran terrorismriskinsuran civilright insurancein-dustry incomepeople housingmarket lowin-come lowincomepeople republicanparty hatecrime federtrade-commission federtrade voterregistration strongsupport appro-priationbil billiondollar palestinianauthority tradecommission bor-rowmoney lookforward communitydevelop-ment jointcommitte spendmoney so-cialsecurityprogra worldtrade reformpro-gram minimumwage middleincomefamili	0.688690496	0.555634168

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
10	15	19.5	prescriptiondrug patientsafety classaction commonlyprescribeddr medicaremedicaid baseclosing medicareprescription environmentlaw republiccypru pharmaceuticcompani urgesupport costprescriptiondrug pharmaceuticindustry drugbil medicaremodernizatio laborlaw lookforward blockgrant lowincome marketaccess seniorcitizen centralamerican prescriptiondrugbil reformprogram savemoney appropriationbil federtradecommission committebusinessentr senatecommittebusine federtrade	classaction actionlawsuit pharmaceuticindustry cuttaxe classactionreform madisoncounty classactionfairness financecommitte californiasupremecou americanpeople createjob actionfairnessact courtappeal nationalguard taxrelief democraticleader bankruptcycourt legalsystem pharmaceuticcompani illegalalien taxincrease jointcommitte lawreview iraqipeople washingtondc creditcard judgerobert iraqafghanistan federbudget bankruptcyreform	0.568862886	0.055971025

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
26	26	18.9	nationaldebt spending cut middleclass studentloan loanprogram presidentbudget iraqafghanistan taxbreak securitytrust securitytrustfund socialsecuritytrust veteranhealthcare veteranhealth bottomline congresswhitehouse alternativeminimumtax childsupport minimumtax budgetdeficit cutbil foodstamp estatetax billiondeficit tradedeficit spendingcutbil budgetcut endangeredspeciact billiontax studentaid nationalguard	militaryfamili iraqipeople nationalguard servingcountry forceiraq bodyarmor wariraq operationiraqifreedo democratrepublican loanprogram remembereddfamilyfrie americanpeople guardreserve iraqafghanistan policyiraq coalitionforce missionaccomplished troopiraq federelection whitehousecongress troophome bringtroop growingeconomy saddamhussein bushwhitehouse bringtroophome abughraib fuelefficiencystanda presidentcheney vicepresidentcheney	1.348307196	0.079197108

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
14	7	8.2	bordersecurity borderpatrol committehomelandsecu illegalimmigration driverlicense illegalimmigrant terrischiavo guestworker securityplan hurricanekatrina chemicplant secretaryairforce unofficial millioncut chiefjustice nationalsecurityissu programhelp naturaldisaster joblost bottomline cutfunding immigrationreform americanpeople lookforward worldtrade governmentresponse socialsecuritycard californiasupremecou guestworkerprogram gulfcoast	oilfood oilfoodprogram foodprogram unreform neareathobject saddamhussein humanrightbody stpaul foodscandal oilfoodscandal reformunitednation americanpeople unitednationreform americancommunity internationalatomice atomicenergyagency sovietunion unofficial foreignrelation iraqipeople unpeacekeeper nationoilfood unitednationoil strongsupport internationalpeacese tencommandment humanembryo iraqiwomen largestexportmarket nunnlugarprogram	0.458592364	0.497009605

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
15	24	3.7	asianpacific asianpacificamerican heritagemonth columbiariver-gorge pacificislander pacificamericanherit americancommunity americanheritagemont asianpacificislander asianamerican japaneseamerican medicmalpractice civilright asianamericanpacific loanprogram strongsupport americanpacificislan nationaldebt insuranceindustry coalitionforce lookforward urgeswiftpassage committegovernmentre ohioelectoralvote patientsafety globalwar securitytrustfund socialsecuritytrust securitytrust abusepower	gunviolence gundeal guninduistry buygun terroristwatchlist assaultweapon watchlist rhodeisland calibersniperrifle sniperrifle gunsafety ablebuygun transitsystem americanpeople deepseacoral creditcard assaultweaponban flagdrapedcoffin nuclearoption buleye changerule iraqafghanistan climatechange victimgunviolence backgroundchecksyste lowincome gunlobby defenseauthorization troophome millionamerican	1.388433622	0.173919512

...continued

News1 Index	News2 Index	Similarity Est.	News1 Text	News2 Text	Observed Distance	Null Distance
22	14	1.6	boyscout boyscoutamerica supportboyscout chiefjustice nationalfloodinsuran appropriationbil floodinsuranceprogra permanentresidentsta iraqipeople greencard tencommandment iraqafghanistan businessowner militaryfamili healthsavingaccount healthsaving unitedsupremecourt iraqweaponmass savingaccount americanworker sovietunion senateintelligenceco parentright iraqiwomen nationaldefense urgesupport iwojima globalwarterrorism districtjudge personalaccount	passengerrail railsystem railservice passengerrailservice passengerrailsystem hazardoumaterial communitydevelopment nationalpassengerrai reverserobinhood clinicsocialworker missionaccomplished taxbreakrich americanpeople japaneseamerican blockgrant billiondollar taxbreak creditcard provisionalballot appropriationbil rosapark indiantribe civilright postoffice livingpoverty gulfcoast republicanparty peoplelivingpoverty cutfunding communitydevelopmen2	0.927622807	0.149907892

Table B.5: Automated Topic Summaries for SOTS–SOTU Topic Model

Topic	Phi	Lift	Relevance	FREX
1	waste, environmental, environment, clean, pollution, site, issue, hazardous, recycle, solid	recycle, toxic, pollutant, pollute, hazardous, aquifer, mitigate, waste, disposal, cleanup	waste, recycle, environmental, hazardous, pollution, toxic, clean, disposal, pollute, environment	institution, enterprise, envision, energize, resilient, incubator, entrust, embrace, privately, maximize
2	crime, criminal, police, enforcement, drug, prison, offender, juvenile, system, public	abuser, vicious, harden, prosecutor, illicit, trooper, revoke, habitual, violator, prosecution	crime, police, criminal, prosecutor, trooper, enforcement, offender, abuser, juvenile, violent	mansion, quality, supervise, reluctant, crash, esteem, portfolio, equalize, taxable, corporation
3	mental, institution, hospital, mentally, health, service, developmental, retardation, staff, treatment	remiss, custodial, concurrence, developmental, cation, mental, retardation, psychiatric, alcoholism, alcoholic	mental, developmental, concurrence, remiss, retardation, cation, custodial, mentally, psychiatric, alcoholism	recreation, formally, simultaneously, local, tyranny, ratification, goal, reflect, penalize, outstrip
4	citizen, problem, economic, responsibility, public, service, nation, challenge, goal, past	substandard, eral, realization, tyranny, enviable, translate, shirk, solemn, declaration, gratify	eral, substandard, realization, translate, tyranny, restraint, enviable, shirk, citizen, declaration	empower, mat, darkness, country, roof, commodity, tit, praise, overwhelmingly, wholesale
5	insurance, fault, public, lawsuit, tort, sue, cost, liability, automobile, incur	frequency, fault, tort, justifiable, severity, incur, sue, execute, punitive, lawsuit	fault, tort, sue, frequency, incur, lawsuit, justifiable, execute, severity, punitive	hate, dimension, esteem, fairly, steve, lou, tory, firm, pile, directive
6	taxpayer, income, percent, reduce, cost, pension, rate, health, family, save	pur, transparent, intangible, pension, overhead, loophole, simplify, median, taxed, furlough	pur, pension, taxpayer, transparent, loophole, simplify, overhead, corporate, intangible, cap	sanction, counting, designation, lower, invention, prudently, endure, manageable, fairness, impasse
7	coastal, textile, deduction, mend, exemplary, holiday, inheritance, pocketbook, mac, sentiment	textile, exemplary, holiday, mac, pocketbook, coastal, sentiment, conformity, deduction, inheritance	textile, exemplary, coastal, holiday, mac, pocketbook, deduction, sentiment, inheritance, mend	personality, payment, systematic, deaf, complacency, waive, multiply, complacent, speedy, occupational
8	governor, legislature, commission, citizen, system, constitutional, race, public, agency, dirt	dirt, commence, conscientious, maze, distinct, ratification, proponent, tri, printing, governance	dirt, commence, conscientious, distinct, respective, maze, printing, governance, ratification, tri	recipient, ward, heretofore, reimbursement, dwell, originate, lip, lessen, regulation, educate
9	thank, life, let, today, family, help, governor, serve, join, men	selfless, laughter, uphold, glory, freedom, tha, uncommon, chris, ethnic, deployment	selfless, uphold, laughter, thank, courage, legacy, freedom, please, tha, inspire	bigotry, bob, live, page, goodness, facet, embody, whatsoever, neighbor, energize
10	fuel, agriculture, production, renewable, electricity, clean, consumer, produce, farm, ethanol	soybean, diesel, bushel, ethanol, electricity, electrical, consumption, fleet, producer, wine	fuel, ethanol, electricity, renewable, diesel, soybean, producer, electrical, production, emission	ski, risk, recycle, learner, persevere, homemaker, guideline, remarkable, invaluable, haul

(continued)

Topic	Phi	Lift	Relevance	FREX
11	road, highway, bridge, system, interstate, maintenance, construction, turnpike, traffic, fund	turnpike, pavement, bridge, roadway, pave, ministration, road, resurface, widen, freight	road, bridge, turnpike, pave, roadway, pavement, interstate, widen, highway, ministration	military, lust, forfeiture, range, legislature, downstate, heretofore, news, lesser, goodness
12	map, sail, flu, rain, andrea, timid, acid, seniority, trauma, flower	flu, sail, timid, acid, andrea, seniority, map, flower, rain, unavoidable	flu, sail, timid, map, acid, rain, andrea, seniority, flower, trauma	laboratory, hub, donate, seniority, rum, tutor, elector, pastor, educational, flight
13	cut, chart, revenue, sale, look, reduction, problem, today, criticize, fund	criticize, tab, criticism, chart, nat, blame, basically, straightforward, incredibly, genuinely	chart, criticize, cut, criticism, blame, basically, tab, incredibly, nat, headline	smoking, plate, staffed, residence, seniority, servant, lastly, enrich, reduce, sea
14	insurance, auto, rate, premium, border, fun, lull, company, unsustainable, unfunded	lull, hide, asa, unsustainable, paradox, fun, radical, auto, immigration, dick	auto, lull, hide, unsustainable, fun, insurance, immigration, asa, radical, paradox	institutional, solid, football, hope, empower, religion, earnings, larry, guide, ich
15	ton, nation, fight, help, wan, today, disaster, economic, terrorist, threat	wan, diane, dig, shy, wash, structurally, ton, devastation, terrorist, cheap	wan, ton, wash, dig, terrorist, diane, shy, structurally, devastation, terrorism	worker, pursuit, sat, muster, direction, dentist, elector, geography, pride, handicapped
16	highway, transportation, fund, bond, construction, project, system, capital, improvement, road	expressway, gentle, upstate, plastic, route, artery, lane, parking, bond, hasten	highway, bond, expressway, transportation, construction, route, upstate, gentle, construct, interstate	recruiting, early, duo, revolution, physically, edge, dime, mediocrity, dynamic, overdue
17	institution, regent, college, university, faculty, system, mode, award, public, subcommittee	ewe, mode, suburban, bum, endorsement, regent, intermediate, subcommittee, faculty, mat	mode, ewe, regent, suburban, endorsement, subcommittee, faculty, bum, intermediate, institution	halt, reach, headline, guardsman, frugality, grader, fun, revolution, finding, lure
18	addiction, addict, heroin, narcotic, treatment, locality, decisively, lifeline, caliber, accordingly	heroin, addiction, caliber, addict, lifeline, decisively, narcotic, locality, suburb, jobless	heroin, addiction, addict, caliber, lifeline, narcotic, decisively, locality, suburb, jobless	emphasis, edge, obvious, quietly, expedient, freeze, importantly, earn, productive, exposure
19	tag, yard, battery, service, mouth, installment, keith, funeral, salesman, immigrant	yard, mouth, battery, edge, keith, salesman, tag, funeral, installment, batter	tag, yard, battery, mouth, keith, salesman, edge, funeral, installment, batter	paraphrase, toxic, sample, intellectual, physical, distinct, listen, reward, meat, corporate
20	legislature, february, sworn, message, bully, sage, bullying, governor, oath, issue	bullying, bully, uncommon, sage, sworn, prerogative, commemorate, darkness, scan, label	bully, bullying, sage, uncommon, sworn, label, prerogative, commemorate, scan, darkness	dignity, seek, prize, cup, strongly, curve, livelihood, dirt, talented, dedicate
21	reallocation, manageable, service, pension, citizen, support, wholesale, resource, growth, system	reallocation, manageable, wholesale, roadblock, bravely, scrutinize, turnover, counterpart, solvency, rarely	reallocation, manageable, wholesale, roadblock, bravely, scrutinize, turnover, counterpart, rarely, solvency	complexity, migrant, extract, conscience, incarcerate, uniquely, remarkably, commence, principally, rapidly

(continued)

Topic	Phi	Lift	Relevance	FREX
22	governor, journal, joint, commonwealth, responsibility, quality, history, citizen, deliver, life	borne, digitize, disappointment, reinforce, motto, humility, considerably, separation, theme, crumble	borne, digitize, reinforce, disappointment, journal, commonwealth, theme, wednesday, motto, considerably	panel, selection, land, excitement, structure, entertainment, prudence, film, judiciary, poet
23	deficit, fiscal, spending, revenue, face, cut, balance, problem, billion, recession	wiring, pet, grid, steven, deficit, unpopular, structural, jessica, trillion, ink	deficit, wiring, pet, grid, debt, recession, structural, rating, spending, fiscal	defender, disappear, examine, partnership, tri, diligent, deduct, compete, impression, competitive
24	dairy, telecommunication, milk, warm, sewer, revolve, support, cost, cheese, southwestern	warm, cheese, dairy, southwestern, smooth, milk, rewrite, telecommunication, vain, strategically	warm, dairy, cheese, milk, telecommunication, southwestern, smooth, rewrite, revolve, sewer	billy, rain, arrive, diploma, inception, fitting, execution, shop, disappointment, illustrate
25	teacher, educational, salary, support, quality, secondary, public, elementary, professional, system	col, resistance, certification, doc, impasse, enrichment, characteristic, profession, elevate, secondary	col, teacher, certification, secondary, elementary, resistance, salary, profession, enrollment, educational	rehabilitate, discourage, force, poorly, janet, everywhere, faster, endorse, past, era
26	get, let, want, look, back, money, lot, help, job, start	laugh, grandmother, hey, rally, amaze, winner, eve, folk, dress, joke	get, folk, laugh, amaze, grandmother, lot, hey, frankly, rally, eve	warrant, teacher, touch, randy, fare, endless, home, recreation, museum, symptom
27	housing, system, problem, area, unit, legislation, income, legislature, public, development	patron, worsen, housing, moderate, price, archaic, uniformity, rental, mortgage, imaginative	housing, patron, worsen, moderate, mortgage, rental, machinery, archaic, rent, price	diverse, homework, player, standard, pocketbook, abandon, displace, grim, doorkeeper, hem
28	worker, wire, sacrifice, fiber, family, specialist, afternoon, abraham, bonus, military	chaplain, wire, optic, airman, minister, abraham, unleash, fiber, bear, spouse	wire, chaplain, optic, airman, abraham, fiber, minister, unleash, bear, spouse	endowment, temper, healthy, supplemental, matt, mon, fault, forefather, initiate, job
29	guard, serve, military, men, iraq, woman, family, son, governor, army	hank, pastor, jennifer, tommy, commander, ian, guardsman, iraq, troop, helicopter	guard, iraq, pastor, hank, troop, jennifer, army, commander, tommy, ian	voucher, exploration, justification, absorb, rig, evaluate, hopefully, indispensable, hygiene, forever
30	wage, minimum, beverage, cal, crew, labor, public, respectful, pub, brick	beverage, respectful, allotment, rededicate, brick, pub, cal, mortar, crew, unwilling	beverage, cal, respectful, crew, allotment, wage, brick, pub, rededicate, mortar	spur, diesel, doorkeeper, temptation, steadily, dilemma, issue, grateful, silence, december
31	haul, sec, dog, waste, autism, dump, foolish, carve, explosion, civilian	sec, haul, dog, recapture, output, autism, foolish, explosion, carve, manifest	sec, haul, dog, autism, foolish, recapture, output, explosion, carve, aspire	diagnostic, deserve, deliberate, deplete, default, discus, delegation, rehabilitate, deterrent, delinquent
32	quo, status, marriage, thee, pat, championship, child, let, wine, fashion	thee, championship, quo, marriage, roster, fame, pat, crusade, wine, status	thee, quo, championship, marriage, status, pat, roster, wine, fame, crusade	disappointed, secretary, consumer, outmode, recently, nationally, plea, tragically, remainder, push

(continued)

Topic	Phi	Lift	Relevance	FREX
33	growth, economic, sion, revenue, grow, till, difficult, recession, past, tho	trailer, till, accountant, sion, rethink, hinder, tho, undertook, tin, analyst	till, sion, trailer, tho, accountant, rethink, tin, hinder, frame, undertook	rethink, surpass, employment, emotion, excessive, agent, alternative, newspaper, phrase, exceed
34	family, help, care, cost, child, income, affordable, credit, job, homeless	reinvest, homelessness, grab, refocus, donor, homeless, centerpiece, slide, discount, crunch	homelessness, homeless, reinvest, refocus, grab, donor, discount, slide, centerpiece, affordable	fight, train, link, mam, toxic, frail, cell, coach, hotel, mother
35	time, thy, anxiety, adversity, thou, face, fear, family, job, blessing	thy, anxiety, thou, kitchen, adversity, blessing, daunt, bought, mighty, counter	thy, anxiety, thou, adversity, blessing, kitchen, daunt, bought, mighty, counter	inspector, ravage, intensive, treasure, donation, enrol, downstate, fruit, camp, plow
36	ing, stance, cant, surprisingly, mat, mil, value, lion, divorce, problem	surprisingly, stance, cant, selfish, divorce, mat, drill, noise, instant, injustice	surprisingly, stance, cant, mat, divorce, selfish, mil, drill, injustice, lion	radio, irrigation, evident, samuel, ever, diverse, dirt, estimate, digest, supplier
37	local, subsidy, transmission, comfortable, renewable, bipartisanship, intolerable, abandonment, percent, cost	bipartisanship, intolerable, comfortable, freed, abandonment, transmission, ugly, proclamation, reinvestment, insulate	bipartisanship, comfortable, intolerable, transmission, subsidy, freed, abandonment, reinvestment, ugly, freight	pan, star, directly, dignity, governmental, democratic, asse, precisely, reduction, learner
38	partnership, tain, public, resource, community, innovation, optimist, forum, stakeholder, ingenuity	tain, optimist, superb, undeniable, energize, spawn, pipe, stakeholder, uniquely, pitch	tain, optimist, superb, undeniable, energize, stakeholder, pipe, uniquely, spawn, ingenuity	ideology, injury, geothermal, escalate, excess, fierce, derive, extreme, prerogative, escort
39	consumer, problem, area, public, legislation, can, proposal, local, citizen, legislature	can, cum, candidly, logically, repay, lieu, bottle, lien, decree, impair	can, cum, repay, candidly, lieu, consumer, bottle, logically, decree, lien	decree, perspective, landscape, pristine, designate, employ, mind, enjoyment, namely, reluctant
40	enrol, ary, correctly, contingent, bachelor, playground, food, married, spill, partly	correctly, ary, playground, bachelor, contingent, partly, married, spill, enrol, bidding	ary, correctly, playground, bachelor, contingent, partly, married, enrol, spill, bidding	producer, feel, fairly, period, exempt, finger, undertaking, existence, plenty, fight
41	brain, want, text, prison, michael, maybe, get, unbelievable, michelle, rein	unbelievable, eric, michelle, fantastic, text, brain, rein, dinner, appliance, surroundings	unbelievable, brain, eric, michelle, text, fantastic, rein, dinner, endow, appliance	leadership, transform, exit, flood, consult, fatal, prohibition, register, observe, monthly
42	development, economic, industrial, area, trade, expand, job, product, agricultural, labor	promotional, ism, diversification, grower, industrial, mold, poultry, ployment, incidentally, mug	industrial, trade, promotional, diversification, development, agricultural, marketing, promotion, ism, economic	unpleasant, namely, predictable, county, dimension, incentive, dedicate, correctly, vastly, participate
43	student, passionately, underestimate, reinvigorate, challenge, community, improve, public, panel, life	passionately, reinvigorate, underestimate, unprepared, academically, internship, socially, thankfully, hearten, unveiled	passionately, underestimate, reinvigorate, unprepared, internship, academically, socially, thankfully, farther, hearten	eminent, diversify, locality, differently, gram, kept, reality, destruction, reckless, architectural

(continued)

Topic	Phi	Lift	Relevance	FREX
44	typically, tim, redesign, beginning, trail, roger, pit, child, wildfire, deaf	typically, bike, beginning, roger, pit, wildfire, redesign, tim, trail, pillar	typically, beginning, redesign, roger, bike, pit, wildfire, tim, trail, deaf	lastly, distribute, threat, regularly, elector, christian, double, diagnostic, overlook, violent
45	problem, resource, revenue, ant, agriculture, directive, account, wit, rancher, ongoing	ant, directive, packer, cultivate, wit, inch, twentieth, sane, conservancy, rancher	ant, directive, wit, packer, cultivate, inch, rancher, specie, twentieth, sane	joe, image, diagnosis, fringe, irrigation, jewel, operator, roll, promote, enemy
46	job, create, economic, help, nation, growth, rate, today, unemployment, grow	powerhouse, invention, spin, backwards, portfolio, steel, supplier, unmatched, creator, reclaim	job, unemployment, create, steel, low, powerhouse, creator, ranked, supplier, manufacturer	tha, environment, satisfaction, satisfy, notable, colonel, tier, asian, clearinghouse, unleash
47	welfare, recipient, assistance, system, area, roll, problem, benefit, support, responsibility	identical, fur, drainage, unrealistic, welfare, recipient, nut, multitude, roll, accuracy	welfare, recipient, fur, identical, roll, drainage, unrealistic, nut, multitude, needy	satisfactory, danger, theater, magnificent, honest, issue, prove, deterioration, reorganize, simpler
48	job, development, company, economic, create, technology, help, investment, growth, grow	automotive, entrepreneurial, breakthrough, entrepreneur, spur, weld, geothermal, incubator, uncle, precision	company, entrepreneur, technology, automotive, industry, entrepreneurial, venture, job, innovation, telecommunication	multitude, regain, deliberately, discrimination, uniform, december, ineffective, lose, deficit, death
49	transportation, project, development, community, system, economic, growth, infrastructure, resource, area	bolster, ramp, congestion, showcase, commuter, replenish, derive, surroundings, capped, waterway	bolster, transportation, congestion, commuter, ramp, infrastructure, boom, mid, derive, showcase	pill, boast, maturity, push, deficient, deal, governmental, independent, contributor, planner
50	teen, care, health, community, immense, pregnancy, abortion, greenhouse, economic, challenge	immense, lately, unlock, abortion, collar, predictable, erase, greenhouse, cherish, pick	immense, lately, abortion, collar, predictable, greenhouse, unlock, teen, cherish, pregnancy	intend, father, culture, matt, unscrupulous, substantial, meaning, progressive, gamble, understandable
51	discrimination, employment, lam, renaissance, accommodation, old, lit, donate, religion, surcharge	lam, accommodation, renaissance, lit, discrimination, hem, donate, allied, religion, metal	lam, renaissance, accommodation, discrimination, lit, donate, hem, religion, metal, allied	inequitable, salute, lisa, fairness, deed, rig, uniquely, formation, adequacy, dramatic
52	development, fund, economic, ring, credibility, heretofore, borough, per, legislature, indebtedness	borough, credibility, heretofore, maturity, ring, ting, indebtedness, furniture, refinance, acceleration	borough, credibility, ring, heretofore, indebtedness, ting, maturity, furniture, acceleration, refinance	ire, dental, mistake, gang, parochial, misery, score, distinct, residence, resident
53	youth, appalachian, juvenile, crime, community, support, delinquency, ged, vocational, public	appalachian, ged, beating, delinquency, outrage, billy, moratorium, protective, steal, seamless	appalachian, ged, delinquency, beating, youth, moratorium, outrage, billy, protective, juvenile	cornerstone, retailer, receipt, inefficient, rush, dam, transitional, constantly, context, cow

(continued)

Topic	Phi	Lift	Relevance	FREX
54	get, want, town, problem, city, community, local, real, issue, slogan	elude, loudly, slogan, backward, trash, bite, knock, hum, garage, drove	elude, loudly, slogan, trash, hum, knock, argue, backward, nobody, bite	container, concurrent, modern, trauma, teeth, noon, demonstration, culture, activate, individually
55	ate, embarrass, sixty, owe, hotel, crash, sovereign, slum, operator, dimension	embarrass, slum, valley, ate, sovereign, hotel, crash, ski, fool, independent	embarrass, ate, slum, hotel, valley, sovereign, crash, sixty, ski, dimension	heretofore, diversify, poise, disclose, structural, device, statue, diversity, folk, hope
56	fund, revenue, percent, current, reserve, share, total, unpleasant, appropriation, growth	unpleasant, statesmanship, caution, reallocation, frugality, consequently, aggregate, reserve, commonly, heretofore	unpleasant, statesmanship, caution, reserve, consequently, reallocation, predecessor, frugality, heretofore, aggregate	federally, economical, correctly, per, health, consume, menu, dedicate, constantly, conscience
57	mum, welfare, family, neighborhood, help, job, crime, child, want, get	beet, mum, jam, menu, lea, saint, male, distressed, inner, victimize	mum, jam, beet, menu, lea, saint, inner, male, distressed, neighborhood	skip, today, relieve, appendix, grip, drought, wife, notion, toxic, courthouse
58	district, student, kindergarten, grade, assessment, dental, aid, health, confident, percent	twelfth, remediation, collaborate, dentist, slot, dental, hallmark, adequacy, outmode, unchanged	twelfth, dental, remediation, slot, dentist, collaborate, district, kindergarten, hallmark, collaboration	expression, stabilize, fantastic, partially, keen, failure, expense, foreseeable, suburban, exceed
59	development, commission, visitor, site, regional, facility, area, public, historical, expansion	impetus, impede, vacation, accrue, parcel, mainly, preclude, historical, attraction, focal	historical, impetus, visitor, vacation, attraction, inter, impede, accrue, parcel, airport	shorten, match, fun, fairly, flee, exclude, divert, initiative, enlist, operational
60	license, driver, traffic, highway, patrol, enforcement, telephone, death, drive, system	turbine, whip, drunken, speed, confirmation, license, patrolman, telephone, violate, driver	license, driver, turbine, telephone, whip, speed, drunken, traffic, patrol, violate	love, deed, mount, decent, relation, decay, cynic, mental, recruitment, decade
61	january, tuesday, jon, predator, percent, president, today, generosity, indeed, december	jon, tuesday, verify, generosity, predator, skip, january, delete, prey, subsistence	tuesday, jon, january, predator, generosity, verify, skip, delete, prey, subsistence	plug, grief, harry, guideline, suspect, summarize, stuff, evident, external, residency
62	child, community, student, educational, kindergarten, teacher, life, help, childhood, system	unveil, learner, out, youngster, preschool, continuum, preventable, readiness, dropout, spearhead	unveil, learner, youngster, preschool, out, kindergarten, dropout, readiness, childhood, continuum	life, cocaine, corp, gift, room, damage, relevant, son, entrepreneurial, housing
63	resource, conservation, management, land, commission, supply, legislation, flood, area, development	dependable, enjoyment, scatter, conversion, land, quantity, conservation, watershed, fishery, forest	land, conservation, management, supply, flood, dependable, conversion, enjoyment, quantity, forest	join, predator, pool, manifest, minimum, status, innovator, ratio, homemaker, fulfill
64	code, estate, tribal, real, ill, native, revenue, relation, system, cation	hasten, terribly, ace, tribal, estate, undo, processor, serviceman, code, split	code, estate, hasten, tribal, terribly, ace, undo, processor, cation, serviceman	export, land, realize, pound, enormous, disruption, effective, emission, discriminatory, generate
65	veteran, refugee, complain, judy, fund, herein, herewith, senior, service, confrontation	refugee, herewith, judy, complain, herein, confrontation, easement, torn, veteran, tent	veteran, refugee, complain, judy, herewith, herein, confrontation, easement, torn, clearinghouse	crop, impede, enemy, contribution, struggle, cutback, eighty, regularly, drought, lastly

(continued)

Topic	Phi	Lift	Relevance	FREX
66	loan, banking, advertising, system, highway, legislature, credit, legislation, tourist, administration	publicity, banking, charter, regulator, advertising, exhibit, eradicate, periodic, architect, ail	banking, loan, publicity, advertising, charter, exhibit, regulator, periodic, architect, eradicate	favorite, reminder, penal, piece, drag, enterprise, expire, rim, serf, dioxide
67	appropriation, building, fund, construction, facility, capital, total, project, area, department	heretofore, dormitory, apparently, din, declaration, architectural, outlay, fro, thereof, accreditation	appropriation, heretofore, outlay, apparently, dormitory, declaration, gram, building, din, crease	roof, dedicate, low, salute, dust, prohibit, resistance, despair, attorney, gubernatorial
68	ion, inn, gen, nee, noon, gas, incredibly, ron, earthquake, lisa	gas, gen, nee, earthquake, lisa, cup, inn, ion, ron, trace	gen, ion, gas, nee, inn, earthquake, lisa, noon, ron, cup	transparent, dialogue, downward, utility, simplify, decisive, homemaker, employee, mission, premier
69	governor, defend, domestic, flag, thank, senator, please, president, police, allegiance	barbara, fighter, humble, corporal, nam, stranger, devotion, flag, allegiance, mourn	barbara, flag, fighter, humble, allegiance, defend, corporal, devotion, stranger, domestic	pain, unanimously, surgery, goodness, edge, evolve, knee, immunize, ombudsman, widely
70	strip, super, damage, jesse, finalize, fraud, highway, aftermath, storm, bigotry	jesse, strip, aftermath, bigotry, finalize, prestigious, super, spotlight, federalism, morally	strip, jesse, finalize, aftermath, super, bigotry, prestigious, federalism, spotlight, morally	distance, dozen, patron, instrument, disabled, deployment, tuesday, devastate, rear, distant
71	television, proclaim, destination, station, donation, beach, leo, public, child, scene	proclaim, leo, buddy, donation, television, pristine, decay, displayed, destination, scene	proclaim, television, leo, donation, buddy, destination, pristine, beach, decay, scene	conscious, contaminate, trim, shin, welcome, creatively, consensus, sell, favorable, emphasize
72	cancer, public, attorney, criminal, fin, enforcement, vaccine, bullock, youthful, agent	vaccine, fin, bullock, youthful, confinement, corrupt, inequality, wheel, cancer, diagnose	vaccine, fin, bullock, cancer, youthful, confinement, corrupt, inequality, wheel, agent	glow, courthouse, recruit, math, editor, secret, understandable, correctly, courageously, shoulder
73	prison, system, inmate, criminal, sentence, community, correctional, offender, facility, bed	imprisonment, incarceration, incarcerate, felon, inmate, overcrowd, recidivism, prisoner, prison, correctional	prison, inmate, incarceration, prisoner, sentence, correctional, bed, overcrowd, jail, imprisonment	excellence, dinner, lion, structurally, knowledgeable, combine, comment, michelle, collect, creek
74	department, service, agency, public, commission, problem, area, development, system, administration	oat, tuberculosis, liaison, parallel, leo, outmode, billboard, overlap, centralize, seasonal	department, oat, agency, tuberculosis, commissioner, liaison, reorganize, function, consolidation, centralize	hunger, economic, decency, restrain, dramatically, poor, defy, instill, structurally, unfair
75	student, teacher, technology, help, rep, learn, child, initiative, computer, fund	rep, virtual, breakfast, scholar, sen, interactive, remediation, tutor, competency, proficiency	rep, sen, technology, virtual, scholar, breakfast, learn, computer, student, teacher	defy, prescribe, embody, appointee, association, overseas, new, dime, annually, fare
76	utility, solar, heating, consumer, rate, conserve, tenant, home, cost, impression	solar, rejection, impression, heating, utility, landlord, tenant, bare, thwart, refusal	solar, utility, heating, impression, rejection, tenant, landlord, bare, thwart, conserve	fourteen, fundamental, fund, frugal, sideration, fiber, thoughtful, guardsman, formerly, reve

(continued)

Topic	Phi	Lift	Relevance	FREX
77	drug, abuse, treatment, substance, alcohol, dealer, community, fund, prevention, effort	token, marijuana, scourge, dealer, drug, ladder, substance, pusher, illicit, abuse	drug, token, marijuana, dealer, abuse, scourge, substance, ladder, treatment, alcohol	mandate, home, investigative, regulatory, disability, ongoing, stop, seek, execute, solid
78	child, birth, adult, litigation, baby, mother, round, development, care, childhood	round, female, outdoors, litigation, birth, destructive, conducive, mobility, consortium, sufficiently	birth, round, litigation, female, outdoors, destructive, baby, conducive, mobility, prenatal	intervene, dress, tool, expect, stroke, heart, peace, retailer, orderly, forefather
79	today, want, governor, get, challenge, thank, look, back, life, legislature	appreciative, humor, wiser, passion, jimmy, arrival, tyranny, gavel, partisanship, mercury	appreciative, humor, passion, arrival, wiser, yes, partisanship, jimmy, politics, understood	unusual, quantity, disproportionate, largely, huge, discovery, outmode, dentist, mindful, dilemma
80	fiscal, fund, revenue, expenditure, appropriation, public, estimate, total, current, educational	chemistry, downstate, earnestly, contingency, unavoidable, multiply, approximate, deliberately, anticipate, commendable	fiscal, expenditure, estimate, anticipate, downstate, chemistry, surplus, appropriation, wednesday, college	deliver, deliberation, dig, presentation, inequitable, supporter, rapidly, desperately, dedication, rehabilitative
81	spending, governor, money, fiscal, service, excellency, get, fund, want, spend	foresee, uncle, dip, binding, apologize, excellency, unsustainable, raid, bravery, bullet	foresee, excellency, uncle, binding, dip, unsustainable, raid, apologize, trim, bullet	conservation, fail, fortunate, commemorate, justification, core, deployment, thats, experienced, planner
82	local, government, aid, district, county, municipality, problem, circuit, municipal, city	stem, circuit, intergovernmental, breaker, municipality, inevitable, thoughtfully, government, revamp, exclusive	local, government, circuit, municipality, stem, intergovernmental, breaker, inevitable, municipal, levy	light, pinto, heat, shirk, revoke, surgery, freedom, entertainment, grateful, naturally
83	remedy, kill, insurer, duo, plead, upcoming, catastrophe, quick, rare, ultimately	remedy, plead, insurer, catastrophe, duo, upcoming, kill, rare, achievable, systematic	remedy, insurer, plead, duo, catastrophe, upcoming, kill, rare, quick, achievable	inevitable, resort, inspiration, abandon, reluctance, instant, integrate, innovator, japan, imaginative
84	library, strange, garbage, grip, reallocate, hospitality, dwell, dominate, public, sophisticated	strange, hospitality, garbage, library, dwell, hick, reallocate, grip, sophisticated, dominate	strange, library, garbage, hospitality, dwell, reallocate, grip, hick, sophisticated, dominate	shed, tiny, fail, injury, energetic, thereafter, underway, interactive, geographical, fierce
85	public, voter, legislature, reapportionment, legislation, political, proposal, commission, campaign, group	electoral, reapportionment, dominant, apportionment, assert, representation, reapportion, dismiss, intervene, scandal	reapportionment, electoral, representation, apportionment, assert, dominant, voter, reapportion, ethic, intervene	farther, disability, discretionary, minimize, refund, deployment, diligence, extraordinary, exploit, reach
86	child, family, care, parent, health, help, service, support, life, abuse	stroke, caseworker, setting, sexual, immunization, detect, killer, child, discriminate, parent	child, stroke, parent, family, caseworker, abuse, care, sexual, neglect, healthy	political, wind, volume, hurdle, rape, institutionalize, visitor, trim, super, literacy

(continued)

Topic	Phi	Lift	Relevance	FREX
87	pipeline, collaboration, sir, want, create, township, thank, barrel, kathy, leader	kathy, generator, sea, collaboration, sir, offshore, acreage, platform, dust, township	collaboration, kathy, generator, pipeline, sir, sea, township, platform, drilling, dust	freely, shelf, effectiveness, thoughtful, historically, economist, conversion, element, danger, robust
88	employee, retirement, service, system, personnel, public, salary, benefit, retire, civil	survivor, attrition, opt, multiply, vain, turnover, corruption, supervisor, retirement, merger	employee, retirement, survivor, attrition, corruption, opt, multiply, personnel, retire, vain	lord, violent, role, spar, indispensable, registry, turn, faculty, efficient, enemy
89	health, care, service, department, infant, cost, rate, mortality, percent, nursing	immunize, mortality, ambulance, immunization, detection, infant, prenatal, pregnant, aide, lodge	mortality, immunization, immunize, infant, ambulance, detection, prenatal, pregnant, nursing, screen	ranked, fought, exhibit, heart, rain, emission, immigration, meaningful, prohibition, doorkeeper
90	dam, sewage, foot, stuff, hut, stream, negotiation, project, journal, legislature	stuff, mate, hut, wholesome, sewage, dam, fraction, tin, withdrawal, mismanagement	stuff, sewage, hut, mate, dam, wholesome, foot, tin, negotiation, fraction	fraction, inefficiency, sight, despite, diligently, discretion, leaf, dependence, derive, disastrous
91	student, teacher, child, system, classroom, parent, public, grade, teach, standard	homework, graduation, classroom, math, grade, educator, lifelong, textbook, teach, score	student, teacher, classroom, grade, teach, math, graduation, educator, learn, parent	excitement, personality, mung, recapture, karen, poor, fragment, frankly, crew, source
92	get, really, everybody, want, child, life, let, problem, back, help	pamphlet, tum, clothes, yeah, slam, thrill, tent, reinvent, quit, everybody	pamphlet, tum, everybody, clothes, quit, yeah, reinvent, slam, thrill, tent	turnover, implication, remedial, ensure, rein, toughen, engine, dock, lapse, greatness
93	populate, event, per, clerical, installation, beer, cent, procurement, cam, favorably	populate, clerical, cam, beer, procurement, installation, empowerment, favorably, unduly, illustrate	populate, clerical, beer, cam, procurement, installation, favorably, empowerment, unduly, event	practitioner, ons, fraudulent, grab, examiner, eon, gram, gift, drag, honor
94	service, care, health, elderly, citizen, age, cost, senior, assistance, disabled	hint, gin, developmentally, transitional, medically, age, frail, institutionalization, isolation, disabled	hint, age, gin, elderly, disabled, needy, developmentally, medically, senior, nursing	energize, everybody, injustice, fully, partly, delinquency, prosperous, lord, encourage, enrich
95	compensation, worker, injured, workman, cow, adjutant, labor, fund, memorial, unemployment	cow, soundness, spare, adjutant, armory, activate, workman, injured, slight, compensation	cow, compensation, workman, injured, adjutant, spare, soundness, armory, activate, slight	foresight, preparedness, perpetuate, fragment, hammer, resilient, secondary, forfeiture, prize, petty
96	health, care, cost, insurance, system, coverage, access, affordable, child, improve	chil, preventative, obesity, uninsured, diabetes, wellness, tobacco, preventable, coverage, provider	chil, health, coverage, uninsured, care, tobacco, provider, insurance, wellness, obesity	lot, premise, signify, county, dilemma, quick, supplier, earthquake, area, indigent
97	judge, judicial, barrel, menu, silent, appellate, nineteen, circuit, arbitration, bench	menu, barrel, appellate, silent, nineteen, bench, arbitration, mediation, livable, version	menu, barrel, appellate, silent, nineteen, bench, judicial, judge, arbitration, livable	grip, secondary, plight, idly, hamper, hike, rancher, heating, historical, incubator

(continued)

Topic	Phi	Lift	Relevance	FREX
98	governor, citizen, public, group, judiciary, legislature, incumbent, gentleman, official, problem	incumbent, tension, entrance, reiterate, concurrent, display, unanimously, editor, willie, posse	incumbent, concurrent, judiciary, entrance, tension, reiterate, display, unanimously, chosen, deem	stifle, supportive, impair, stem, crash, haul, proudly, stayed, constituency, praise
99	senator, joint, governor, president, escort, honorable, message, legislature, january, rostrum	rostrum, reconvene, escort, ralph, fred, duly, fir, splendid, honorable, senator	senator, escort, rostrum, joint, honorable, reconvene, president, fred, duly, ralph	refinance, passionate, tighten, ensure, penal, environment, feel, era, segment, ewe
100	fund, revenue, biennium, appropriation, spending, service, percent, total, fee, expenditure	reallocate, mon, wont, plea, sine, lump, mere, unappropriated, recur, similarly	reallocate, mon, wont, plea, biennium, mere, sine, lump, recur, biennial	drought, inception, greatness, guest, envision, mankind, novel, midwestern, pitch, expansion
101	lottery, compact, cigarette, nuclear, pack, businessman, intangible, tip, meat, immunity	tip, intangible, immunity, businessman, cigarette, inadequacy, pack, charlie, compact, lottery	cigarette, compact, tip, intangible, lottery, pack, businessman, nuclear, immunity, inadequacy	jewel, railroad, lightly, exam, plentiful, gentle, injustice, positive, multiply, split
102	ing, fore, humanity, commonwealth, community, system, educational, public, ill, quality	attest, fore, theater, humanity, seamless, nomic, uranium, mold, wheel, delinquent	fore, humanity, attest, seamless, ing, nomic, theater, mold, uranium, wheel	depressed, persevere, exhibit, thought, digital, decree, mortar, gear, anyway, news
103	defense, department, ton, service, til, auditor, emergency, problem, jurisdiction, commission	til, eventual, booklet, impend, reshape, archive, sanitation, precedent, cur, sorely	til, reshape, eventual, booklet, impend, archive, cur, ton, auditor, precedent	married, mate, excellency, soil, lieu, drill, healthy, week, escape, pile
104	percent, reveal, measurement, wear, aid, commitment, economic, belt, sharpen, current	pill, measurement, sharpen, extract, passport, bind, likelihood, reveal, supervisory, pile	measurement, pill, sharpen, extract, reveal, passport, percent, bind, likelihood, pile	profoundly, plastic, factory, sad, eye, essence, genius, horizon, badly, auto
105	tourism, film, economic, development, sin, help, ing, private, lou, job	lou, sin, film, fro, idly, outreach, maturity, planner, ting, breakfast	lou, sin, film, fro, outreach, idly, coalition, tourism, ting, lender	uniquely, civic, upcoming, retain, union, compact, jeopardize, economic, commend, forecast
106	fund, investment, account, billion, build, infrastructure, project, money, support, invest	rev, defray, pothole, continuously, anne, lust, routine, fruit, unpaid, overwhelmingly	rev, account, defray, fund, stimulus, pothole, anne, continuously, earmark, routine	pharmacy, refocus, per, extent, feasible, professional, bipartisanship, boast, recreation, gamma
107	god, bless, let, courage, thank, america, band, life, forward, look	band, shed, passionate, bag, shin, bless, appendix, god, territory, statehouse	bless, god, band, shed, shin, passionate, bag, journey, territory, seize	harm, week, wash, eliminate, meat, everything, gallon, function, relief, increasingly
108	fund, per, revenue, cent, cost, available, service, local, sale, estimate	lesser, partially, quasi, assessor, earmark, increment, equalization, offset, successive, principally	lesser, per, partially, earmark, cent, quasi, equalization, revenue, levy, offset	possibly, participation, logic, squeeze, quality, great, isaac, ethic, police, critic

(continued)

Topic	Phi	Lift	Relevance	FREX
109	boat, reservoir, sport, waterway, area, child, commission, lunch, registry, terminal	boat, waterway, neal, registry, attainable, stall, reservoir, terminal, lunch, curve	boat, waterway, reservoir, registry, neal, sport, stall, terminal, attainable, lunch	dress, emphasize, inspiration, salute, dismiss, forest, exploitation, drastically, distressed, align
110	research, technology, scientist, development, economic, fund, investment, lab, engineering, foundation	scientist, researcher, physic, discovery, magic, renowned, research, designation, scientific, professor	research, scientist, researcher, physic, scientific, magic, lab, science, discovery, renowned	unit, editor, japan, eradicate, overcome, waive, prolong, monopoly, debt, depart
111	get, want, help, child, life, prescription, coach, senior, today, team	coach, suicide, breast, glass, football, christian, angry, barber, syndrome, player	coach, football, suicide, breast, prescription, glass, player, christian, angry, syndrome	energy, partisanship, paradox, resilience, overlook, vigilance, organizational, engine, unit, fabric
112	anti, thankful, religious, teen, smoking, racial, charity, prejudice, signal, harsh	thankful, prejudice, signal, religious, charity, racial, interview, anti, teen, smoking	thankful, religious, prejudice, charity, racial, anti, signal, teen, smoking, interview	sovereignty, widespread, ing, oath, gas, formally, neighbor, gap, frustrate, greener
113	benefit, employee, compensation, unemployment, employment, salary, worker, cost, ment, living	duration, weekly, ward, lation, fringe, successive, depart, jobless, bread, workman	weekly, lation, duration, compensation, ward, fringe, depart, benefit, unemployment, workman	fuel, wit, week, restoration, incorporate, slide, son, sound, incarcerate, suffer
114	job, training, worker, skill, help, employer, technical, train, career, skilled	jewel, apprenticeship, unfilled, globally, retrain, skilled, training, exit, skill, dual	training, apprenticeship, skill, jewel, skilled, worker, technical, employer, job, train	constructive, ill, institutional, wonder, disruptive, repeatedly, unpleasant, controversial, acid, drive
115	rainy, swim, voucher, invent, squander, help, slip, finger, randy, family	randy, voucher, swim, finger, squander, soft, invent, tiny, din, complacent	swim, voucher, randy, finger, squander, rainy, invent, soft, tiny, livelihood	dream, document, impede, hub, hey, residence, bigotry, creatively, travel, tour
116	resource, quality, area, environment, preserve, tree, recreation, development, develop, environmental	power, tank, tree, underground, outdoor, leak, reclamation, external, mine, erosion	tree, power, outdoor, tank, underground, wildlife, mine, reclamation, recreation, recreational	foreign, shift, signal, pharmaceutical, yeah, jon, hearing, foremost, unlike, irrigation
117	vehicle, motor, highway, legislation, commission, public, legislature, system, department, statute	landowner, mileage, domain, elude, lapse, compulsory, motor, vehicle, therein, mansion	vehicle, motor, landowner, accident, domain, mileage, issuance, compulsory, inspection, registration	mental, hotel, trailer, fulfillment, new, leader, enhance, living, favorite, disappointment
118	drive, possess, alcohol, drunk, test, drinking, blood, fatality, gamble, successive	possess, successive, megawatt, fatal, bound, refuge, bail, fatality, drunk, blood	possess, successive, megawatt, drunk, blood, fatality, bail, fatal, alcohol, drinking	neighbor, ell, endeavor, environmental, reliant, expectation, trooper, supplement, discriminate, upgrade
119	income, revenue, rate, proposal, legislature, governor, inflation, message, believe, relief	muster, theory, hereby, bracket, rebate, index, regressive, simpler, adjust, excise	theory, muster, income, hereby, rebate, bracket, adjust, inflation, regressive, index	elementary, reach, meant, overseas, enroll, donate, mixed, sharply, everywhere, duo

(continued)

Topic	Phi	Lift	Relevance	FREX
120	health, care, hospital, community, service, facility, cost, patient, treatment, mental	residency, clinic, physician, acute, hospitalization, hospital, clinical, nurse, afflict, doctor	hospital, clinic, physician, health, residency, doctor, patient, care, nurse, acute	earnestly, illiteracy, tear, district, document, gather, pocket, deploy, persevere, related

Table B.6: Topic Labels Assigned to Sample of Cleaned Document Excerpts

Text	Topic Assn. 1	Topic 1 Prop.	Topic Assn. 2	Topic 2 Prop	Topic Assn. 3	Topic 3 Prop
resource laboratory benefit press urge building public boat launch site camp ground area facility citizen emphasis locate facility convenient city dock dock lose money bail turn efficient money operation throughout administration earn obligation earn handsome net profit dock management testament ability dock consistent operational profit dock drawn trade dock revenue million operational expense total million outstanding bond million million aside depreciation left sizeable net profit almost million tax money involve profit retain dock build improve facility network dock dock get start dock facility dock near undoubtedly installation aid materially industrial growth dock shipping good create dock attract shipping barge tow foreign generate revenue dock system highly competitive port overlook possibility secure tonnage promote develop trade dock participate successful trade mission area america encourage domestic shipment via sale permanently station louis base canvas lead city engage hit campaign insure role growth dock people believe reach strongly urge dock kept governor might unresponsive public public dock deficit fund obligate treasury cramp dock governor powerless remedial facility construct handle dock competitive position port stimulate shipping trade tour trade area trade possibility latin nation waterway development construction mid waterway cut across navigable upstream modernization almost improve navigation campaign build waterway gain momentum congressional fund	42 (development,economic,industrial,area,trade)	26.9%	78	22.6%	67	11.1%

Table B.6: Topic Labels Assigned to Sample of Cleaned Document Excerpts *(continued)*

Text	Topic Assn. 1	Topic 1 Prop.	Topic Assn. 2	Topic 2 Prop	Topic Assn. 3	Topic 3 Prop
safe value neighborhood human stranger vacant abandon house haunt reminder preoccupation economic development fail relationship share conservation community establish community conservation companion economic development administration advocate conserve enhance treasure ingenuity generation money material resource carefully build ing get yes development area initiative help target service resource encourage support private reinvestment distressed community focus activity job community people designate neighborhood city development area money available community grant total area believe appropriation million development area administer department community affair couple million appropriation assist worker develop specific recovery strategy help care quality life community commitment community technical definition distressed today resident human definition unemployed worker distressed family face potential loss residential fault mortgage foreclosure credit respond emergency mortgage assistance misgiving financial base base rickety sent million fund million annually corporate contribution neighborhood assistance tax credit conclusion inescapable mortgage assistance possibly match passage withdrawal tax credit neighborhood assistance clearly look elsewhere logical turn revenue source continued grow recession lottery borrow million lottery specifically mortgage foreclosure fourth initially borrowing hoax fail sort phantom dime borrow lottery fund homeowner loan return specter sheriff door reality pocket par press community revitalization job expansion goal pursue citizen environment breathe drink enjoy mountain forest perishable treasure touch life creator	34	13.2%	116	11.2%	57	11.0%
visit join conference innovation issue visit issue individual community nurture innovation nation economic showcase help get idea laboratory ara job ing product service market enjoy economic choke congestion skyrocket cost housing average life skill knowledge citizen left school challenge guarantee past challenge invite citizen join chart guide commonwealth throughout history understood meaning thanksgiving forge revolution nation classroom public help establish fine college university seed bed revolution tech revolution vision nation chi labor create hospital ill legacy people era past challenge era child let building tradition people lead thank	48	36.4%	9	35.4%	102	9.7%

Table B.6: Topic Labels Assigned to Sample of Cleaned Document Excerpts *(continued)*

Text	Topic Assn. 1	Topic 1 Prop.	Topic Assn. 2	Topic 2 Prop	Topic Assn. 3	Topic 3 Prop
development commissioner insurance recently resist implement resistance predicate lack concern availability premium cost problem encounter citizen rather emanate belief problem met efficiency effectiveness system objective joint effort insurance department insurance implement premium substantially reduce objectionable restriction remove consequence insurance administration almost crime insurance subsidize tax revenue disagree philosophy commissioner citizen purchase crime insurance coverage facility deliberation contact appropriate official insurance administration initiate implement crime insurance flood insurance tragic loss life flood water hit past area reveal percentage citizen purchase economic available flood insurance community citizen participate insurance department initiate public inform urge join effort community participation citizen purchase flood insurance event portray quite vividly flood insurance lack communication community citizen explore possibility related situation credit insurance statute regulate rate credit life credit health insurance value necessity type insurance standpoint consumer merchant institution danger temptation extent abuse accompany transaction practical responsible system regulate credit life health insurance rate enact legislation adversely afford consumer supervision credit life health insurance framework legislation commissioner insurance legislation understand available financial institution past continued effort develop acceptable uniform consumer credit code regulatory applicable credit insurance rate proposal supervision type insurance uniform consumer credit code enact legislation finance cost consumer strongly regulation credit life health insurance rate	39	17.8%	14	17.0%	74	12.2%

Table B.6: Topic Labels Assigned to Sample of Cleaned Document Excerpts *(continued)*

Text	Topic Assn. 1	Topic 1 Prop.	Topic Assn. 2	Topic 2 Prop	Topic Assn. 3	Topic 3 Prop
academic proficiency quality conjunction teacher salary fundamental effort strengthen quality classroom intern approximately million fund current available revenue generate sale tax budget fund administration certification test teacher support remember enhance quality translate requirement choose teach profession teacher salary compound problem few quality teacher applicant school aid tenure governor education emerge administration struggle depressed strength commitment maintain quality competitive system education although goal met support school district significant commitment difficult time secure education attain enactment tax production inception tax generate million revenue fund initiative school fund appropriation school district grown school budget limitation varied period cumulative outcome percent school district budget school finance school employee retirement system non contributory budget percent aid income tax rebate support million million fund sale tax proposal implement approximate eight percent teacher salary tax million tax approximately million homeowner experience current school transportation aid school district fund percent cost current resource percent support resource percent available sale tax cost public education met aid local tax tax overburden face challenge support ensure maintenance quality system minimize tax doubly difficult task resource available revenue sale tax support school district build option maximize teacher tax realistic budget anticipate inflation rate educational service population administration sought ensure range educational develop	108	33.8%	25	25.7%	104	7.6%

Table B.6: Topic Labels Assigned to Sample of Cleaned Document Excerpts *(continued)*

Text	Topic Assn. 1	Topic 1 Prop.	Topic Assn. 2	Topic 2 Prop	Topic Assn. 3	Topic 3 Prop
box symmetric asymmetric symmetric meaning run get outside computer maximum back tremendous symmetric get phone direction horrendous stuff forward technology forward sure symmetrical cost effective incredibly performance requirement policy access interconnection anybody turf regulatory public policy sort deal citizen everything school economics daily life citizen pot elain service developmental community system credit card easy downtown back insurance company easy feel citizen system fully switch get feature survivable band band get network handle voice data video communication citizen chart let back highlight anybody school system want get school system get type hit accurate dot back slash current sound band aid arm want condition type standard want get dial phone understand system folk ability human being understand policy system telephone provider structure tremendous chart band width left telephone people pot telephone service band telephone today technology symmetrical remember baud statistic percent phone company america phone line america get handle handshake scratchy noise somebody handshake speed effectually function phone line telephone service pipe size band integrate service digital network today technology maximum remember max technology today thousand line per band technology really transitionary technology phase phase phase faster everybody get policy forward get capability today million per million technology pipe pretend around technology	26	16.9%	92	16.6%	60	16.0%

Table B.6: Topic Labels Assigned to Sample of Cleaned Document Excerpts *(continued)*

Text	Topic Assn. 1	Topic 1 Prop.	Topic Assn. 2	Topic 2 Prop	Topic Assn. 3	Topic 3 Prop
system plow ahead aimlessly understand might winner loser environment look forward partner push redevelopment unused industrial enact legislation instrumental create job display job employment expect grow week senator receipt help revitalize contaminate parcel fund recovery building revolve loan fund assist clean unused industrial site tax credit rehabilitation industrial task climate tax regulatory issue significant improvement forth significant streamline regulate septic system available fund completely review process environment unnecessary burden expense developer homeowner encourage support labor community mall economic growth issue look forward group economic issue mall encourage mall expansion improvement congratulation attractive competitive location region grow develop everyone job foundation family foundation family wife marilyn source tremendous support throughout life past motivation family family boost expand availability job ensure service affordable ensure policy support family rather tear apart week welfare package help family help examination convince ever welfare system life child quarter mother welfare today welfare system child apartment independently proposal proposal system pregnant wrong message message woman life barely pregnant longer want punish want explore potential adult child men child equally responsibility let abrogate proposal shift spend welfare longer system subsistence benefit individual earnings benefit supplement earnings income family eligible help maintain current benefit structure expense want president proposal	46	13.9%	42	13.4%	86	12.5%

Table B.6: Topic Labels Assigned to Sample of Cleaned Document Excerpts *(continued)*

Text	Topic Assn. 1	Topic 1 Prop.	Topic Assn. 2	Topic 2 Prop	Topic Assn. 3	Topic 3 Prop
reform streamline modernize agency payment faster efficient save around million user stop shop entrepreneur look easy college university individual card system onto system save million annually accountability transparency process president leadership issue towards system consolidate commitment department health agriculture system thank agency dedication efficient building suite solution pursue save taxpayer waste serve customer people efficiently idea share administration idea look waste abuse besides serious budget problem face unsustainable pension system currently unfunded billion pension system money consequence employee severe public employee legislator want path pension problem look forward passing reform solvency pension system reform commitment current public employee health turn health welfare citizen study nation health citizen unacceptable productivity hundred million thousand preventable death forward get access affordable health care market initiative encourage people november sent message initiative citizen purchase private heal insurance health pursue challenge local solution help citizen gain access affordable health insurance support innovative base insure public private partnership help affordable health insurance option employee building suite likewise foundation base potential health insurance consumer ago pass legislation creation health build low consumer purchase health insurance ultimately choice choice department health innovative public private initiative certify healthy encourage healthy living eat initiative grow certify healthy school certify healthy	96	52.7%	6	19.3%	88	8.1%

Table B.6: Topic Labels Assigned to Sample of Cleaned Document Excerpts *(continued)*

Text	Topic Assn. 1	Topic 1 Prop.	Topic Assn. 2	Topic 2 Prop	Topic Assn. 3	Topic 3 Prop
diligence parent welfare recent crime drug incentive unprotected border mandate recommitment enough restore hun balance criminal system family dignity resource prevention detection arrest prosecution combat threat security employment welfare client legislator senator nave chris whatever crime help investment dividend formation crack anti million save public assistance cost team fugitive felon want safe investment quantify cent friend poignant investment friend herb jennifer jamie return occasion participate investment health care ing april journal grown level friend thank economic political influence imagine mother father certainly governor accomplishment achieve escort governor rostrum create unparalleled environment quality lieutenant governor cabinet people around chosen advance human understand thank senator joint creator senator escort planet share arm gift squander repression greed pander remainder joint past human asset quality quality aware foster ability individual men achieve assure orderly maintain enhance birthright senator hair democratic meek grant today across personally neal loss freedom loss freedom beverage nation naturally bless nation licensee reunite dream thank generous people ground effective margie family want thank friend margie senator neal enthusiasm dedication emulate reward experience road drive create drunk drive commission wife girl share crisis triumph establishment policy daughter grown child adult standard educational happiest life reimbursement enforcement stood create drunk lead dream	9	20.7%	2	14.5%	99	13.0%

Table B.6: Topic Labels Assigned to Sample of Cleaned Document Excerpts *(continued)*

Text	Topic Assn. 1	Topic 1 Prop.	Topic Assn. 2	Topic 2 Prop	Topic Assn. 3	Topic 3 Prop
thousand educator elementary school care deeply child student teach standard accountability choice money quick magic cure student fought front toward goal comprehensive building block student achievable workable building block standard assessment school educator accountable implement commission student fully fund budget want thank leadership willingness commission fully implement system standard assessment accountability among nation appreciate extraordinary commission legislator educator employer parent kay teacher teacher today kay thank share classroom experience importantly commitment student let commend standard accountability thorough efficient system school standard fully clearly understood educator student parent building block prepare building block childhood education health care child school child foundation lifelong learn elementary school refocus help grow assure infant toddler life commitment child health insurance upgrade skill teacher leader childhood education public school learn student behind school across kindergarten school district serve child evidence kindergarten help ensure student guest ecole alternative elementary school along teacher join kindergartener abigail sow thanks building block classroom equip support legislature implement billion child safe secure school building governor invest nearly million school building renovation repair billion school building average project school project school construction steady ambitious school building nation thank fourth building block partner ensure highly train teacher classroom principal almost million teacher training	91	71.3%	67	8.3%	96	6.4%

B.7 Extensions of the Delta Statistic Approach

The delta-statistic method proposed in the first paper is used in papers two and three to test for the comparability of text sources. The approach, however, is extensible to several other research methods in cases where high-dimensional, sparse data such as text are available. The method may be used for the benefit of balance checking, and block randomization to reduce sampling variability, in experimental designs. We should expect that if a randomization has by blocking or otherwise properly eradicated unobserved heterogeneity in a sample of subjects, then subjects assigned to each condition should on average speak similarly; similarly, if we take latent traits to be useful to block on, then the models may produce a blocking design that reduces sampling variability by using text. In a time of large-scale RCTs, A/B testing, and experimentation that is linked to unstructured data, this alternative use case for the delta-statistic may find some use.

B.7.1 Application to Balance Checks in Randomized Controlled Trials

Researchers usually run a balance check after they randomly assign subjects to the arms of an experiment. The purpose of the balance check is to indicate if random assignment might have been unsuccessful. In a balance check, the researcher compares the central tendencies of every pre-treatment covariate in one arm to the central tendencies in every other arm. If any of these differences are greater than what one would expect by random chance, the balance check suggests that random assignment

might have been unsuccessful. "Unsuccessful" random assignment can challenge claims on causal inference, because it suggests either an induced relationship between the mechanism of assignment and the outcome, or the potential for spurious correlation in downstream analyses.

Text can be used in conjunction with pre-treatment covariance, or as an alternative to them when data are unavailable, to run balance checks. Language enhances balance checks because it can capture variance resulting from omitted variables. For instance, in studies using social media or blogs to study fake news (Guess et al. 2018, *e.g.*), data on the political preferences of the participants may not be available (the only available data may be the observed text); text collected from social media posts can be used as an indicator for the political preferences of the subject pool prior to the intervention.

It is also useful because it can be highly related to outcomes of interest. For example, if a Parkinson's disease researcher wishes to study the effect of vocalization therapy on physician assessments of disease progression (Mozer et al. 2018, *e.g.*), then the text from the patient's chart notes are likely highly related to which patients will see the best improvements, especially from a heterogeneous treatment effects perspective.

To use text as part of a balance check, the researcher should compare the language among the subjects in one arm of the design to the language among subjects in the other arms. Text, however, is not useful for balance checking in its sparse structured form. The reasons are two. First, any given word, token, or feature may be distributed sparsely with respect to experimental arms, and there may not be enough information

contained in any arm to compute notions of statistical difference. There may not be enough patient charts using the word "exacerbate" in the treatment group to compute a standard error over that word.

Second, the comparison of hundreds of thousands of features across experimental arms increases the probability of a false positive; it increases the probability that the researcher may incorrectly reject the random assignment because there is a false difference between the usage rate of the token in one arm versus another arm. The delta statistic approach detailed in this chapter overcomes these complications, reducing the dimensionality of the sparse text matrix down to fewer, "clustered" pre-treatment co-variates, and then simulating the null distribution to use.

B.7.2 Application to Statistical Blocking

Statistical blocking is a technique which reduces the effective sample size needed to draw population inferences. The technique breaks an experiment into several smaller experiments, each of which will experience a smaller degree of sampling variability with respect to the outcome. It matches treatment subjects with high expected outcomes to control subjects with high expected outcomes, and lows to lows, such that the difference between treatment and control for the highs and the lows is less variable than it would have been if the highs and the lows were compared to each other.

Researchers may block on pre-treatment text, just as they would normally do using pre-treatment covariates. Using text for blocking is advantageous for the same reasons discussed earlier, in appendix B.7.1 on applications to balance checks. Language

can enhance pre-treatment blocking because it can capture variance resulting from omitted variables. Language can also enhance pre-treatment blocking because it can be highly related to outcomes of interest. See section appendix B.4.1 for examples.

The delta statistic approach detailed earlier in this chapter can be used to block on pre-treatment text because it overcomes the complications of sparsity and errors in inferences. The dimension reduction step is most important, since it produces a number of features that can be used for blocking or matching in the usual way. The test statistic is important, however, in the case the outcome of interest is also text. If a strong theory drives the hypothesis that the experimental intervention should change the way the subjects verbal or text-based response, then the statistic may be used to test for differences in the speech patterns of any two arms of the experiment.

Application When Neither Data Source Has Annotations

It is worth pointing out that the delta-statistic method may be used in cases when the corpora of interest entirely lack annotations. The techniques established here do not require annotation to establish the comparability of corpora. This is important because it enables unsupervised learning techniques to check for latent structure patterns, using the theory developed in section 1.1. Though the purpose of developing the method is to validate joint scaling exercises involving text and observed dimensions like ideology, the extensibility of the method to cases in which there are not annotations – ontology discovery – merits further investigation.

B.8 Software Produced in the Course of this Dissertation

The dissertation contributed as a bonus several discrete computational advancements to assist in social scientific research. The first is *doc2text*. Historical data development is critical for the viability of novel NLP analyses and products, but developing text corpora can be quite difficult and expensive. Much of the information we are interested in are locked away in poorly scanned images. Enter the magic of computer vision. *doc2text* harnesses the power of computer vision to assist in the rapid development of research ideas and new data products. Appendix B.8 demonstrates how the program was used to speed up data processing for the State of the States collection. The second is *RA Booster*. RA Booster quickly scales the ability of researchers to complete projects that entail the annotation or quality assurance of text data. This project helped speed up quality assurance work by research assistants assigned to the State of the States project.

Figure C1: Scaling Text Research with *RA Booster!*

Bibliography

APSA Committee on Political Parties. 1950. "A Report of the Committee on Political Parties: Toward a More Responsible Two-Party System." 45 (2): 488–499.

Abramowitz, Alan I, and Steven Webster. 2016. "The rise of negative partisanship and the nationalization of US elections in the 21st century." *Electoral Studies* 41:12–22.

Achen, Christopher H. 1977. "Measuring representation: Perils of the correlation coefficient." *American Journal of Political Science*: 805–815.

— . 2002. "Parental Socialization and Rational Party Identification." *Political Behavior* 24 (2): 151–170.

Achen, Christopher H, and Larry M Bartels. 2017. *Democracy for realists: Why elections do not produce responsive government*. Vol. 4. Princeton University Press.

Ahler, Douglas J, and David E Broockman. 2018. "The delegate paradox: Why polarized politicians can represent citizens best." *The Journal of Politics* 80 (4): 1117–1133.

Airoldi, Edoardo M. 2014. *Presentation on Coherence in Topic Models*.

Airoldi, Edoardo M, et al. 2014. "Introduction to Mixed Membership Models and Methods." *Handbook of mixed membership models and their applications* 100:3–14.

Aldrich, John H, and John D Griffin. 2003. "The presidency and the campaign: creating voter priorities in the 2000 election." *The presidency and the political system* 7:239–256.

Alesina, Alberto, and Howard Rosenthal. 1995. *Partisan politics, divided government, and the economy*. Cambridge University Press.

Alvarez, R Michael. 1998. *Information and Elections*. University of Michigan Press.

Anderson, D, and K Burnham. 2004. *Model selection and multi-model inference*. 2nd. Vol. 63. NY: Springer-Verlag.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Ansolabehere, Stephen, Jonathan Rodden, and James M. Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102 (2): 215–232.

Ansolabehere, Stephen, James M. Snyder, Jr., and Charles Stewart. 2001. "Candidate positioning in US House elections." *American Journal of Political Science* 45 (1): 136–159.

Arora, Sanjeev, et al. 2013. "A practical algorithm for topic modeling with provable guarantees." In *International Conference on Machine Learning*, 280–288.

Ash, Elliott. 2015. *The political economy of tax laws in the US states*. Tech. rep. Working Paper, Columbia University.

Asuncion, Arthur, et al. 2009. "On smoothing and inference for topic models." In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, 27–34. AUAI Press.

Bachrach, Peter, and Morton Baratz. 1962. "Two Faces of Power." *American Political Science Review* 56 (4): 947–952.

Bafumi, Joseph, and Michael C. Herron. 2010. "Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress." *American Political Science Review* 104 (03): 519–542.

Bafumi, Joseph, et al. 2005. "Practical issues in implementing and understanding Bayesian ideal point estimation." *Political Analysis* 13 (2): 171–187.

Bailey, Michael A. 2007. "Comparable preference estimates across time and institutions for the court, congress, and presidency." *American Journal of Political Science* 51 (3): 433–448.

Bailey, Michael, and Kelly H Chang. 2001. "Comparing presidents, senators, and justices: interinstitutional preference estimation." *Journal of Law, Economics, and Organization* 17 (2): 477–506.

Ban, Pamela, et al. 2017. "How newspapers reveal political power." *Political Science Research and Methods*.

Barber, Michael, and Nolan Mccarty. 2013. "Causes and Consequences of Polarization": 19–53.

Bartels, Larry M. 1991. "Constituency Opinion and Congressional Policy Making: The Reagan Defense Build Up." *The American Political Science Review* 85 (2): 457.

Bartels, Larry M. 2000. "Partisan and Voting Behavior, 1952-1996." *American Journal of Political Science* 44 (1): 35–50.

Barth, Jay, and Margaret R Ferguson. 2002. "American governors and their constituents: The relationship between gubernatorial personality and public approval." *State Politics & Policy Quarterly* 2 (3): 268–282.

Baumgartner, Frank R, and Bryan D Jones. 2010. *Agendas and instability in American politics*. University of Chicago Press.

Bawn, Kathleen, et al. 2012. "A theory of political parties: Groups, policy demands and nominations in American politics." *Perspectives on Politics* 10 (03): 571–597.

Beck, Nathaniel. 1983. "Time-varying parameter regression models." *American Journal of Political Science*: 557–600.

— . 1985. "Estimating dynamic models is not merely a matter of technique." *Political Methodology*: 71–89.

Bennett, W Lance, and Shanto Iyengar. 2008. "A new era of minimal effects? The changing foundations of political communication." *Journal of communication* 58 (4): 707–731.

Benoit, Kenneth, Kevin Munger, and Arthur Spirling. 2019. "Measuring and explaining political sophistication through textual complexity." *American Journal of Political Science* 63 (2): 491–508.

Benoit, Kenneth, et al. 2018. "quanteda: An R package for the quantitative analysis of textual data." *J. Open Source Software* 3 (30): 774.

Berelson, Bernard. 1952. *Content analysis in communication research*. Free Press.

Berelson, Bernard, and Sebastian De Grazia. 1947. "Detecting collaboration in propaganda." *Public Opinion Quarterly* 11 (2): 244–253.

Berelson, Bernard, and Paul Felix Lazarsfeld. 1948. *The analysis of communication content*. Universitetets studentkontor.

Bergan, Daniel E. 2009. "Does grassroots lobbying work? A field experiment measuring the effects of an e-mail lobbying campaign on legislative behavior." *American politics research* 37 (2): 327–352.

Berkman, Michael, and Eric Plutzer. 2018. *Mood of the Nation Poll*. McCourtney Institute for Democracy.

Berry, William D, and Brady Baybeck. 2005. "Using geographic information systems to study interstate competition." *American Political Science Review* 99 (04): 505–519.

Binder, Sarah A. 2004. *Stalemate: Causes and consequences of legislative gridlock*. Brookings Institution Press.

— . 2017. "Polarized we govern?" In *Governing in a polarized age: Elections, parties, and political representation in America*, ed. by Alan S Gerber and Eric Schickler. Cambridge University Press.

Binder, Sarah. 2015. "The dysfunctional congress." *Annual Review of Political Science* 18:85–101.

— . 2016. "Congress and Policy Making in the 21st Century," no. February: 350.

Bischof, Jonathan, and Edoardo M Airoldi. 2012. "Summarizing topical content with word frequency and exclusivity." In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 201–208.

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

Boehmke, Fred, et al. 2018. "Influencers, Innovators, and Ideology: How Policy Innovativeness and Leadership Vary by Policy Ideology." Conference on State Policy and Politics, St. Louis, Missouri.

Bond, Jon R, and Richard Fleisher. 2001. "The polls: Partisanship and presidential performance evaluations." *Presidential Studies Quarterly* 31 (3): 529–540.

Bonica, Adam. 2013. "Ideology and interests in the political marketplace." *American Journal of Political Science* 57 (2): 294–311.

Borges, Jorge Luis. 1941. "La biblioteca de Babel." *Obras completas* 1.

Bowling, Cynthia J, and Margaret R Ferguson. 2001. "Divided government, interest representation, and policy differences: Competing explanations of gridlock in the fifty states." *Journal of Politics* 63 (1): 182–206.

Brandeis, Louis Dembitz. 1932. "Dissenting opinion." *New State Ice Co. v. Liebmann* 285.

Burden, Barry C, and Amber Wichowsky. 2010. "Local and national forces in congressional elections." In *The Oxford handbook of American elections and political behavior*.

Butler, Daniel M., and David W. Nickerson. 2011. "Can Learning Constituency Opinion Affect How Legislators Vote? Results from a Field Experiment." *Quarterly Journal of Political Science* 6 (1): 55–83.

Butler, Daniel M., and Eleanor Neff Powell. 2014. "Understanding the Party Brand: Experimental Evidence on the Role of Valence." *The Journal of Politics* 76 (2): 492–505.

Campbell, Angus, et al. 1960. "The American Voter." *New York: John Wiley and Sons*.

Carmines, Edward G, and James A Stimson. 1989. *Issue evolution: Race and the transformation of American politics*. Princeton University Press.

Chamberlain, Lawrence H. 1946. "The President, Congress, and Legislation." *Political Science Quarterly* 61 (1): 42–60.

Chiou, Fang-Yi, and Lawrence S Rothenberg. 2003. "When pivotal politics meets partisan politics." *American Journal of Political Science* 47 (3): 503–522.

Chubb, John E. 1988. "Institutions, the economy, and the dynamics of state elections." *American Political Science Review* 82 (1): 133–154.

Civil Rights Digital Library. 2019. "Civil Rights Digital Library." http://crdl.usg.edu/people/p/patterson_john_malcolm_1921.

Clark, Tom S, and Benjamin E Lauderdale. 2012. "The genealogy of law." *Political Analysis* 20 (3): 329–350.

Clinton, Joshua D. 2006. "Representation in Congress: constituents and roll calls in the 106th House." *Journal of Politics* 68 (2): 397–409.

Clinton, Joshua D, and David E Lewis. 2008. "Expert opinion, agency characteristics, and agency preferences." *Political Analysis* 16 (1): 3–20.

Clinton, Joshua D, et al. 2012. "Separated powers in the United States: The ideology of agencies, presidents, and congress." *American Journal of Political Science* 56 (2): 341–354.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The statistical analysis of roll call data." *American Political Science Review* 98 (2): 355–370.

Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics. In Ideology and Discontent, ed. David Apter. New York: Free Press."

Coombs, Clyde H. 1960. "A theory of data." *Psychological review* 67 (3): 143.

Cotter, Cornelius C, et al. 1989. *Party organizations in American politics*. University of Pittsburgh Pre.

Cox, Gary W, and Mathew D McCubbins. 1993. *Legislative leviathan: party government in the House*. Cambridge; New York: Cambridge University Press.

— . 2005. *Setting the agenda: Responsible party government in the US House of Representatives*. Cambridge University Press.

Dahl, Robert. 1961. *Who governs? Democracy and power in an American city*. New Haven, CT: Yale University Press.

Dancey, Logan, and Geoffrey Sheagley. 2013. "Heuristics behaving badly: Party cues and voter knowledge." *American Journal of Political Science* 57 (2): 312–325.

Deerwester, Scott, et al. 1990. "Indexing by latent semantic analysis." *Journal of the American society for information science* 41 (6): 391.

Denny, Matthew J, and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* 26 (2): 168–189.

DiLeo, Daniel. 1997. "Dynamic Representation in the United States: Effects of the Public Mood on Governors' Agendas." *State and Local Government Review* 29 (2): 98–109.

Douglas, Waples, Bernard Berelson, and Franklin R Bradshaw. 1940. *What reading does to people*. University of Chicago Press.

Dragu, Tiberiu, and Xiaochen Fan. 2016. "An agenda-setting theory of electoral competition." *The Journal of Politics* 78 (4): 1170–1183.

Druckman, James N, and Lawrence R Jacobs. 2006. "Lumpers and splitters: The public opinion information that politicians collect and use." *International Journal of Public Opinion Quarterly* 70 (4): 453–476.

Druckman, James N, Lawrence R Jacobs, and Eric Ostermeier. 2004. "Candidate strategies to prime issues and image." *The Journal of Politics* 66 (4): 1180–1202.

Dunkelman, Marc J. 2014. *The vanishing neighbor: The transformation of American community*. WW Norton & Company.

Duranti, Alessandro, and Charles Goodwin. 1992. *Rethinking context: Language as an interactive phenomenon*. Vol. 11. Cambridge University Press Cambridge.

Dynarski, Susan. 2000. *Hope for whom? Financial aid for the middle class and its impact on college attendance*. Tech. rep. National bureau of economic research.

Edwards III, George C, Andrew Barrett, and Jeffrey Peake. 1997. "The legislative impact of divided government." *American journal of political science*: 545–563.

Egami, Naoki, et al. 2018. "How to make causal inferences using texts." *arXiv preprint arXiv:1802.02163*.

Epstein, Lee, and Jeffrey A Segal. 2000. "Measuring issue salience." *American Journal of Political Science*: 66–83.

Erikson, Robert S. 1978. "Constituency opinion and congressional behavior: A reexamination of the Miller-Stokes representation data." *American Journal of Political Science*: 511–535.

Erikson, Robert S, Michael B Mackuen, and James A Stimson. 2002. *The Macro Polity*. Cambridge University Press.

Fiorina, Morris P. 1996. *Divided government*. Allyn & Bacon.

— . 2017. "The (Re) Nationalization of Congressional Elections." *Hoover Institution*.

Fiorina, Morris P, Samuel J Abrams, and Jeremy C Pope. 2005. "Culture war." *The myth of a polarized America* 3.

Gelman, Andrew, and Guido Imbens. 2013. "Why Ask Why? Forward Causal Inference and Reverse Causal Questions."

Gentzkow, Matthew, Bryan T Kelly, and Matt Taddy. 2017. *Text as data*. Tech. rep. National Bureau of Economic Research.

Gentzkow, Matthew, and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78 (1): 35–71.

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy. 2016. *Measuring polarization in high-dimensional data: Method and application to congressional speech*. Tech. rep. National Bureau of Economic Research.

Gerbner, George. 1985. "Mass media discourse: Message system analysis as a component of cultural indicators." *Discourse and communication. new approaches to the analysis of mass media discourse and communication*: 13–25.

Green, Donald P, Bradley Palmquist, and Eric Schickler. 2002. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. Yale University Press.

Griffiths, Thomas L, and Mark Steyvers. 2004. "Finding scientific topics." *Proceedings of the National academy of Sciences* 101 (suppl 1): 5228–5235.

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

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

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21 (3): 267–297.

Grose, Christian R, and Bruce I Oppenheimer. 2007. "The Iraq war, partisanship, and candidate attributes: Variation in partisan swing in the 2006 US House elections." *Legislative Studies Quarterly* 32 (4): 531–557.

Groseclose, Tim, Steven D Levitt, and James M Snyder. 1999. "Comparing interest group scores across time and chambers: Adjusted ADA scores for the US Congress." *American political science review* 93 (1): 33–50.

Groseclose, Tim, and Jeffrey Milyo. 2005. "A measure of media bias." *The Quarterly Journal of Economics* 120 (4): 1191–1237.

Grossmann, Matt. 2014. "Policymaking in Red and Blue: Asymmetric Partisan Politics and American Governance." In *APSA 2014 Annual Meeting Paper*.

Grumbach, Jacob M. 2018. "From backwaters to major policymakers: Policy polarization in the states, 1970–2014." *Perspectives on Politics* 16 (2): 416–435.

Grynaviski, Jeffrey D. 2010. *Partisan Bonds: Political Reputations and Legislative Accountability*. Cambridge University Press.

Guess, Andrew, et al. 2018. "Fake news, Facebook ads, and misperceptions."

Hamilton, James D. 2010. "Regime switching models." In *Macroeconometrics and time series analysis*, 202–209. Springer.

Hand, David J. 2006. "Classifier technology and the illusion of progress." *Statistical science*: 1–14.

Handler, Abram, et al. 2016. "Bag of what? simple noun phrase extraction for text analysis." In *Proceedings of the First Workshop on NLP and Computational Social Science*, 114–124.

Hastie, Tibshirani, R Tibshirani, and J Friedman. 2009. *The elements of statistical learning*. New York: Springer.

Hausman, Jerry A. 1983. "Specification and estimation of simultaneous equation models." *Handbook of econometrics* 1:391–448.

Heckman, James J, and James M Snyder Jr. 1996. *Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators*. Tech. rep. National bureau of economic research.

Hehman, Eric, Samuel L Gaertner, and John F Dovidio. 2011. "Evaluations of presidential performance: Race, prejudice, and perceptions of Americanism." *Journal of Experimental Social Psychology* 47 (2): 430–435.

Hertel-Fernandez, Alexander. 2016. "Explaining liberal policy woes in the states: The role of donors." *PS: Political Science & Politics* 49 (3): 461–465.

Hertel-Fernandez, Alexander, and Konstantin Kashin. 2015. "Capturing business power across the states with text reuse." In *annual conference of the Midwest Political Science Association, Chicago, April*, 16–19.

Herzik, Eric B. 1983. "Governors and issues: A typology of concerns." *State Government* 56 (2): 58–64.

Hogg, David. 2019. "How Can Machine-Learning Methods Help to Make Scientific Inferences?" In *Columbia University's New York Data Science Seminar Series*.

Holsti, Ole R. 1969. *Content analysis for the social sciences and humanities*. Addison-Wesley: Reading, MA.

Hopkins, Daniel J. 2018. *The increasingly United States: How and why American political behavior nationalized*. University of Chicago Press.

Hopkins, Daniel J, and Gary King. 2010. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54 (1): 229–247.

Hornik, Kurt, and Bettina Grün. 2011. "topicmodels: An R package for fitting topic models." *Journal of Statistical Software* 40 (13): 1–30.

Iacus, Stefano M, Gary King, and Giuseppe Porro. 2019. "A theory of statistical inference for matching methods in applied causal research." *Political Analysis*.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, not ideology: A social identity perspective on polarization." *Public Opinion Quarterly* 76:405–431.

Jacobson, Gary C. 2015. "It's Nothing Personal: The Decline of the Incumbency Advantage in US House Elections." *Journal of Politics* 77 (3): 861–873.

Jansa, Joshua M, Eric R Hansen, and Virginia H Gray. 2015. *Copy and paste lawmaking: The diffusion of policy language across american state legislatures*. Tech. rep. Working Paper.

Jelveh, Zubin, Bruce Kogut, and Suresh Naidu. 2015. "Political language in economics."

Jennings, M Kent. 1996. "Political knowledge over time and across generations." *Public Opinion Quarterly* 60 (2): 228–252.

Jensen, Jacob, et al. 2012. "Political polarization and the dynamics of political language: Evidence from 130 years of partisan speech [with comments and discussion]." *Brookings Papers on Economic Activity*: 1–81.

Jerzak, Connor T, Gary King, and Anton Strezhnev. 2018. "Readme2: An R Package for Improved Automated Nonparametric Content Analysis for Social Science."

Jessee, Stephen A. 2009. "Spatial Voting in the 2004 Presidential Election." *American Political Science Review* 103 (01): 59.

Jessee, Stephen A. 2012. *Ideology and spatial voting in American elections*. Cambridge University Press.

Jessee, Stephen. 2016. "(How) can we estimate the ideology of citizens and political elites on the same scale?" *American Journal of Political Science* 60 (4): 1108–1124.

Karol, David. 2009. *Party Position Change in American Politics: Coalition Management*. 326. Cambridge, UK ; New York: Cambridge University Press.

Katz, Daniel, and Samuel J Eldersveld. 1961. "The impact of local party activity upon the electorate." *Public Opinion Quarterly* 25 (1): 1–24.

Key, Valdimer Orlando. 1964. *Parties, politics and pressure groups*.

Key, Valdimir Orlando. 1950. *Southern politics in state and nation*. University of Tennessee Press.

King, Gary, and Will Lowe. 2003. "An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design." *International Organization* 57 (3): 617–642.

Kirkland, Patricia A, and Justin H Phillips. 2018. "Is divided government a cause of legislative delay?" *Quarterly Journal of Political Science* 13 (2): 173–206.

Klarner, Carl. 2003. "The measurement of the partisan balance of state government." *State Politics & Policy Quarterly* 3 (3): 309–319.

Kousser, Thad, and Justin H Phillips. 2012. *The power of American governors: Winning on budgets and losing on policy*. Cambridge University Press.

Krehbiel, Keith. 1998. *Pivotal Politics: A Theory of US Lawmaking*. University of Chicago Press.

Krippendorff, Klaus. 1980. "Validity in content analysis."

— . 2012. *Content Analysis: An Introduction to its Methodology*. 3rd. SAGE Publications, Inc.

Krippendorff, Klaus, and W Paisley PJ Stone. 1969. "The analysis of communication content." *New York: John vliley & Sons*.

Kurtz, Karl T, Alan Rosenthal, and Cliff Zukin. 2003. "Citizenship: A challenge for all generations." In *Denver, CO: National Conference of State Legislatures*.

Lange, Tilman, et al. 2004. "Stability-based validation of clustering solutions." *Neural computation* 16 (6): 1299–1323.

Lasswell, Harold D. 1927. *Propaganda technique in the world war*. Ravenio Books.

— . 1938. "What psychiatrists and political scientists can learn from one another." *Psychiatry* 1 (1): 33–39.

Laver, Michael, et al. 2003. "Extracting policy positions from political texts using words in data." *American Political Science Review* 97 (2): 311–331.

Lax, Jeffrey R., and Justin H. Phillips. 2012. "The democratic deficit in the states." *American Journal of Political Science* 56 (1): 148–166.

Lee, Frances E. 2013. "Presidents and Party Teams: The Politics of Debt Limits and Executive Oversight, 2001-2013." *Presidential Studies Quarterly* 43 (4): 775–791.

Levendusky, Matthew S. 2010. "Clearer cues, more consistent voters: A benefit of elite polarization." *Political Behavior* 32 (1): 111–131.

Levendusky, Matthew. 2009. *The Partisan Sort: How liberals became Democrats and conservatives became Republicans*. Chicago: University of Chicago Press.

Levy, Jacob T. 2007. "Federalism, liberalism, and the separation of loyalties." *American Political Science Review* 101 (3): 459–477.

Lewis, Jeffrey B, and Chris Tausanovitch. 2015. "When Does Joint Scaling Allow for Direct Comparisons of Preferences?" In *Conference on ideal point models*, vol. 1.

Lippmann, Walter. 1922. *Public opinion*. New York: Harcourt, Brace / Company.

Londregan, John. 2000. *Faustian bargains: Legislative institutions and Chile's democratic transition*. Cambridge University Press New York.

Lord, F.M., and M.R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Menlo Park.

Loughran, Tim, and Bill McDonald. 2019. "Textual Analysis in Finance." *Available at SSRN 3470272*.

Lowe, Will, and Kenneth Benoit. 2013. "Validating estimates of latent traits from textual data using human judgment as a benchmark." *Political analysis* 21 (3): 298–313.

Lupia, Arthur. 1992. "Busy voters, agenda control, and the power of information." *American Political Science Review* 86 (2): 390–403.

— . 1994. "Shortcuts versus encyclopedias: Information and voting behavior in California insurance reform elections." *The American Political Science Review* 88 (1): 63–76.

Mann, Thomas E, and Norman J Ornstein. 2016. *It's even worse than it looks: How the American constitutional system collided with the new politics of extremism*. Basic Books.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 1991. *An introduction to information retrieval*.

Manski, Charles F. 1990. "Nonparametric bounds on treatment effects." *The American Economic Review* 80 (2): 319–323.

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

— . 2007. "Assessing preference change on the US Supreme Court." *The journal of law, economics, & organization* 23 (2): 365–385.

Martin, Gregory J, and Joshua McCrain. 2019. "Local news and national politics." *American Political Science Review* 113 (2): 372–384.

Martin, Gregory J, and Ali Yurukoglu. 2017. "Bias in cable news: Persuasion and polarization." *American Economic Review* 107 (9): 2565–99.

The $2.2 trillion stimulus can't be the last thing Congress does on coronavirus. 2020.

Mayhew, David R. 1974. "Congressional Elections: The Case of the Vanishing Marginals." *American Journal of Political Science* 6 (3): 295–317.

— . 2005. *Divided we govern: Party control, lawmaking and investigations, 1946-2002*. Yale University Press.

— . 2011. *Partisan balance: Why political parties don't kill the US constitutional system*. Princeton University Press.

McCarty, Nolan. 2016. "Polarization, Congressional Dysfunction, and Constitutional Change." *Ind. L. Rev.* 50:223.

McCombs, Maxwell E, and Donald L Shaw. 1972. "The agenda-setting function of mass media." *Public opinion quarterly* 36 (2): 176–187.

Mcauliffe, Jon D, and David M Blei. 2008. "Supervised topic models." In *Advances in neural information processing systems*, 121–128.

Michel, Jean-Baptiste, et al. 2011. "Quantitative analysis of culture using millions of digitized books." *science* 331 (6014): 176–182.

Miller, Warren E, and Donald E Stokes. 1963. "Constituency Influence in Congress." *The American Political Science Review* 57 (1): 45–56.

Mimno, David, and Moontae Lee. 2014. "Low-dimensional embeddings for interpretable anchor-based topic inference." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1319–1328.

Mimno, David, et al. 2011. "Optimizing semantic coherence in topic models." In *Proceedings of the conference on empirical methods in natural language processing*, 262–272. Association for Computational Linguistics.

Monroe, Burt L. 2013. "The five Vs of big data political science introduction to the virtual issue on big data in political science political analysis." *Political Analysis* 21 (V5): 1–9.

Monroe, Burt L, Michael P Colaresi, and Kevin M Quinn. 2008. "Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16 (4): 372–403.

Monroe, Burt L, and Ko Maeda. 2004. "Rhetorical ideal point estimation: Mapping legislative speech." *Society for Political Methodology, Stanford University*.

Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62 (3): 760–775.

Mosteller, Frederick, and David Wallace. 1964. "Inference and disputed authorship: The Federalist."

Mozer, Reagan, et al. 2018. "Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality." *arXiv preprint arXiv:1801.00644*.

Nacos, Brigitte L, et al. 1991. "Content analysis of news reports: Comparing human coding and a computer-assisted method." *Communication* 12 (2): 111–128.

Neuendorf, Kimberly A. 2016. *The content analysis guidebook*. Sage.

Orne, Martin T. 1962. "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications." *American psychologist* 17 (11): 776.

Osgood, Charles E. 1959. "The representational model and relevant research materials." In *Trends in content analysis*, 33–88. University of Illinois Press.

Page, Benjamin I, and Robert Y Shapiro. 1983. "Effects of Public Opinion on Policy." *The American Political Science Review* 77 (1): 175–190.

Petrocik, John R. 1996. "Issue ownership in presidential elections, with a 1980 case study." *American journal of political science*: 825–850.

Poole, Keith T, and Howard L Rosenthal. 2007. *Ideology and Congress*. 2 Revised. 361. New Brunswick: Transaction Publishers.

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

— . 1997. *Congress: A Political-economic History of Roll Call Voting*. 314. Oxford University Press.

Popkin, Samuel L. 1994. *The reasoning voter: Communication and persuasion in presidential campaigns*. University of Chicago Press.

Powell, G Bingham, and Guy D Whitten. 1993. "A cross-national analysis of economic voting: taking account of the political context." *American Journal of Political Science*: 391–414.

Putnam, Robert D. 2000. "Bowling alone: America's declining social capital." In *Culture and politics*, 223–234. Springer.

Quinn, Kevin M, et al. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–228.

RePass, David E. 1971. "Issue salience and party choice." *American Political Science Review* 65 (2): 389–400.

Riffkin, Rebecca. 2014. "Public faith in Congress falls again, hits historic low." *Gallup, June* 19.

Riker, William Harrison Riker, et al. 1996. *The strategy of rhetoric: Campaigning for the American Constitution*. Yale University Press.

Robert, Henry Martyn. 1915. *Robert's rules of order revised for deliberative assemblies*. Da Capo Press.

Roberts, Margaret E, Brandon M Stewart, and Edoardo M Airoldi. 2016. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* 111 (515): 988–1003.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2014. "stm: R package for structural topic models." *Journal of Statistical Software* 10 (2): 1–40.

Roberts, Margaret E, et al. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58 (4): 1064–1082.

Rogers, James R. 2005. "The impact of divided government on legislative production." *Public Choice* 123 (1): 217–233.

Rogers, Steven. 2016. "National forces in state legislative elections." *The ANNALS of the American Academy of Political and Social Science* 667 (1): 207–225.

— . 2017. "Electoral accountability for state legislative roll calls and ideological representation." *American Political Science Review* 111 (3): 555–571.

Rogowski, Jon C, and Joseph L Sutherland. 2016. "How Ideology Fuels Affective Polarization." *Political Behavior* 38 (2): 485–508.

Rosenthal, Alan. 1990. "Governors and legislators: Contending powers Washington." *DC: Congressional Quarterly*.

Ruppert, D. 2004. "Trimming and Winsorization." In *Encyclopedia of Statistical Sciences*. John Wiley / Sons, Inc.

Schattschneider, E E. 1960. *The Semi-Sovereign People: A Realist's View of Democracy in America*. Holt, Rhinehart / Winston, New York.

Schattschneider, Elmer Eric. 1942. *Party government*. Transaction Publishers.

Shapiro, Gilbert, and John Markoff. 1997. "Methods for drawing statistical inferences from text and transcripts." *Text analysis for the social sciences*: 9–31.

Shipan, Charles R. 2006. "Does divided government increase the size of the legislative agenda?" In *The Macropolitics of Congress*, 151–70. Princeton: Princeton University Press.

Shor, Boris, and Nolan McCarty. 2011. "The ideological mapping of American legislatures." *American Political Science Review* 105 (3): 530–551.

Shor, Boris, Nolan McCarty, and Christopher R Berry. 2011. "Methodological Issues in Bridging Ideal Points in Disparate Institutions in a Data Sparse Environment." *Available at SSRN 1746582*.

Shor, Boris, and Jon C Rogowski. 2018. "Ideology and the US congressional vote." *Political Science Research and Methods* 6 (2): 323–341.

Siegler, MJ. 2010. "Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003." TechCrunch. https://techcrunch.com/2010/08/04/schmidt-data/.

Sievert, Carson, and Kenneth Shirley. 2014. "LDAvis: A method for visualizing and interpreting topics." In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 63–70.

Simon, Herbert Alexander. 1997. *Models of bounded rationality: Empirically grounded economic reason*. Vol. 3. MIT press.

Sinclair, Barbara. 2016. *Unorthodox lawmaking: New legislative processes in the US Congress*. CQ Press.

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

Smith, Steven S. 2014. *The Senate Syndrome: The Evolution of Procedural Warfare in the Modern US Senate*. Vol. 12. University of Oklahoma Press.

Stokes, Donald E. 1963. "Spatial models of party competition." *American political science review* 57 (02): 368–377.

Sundquist, James L. 1968. "Politics and Policy: The Eisenhower." *Kennedy, and Johnson years* 59.

— . 1988. "Needed: A political theory for the new era of coalition government in the United States." *Political Science Quarterly* 103 (4): 613–635.

Taddy, Matt. 2012. "On estimation and selection for topic models." In *Artificial Intelligence and Statistics*, 1184–1193.

Tausanovitch, Chris, and Christopher Warshaw. 2013. "Measuring constituent policy preferences in congress, state legislatures, and cities." *The Journal of Politics* 75 (2): 330–342.

Teh, Yee W, et al. 2005. "Sharing clusters among related groups: Hierarchical Dirichlet processes." In *Advances in neural information processing systems*, 1385–1392.

Theriault, Sean M. 2008. *Party polarization in congress*. Cambridge University Press.

Tollison, Robert D. 1988. "Public choice and legislation." *Virginia Law Review*: 339–371.

Walgrave, Stefaan, Jonas Lefevere, and Michiel Nuytemans. 2009. "Issue ownership stability and change: How political parties claim and maintain issues through media appearances." *Political Communication* 26 (2): 153–172.

Webster, Steven W, and Alan I Abramowitz. 2017. "The ideological foundations of affective polarization in the US electorate." *American Politics Research* 45 (4): 621–647.

Weinberg, Micah. 2010. "Measuring governors' political orientations using words as data." *State Politics & Policy Quarterly* 10 (1): 96–109.

Wiener, Janet, and Nathan Bronson. 2014. "Facebook's Top Open Data Problems." https://research.fb.com/blog/2014/10/facebook-s-top-open-data-problems/.

Windett, Jason H, Jeffrey J Harden, and Matthew EK Hall. 2015. "Estimating dynamic ideal points for state supreme courts." *Political Analysis* 23 (3): 461–469.

Witten, Ian H, et al. 1999. *Managing gigabytes: compressing and indexing documents and images*. Morgan Kaufmann.

Wlezien, Christopher, and Stuart N Soroka. 2012. "Political institutions and the opinion–policy link." *West European Politics* 35 (6): 1407–1432.

Woon, Jonathan, and Jeremy C Pope. 2008. "Made in Congress? Testing the electoral implications of party ideological brand names." *The Journal of Politics* 70 (3): 823–836.

Zaller, John. 1991. "Information, Values, and Opinion." *American Political Science Review* 85:1215–1237.