Measuring the Issue Content of Supreme Court Opinions*

Abstract

The opinions of the U.S. Supreme Court are central to volumes of research on law, courts, and politics. To understand these complex and oft-lengthy documents, scholars frequently rely on dichotomous indicators of opinion content. Using all U.S. Supreme Court opinions from 1793 to 2010, I first demonstrate the shortcomings of the dominant approach to measuring opinion content, and then propose the use of unsupervised topic models for measuring issue content. Rather than reflecting a single issue dimension, and doing so in their entirety, opinions regularly address multiple topics and devote relative proportions of the Court's attention to those topics.

*All materials necessary to replicate the results presented in this article will be made available on Dataverse upon publication.

Perched atop the judiciary, the U.S. Supreme Court regularly publishes opinions which provide critical information on law and public policy. Understanding the information contained within opinions has therefore served as a principal objective for social scientists and legal scholars for generations. Yet parsing such voluminous and verbose documents poses particular challenges for the researcher, and coding the numerous dimensions of interest only magnifies the time and resource costs. To that end, scholars almost exclusively rely on the high quality coding available from the Supreme Court Database [SCD]. Indeed, Epstein, Knight and Martin (2003) –in their review of research on judicial behavior – identify the SCD as "so dominating in our discipline that it would certainly be unusual for a refereed journal to publish a manuscript whose data derived from an alternative source" (812).

But while the SCD features extensive and uncontroversial coding of a variety of attributes of the Court's opinion – codes which are omnipresent in research on the Court – a few of the variables have been the subject of extensive controversy. At the center of this controversy lies the identification of the issue under consideration by the Court in deciding a case (see, e.g., Shapiro 2009; Harvey and Woodruff 2011), from which the ideological direction of the Court's decision is determined. The issue codes are explicitly intended to measure the policy content of the opinion, reducing lengthy treatises to – in the overwhelming majority of cases – a single issue dimension which provides the basis of the Court's decision. As is frequently recognized, the identification of this dimension is critical to our understanding of the Court, including for example debate over strategic judicial behavior (Benesh and Spaeth 2007; Friedman 2006) and the Court's policymaking role (Baird 2004, 2007).

Though a seemingly simple task, identifying the policy content of a judicial opinion has proven difficult, nesting a series of complex – and increasingly apparent – problems. At the heart of these problems is what I term the *unidimensional assumption*, or the strong preference of prior research to assign each opinion to a single issue or policy area. Such a preference potentially obscures large amounts of important information for scholars interested in the Court, as noted by others (Shapiro 2009). As a sense of the scale of the potential

bias introduced by the *unidimensional assumption*, Shapiro revisited a random sample of 95 opinions and identified each of the issues addressed in the opinions. Of those 95, Shapiro identified 89 – or approximately 94% of the opinions – as addressing multiple issues. In contrast, only one of those 89 opinions was classified as addressing multiple issues by the SCD.

The implications are relatively profound for empirical research on the Court. Consider two recent articles. First, Harvey and Woodruff (2011) suggest systematic bias in issue coding introduced by coder's unconscious beliefs about the decision's ideological direction (Harvey and Woodruff 2011). Here, confirmation bias manifests in decisions over which of many potential issue areas is primary. By manipulating the choice of the single issue classifications – as is required by the coding protocol – the coder shifts the ideological direction of a vote, thereby matching expectations of case disposition and justice ideology. Second, recent work by Rice (2016) suggests dissenting opinion authors strategically emphasize alternative issue dimensions on which the case could be decided. Though consistent with the sort of heresthetical maneuvering identified at all other stages of Court decision-making (Epstein and Shvetsova 2002; Wedeking 2010; Black, Schutte and Johnson 2013), the findings are contrary to other prior work suggesting the Court primarily divides (and decides) solely on a single issue dimension in the overwhelming majority of cases. In both articles, the researchers document how the coding protocol – and specifically the unidimensional assumption – of the dominant measure of opinion content and Supreme Court issue attention yields measures of issue attention that are inappropriate for the purposes of most empirical research.

Alternatively, the classification of issues – or topics – is precisely the sort of task for which computational approaches to text analysis are ideally suited. The potential utility of computational approaches to studies of the Supreme Court and other judicial institutions is well-demonstrated by the volume of research in the area. In the first place, researchers have looked to improve variables already included in the SCD (e.g., McGuire and Vanberg 2005; Evans et al. 2007). For example, McGuire and Vanberg (2005) use Wordscores (Laver,

Benoit and Garry 2003) to measure the ideological direction of Supreme Court opinions, an approach built upon by Lauderdale and Clark (2014) who jointly model votes and issues to extract ideological dimensions within issue areas. In addition to refining existing measures, scholars have proposed expanding the coding scheme to other judicial venues (Szmer and Edwards 2011).

In this article, I leverage a variety of approaches to computational text analysis towards two ends: first, addressing the validity of the SCD's issue codes, and, second, proposing a more robust alternative based on topic models. I address the validity concerns surrounding the issue area codes by seeking to directly predict the classifications using the words actually employed in the opinions. Yet across a series of analyses using the texts of majority opinions across the entire history of the Supreme Court, I find that – no matter how the textual content of the opinion is operationalized – issue area codes are not recoverable at levels typically deemed empirically acceptable. Instead, the codes do not reflect a mapping of opinion content, with divergence typically occurring in a manner consistent with prior criticisms of the issue area codes. In lieu of these standard and crucial measures of the Court's attention, I instead demonstrate the utility of unsupervised topic models as a flexible measurement approach built directly from opinion content, with the resulting measures addressing each of the previously cited criticisms.[1]

The Importance of Opinion Content

Identifying the issues addressed by a given court case is critically important for research on law and courts. Often, scholars restrict their analyses to particular areas of the law in order to better identify a specific effect, say of case facts or precedent, on judicial behavior (e.g., Segal 1984, 1986; Richards and Kritzer 2002; Richards, Smith and Kritzer 2006). In yet other work on judicial behavior, scholars frequently rely on model specifications which

[1]On publication, a series of pre-estimated models of opinion content across different levels of specificity will be made available on the author's Dataverse page.

include dichotomous covariates (fixed effects) for issue areas, seeking to address issue-specific variation. Likewise, and as noted above, specifying the issue at debate serves as the principal means to identify the Court's ideological output (e.g., McGuire et al. 2009) and to identify dimensions of dispute in the case (e.g., Benesh and Spaeth 2007). Finally on judicial behavior, the identification of issues has implications for the rhetorical strategies of the justices, including whether they engage in heresthetical maneuvering (e.g., Rice 2016) or issue creation (e.g., McAtee and McGuire 2007). Moving beyond judicial behavior, measuring issue content also has important implications for cross-institutional studies, including understanding the judicial hierarchy (e.g., Baird 2004, 2007), legal mobilization (e.g., Rice 2014), and the influence of the courts in the policy process (e.g., Flemming, Bohte and Wood 1997; Flemming, Wood and Bohte 1999).

To that end, the identification of the issue continues to generate discussion and disagreement among top scholars in the field, including one recent exchange on the public discussion list of the Law and Courts Section of the American Political Science Association.[2] The SCD provides human-coded public policy and legal issue codes which serve as measures of opinion content. For the public policy issue codes, cases are classified into one of 260 unique codes organized into one of 14 broad categories called issue areas.[3] This issue code is quite explicitly intended to reflect the public policy content of the case; as the coding protocol states, "[t]his variable identifies issues on the basis of the Court's own statements as to what the case is about. The objective is to categorize the case from a public policy standpoint, a perspective that the legal basis for decision (lawType) commonly disregards" (42). Thus, for nearly every case (or vote) of the Court, the issue area code identifies a single policy which the Court (justice) addressed in their opinion; I refer to this as the *unidimensional* assumption.

Like all of the information contained in the SCD, the issue area code for opinion content

[2]Law & Courts List-Serv Digest, August 15-16, 2016.

[3]On rare occasions, the SCD does code multiple topics for a single case.

has been employed extensively in research on the Court. Yet the measure is subject to a number of criticisms. Controversy most frequently stems from the primacy of these codes in understanding ideological motivations in judicial behavior. After identifying the primary public policy dimension under dispute, the codebook outlines a variety of rules for coding the ideological direction of the decision or vote; thus, identification of the policy dimension is primary in identifying the ideological content of the opinion. If the unidimesnional assumption is appropriate – that is, only a single issue is addressed by the court in most cases – then the task of identifying the policy dimension and associated ideological direction should be straightforward. Indeed, the database coding protocol heavily favors this theoretical perspective, assigning only a single issue to nearly every case.

The validity of this assumption, however, is up for debate. As noted above, Shapiro (2009) offers an in-depth critique of this single issue preference and finds little supportive evidence. The presence of multiple issues in opinions classified as pertaining to only a single issue area likewise continues to have ramifications for research on the Court. Harvey and Woodruff (2011) document a trend to code issue areas so that they align with the coder's expectation of how the case was decided on a liberal-conservative dimension. Rice (2016) documents disagreement over policy in dissenting behaviors, which has important ramifications for strategic behavior and legal development. McGuire and Vanberg (2005) attempt to use Wordscores (Laver, Benoit and Garry 2003) – an approach to scaling text in order to identify ideological preferences from speech – but instead recover a number of issue dimensions despite explicitly restricting their analysis to opinions within the SCD's issue areas. As these and other examples show, the preference to assign a single issue has far-reaching ramifications for empirical research on the Court, with those ramifications only likely to grow.

A simple first-cut at examining whether the unidimensional assumption might generate problems is to compare the classifications to alternative coding protocols to see if similar dimensions are recovered. On this front, the Court is unique as the issue content of opinions

is human-coded by the SCD *and* the Policy Agendas [PA] Project, which measures policy attention across American political institutions.[4] For their coding, PA utilizes 19 major topic and 220 minor topic codes as issue categories. The PA issue codes are unlikely to suffer from confirmation bias, as the data explicitly focus solely on policy content and feature none of the ideological variables underlying the confirmation bias story. However, the PA dataset still exclusively assigns only one issue per opinion; thus, the PA codes are not a panacea for the criticisms above. Still, they may provide a valid alternative to the SCD codes if they better reflect the underlying topic structure of opinions. They also offer an opportunity to examine the validity of assuming topical structure is known; if so, scholars should come to similar understandings of the topics given the similarity in the PA and SCD coding protocols. Yet despite the intention of both protocols to reflect policy, they differ profoundly in their classifications. Indeed, while some divergence is inevitable, the scale of this divergence should give researchers pause. As an example, there are at least three opinions classified as "Economic Activity" by SCD in *each and every one* of the 19 different PA issue areas.[5] Such disparities offer support for the concerns expressed in prior research addressing these human-assigned issue codes, and point generally to the difficulties inherent in human-coding enterprises.

Measuring Opinion Content

Of course, one might as well claim the problem lies with the PA classifications and not the SCD. Therefore, I turn now to assessing whether the commonly employed codes offer a valid representation of opinion content. Documenting the validity of human-assigned codes

[4]These data were collected by Frank Baumgartner and Bryan Jones, with the support of National Science Foundation grants 9320922 and 0111611, and are distributed through the Department of Government at the University of Texas.

[5]Similar patterns are observed across issue classifications, as detailed in the supplementary material.

is a daunting, time-consuming, and expensive task. Despite the high costs of coding in the first place, there is little guarantee of reliability as reproducing the classifications duplicates efforts thus introducing additional costs. For example, some scholars have averaged expert codings (Benoit and Laver 2006; Lowe and Benoit 2013), which achieves increased accuracy but only by amplifying costs. In addition to the financial, time, and reliability concerns, one may still also be concerned with the validity of the ex ante determination of codes into which texts are classified (Quinn et al. 2010). By predefining the classification scheme, researchers are occasionally required to adjust the scheme mid-study or to place observations into ill-suited categories. Shapiro's study is instructive; the extensive work completed there amounts to significantly less than 1% of the total cases in the SCD, and the more robust coding scheme proposed would only further compound the costs of coding opinion content.

Instead, methods for computational text analysis – specifically, the intuitions and approaches of supervised learning and cross-validation – provide an avenue to address the validity of the assigned issue area codes and the unidimensional assumption. In supervised learning, a researcher manually codes a subset of data, then partitions the data into training and testing sets. A model is estimated on the training subset, then evaluated by its ability to predict the human-coded labels of the test subset. By learning features which optimize the coding of the test data, researchers may automate the coding of enormous volumes of text at low cost. Options for supervised learning algorithms are extensive (see, for instance, Caruana and Niculescu-Mizil 2006) and accuracy rates are constantly improving.

Notably, supervised learning struggles when the mapping between observable characteristics and assigned classifications is unclear. This occurs in a variety of situations, but is often driven by two related features of human coding. First, most human coding enterprises develop coding schemes appropriate for testing the subject of interest *ex ante* and then employ those schemes in classifying the observations, occasionally though not always updating to incorporate new considerations. This introduces a considerable and exceedingly complex problem for classification, as the mapping may shift across observations if the scheme

changes, or may be inappropriate if the *ex ante* determinations were improper. Second, when the human-assigned codes feature misclassification, the supervised learner is capped from achieving high levels of accuracy. That is, inappropriately classified observations directly influence the weighting of a mapping between observable characteristics and assigned classifications. When observations are "forced" into potentially inappropriate bins the signal for supervised learning is muddled.

In all, supervised learners will identify – within the training subset – observable characteristics which best predict the assigned classifications within that subset. However, if the human-assigned code is inappropriate, the characteristics which prove most predictive within the training subset will not prove predictive in the testing subset *because the signal itself is unclear*. Thus, where assigned issue areas do not sufficiently distinguish between different types of cases based on the actual language used in those cases, supervised learners will fail.

It is this characteristic of supervised learning that provides an avenue for assessing the validity of the SCD's issue area codes. By training a variety of supervised learning models across different observable textual characteristics of the opinions, one can identify whether those characteristics – in fact, *any* textual characterstics – are able to consistently and validly predict the issue area code at standard levels (70 to 90%) of intercoder reliability (Quinn et al. 2010). In the following section, I utilize three different operationalizations of the textual content of opinions to predict the assigned issue areas: word counts, weighted word counts, and n-gram counts. As an additional robustness check, I also document the results across two different but prominent supervised learning tools for classification. In not a single one of the six separate analyses are the issue areas of opinions in the test set predicted at rates which surpass the standard minimum levels of acceptable intercoder reliability.

Predicting Issue Areas Using Opinion Content

To begin, I acquired the texts of all majority opinions in the Court's history from `Justia`, an online repository of legal information.[6] I predict the classifications of opinion issue area as coded by the SCD using each of the above operationalizations of opinion content. *Word counts* reflect precisely that; a count of the number of times the word appears in an opinion. In order to better capture the informational value of particular words, *weighted word counts* are the word counts, weighted by inverse document frequencies, or how frequently the word appears across all documents (Blei, Ng and Jordan 2003). Finally, because isolated words ("rational") may not offer the same informational value as phrases ("rational basis"), I estimate an *n-grams* model that includes both unigrams and bigrams.

A host of algorithms are available for the classification task. Rather than potentially biasing results by selecting only one approach, I instead utilize two of the more common and robust approaches: conditional inference trees (Hothorn, Hornik and Zeileis 2006) and random forests (Breiman 2001). In a standard tree or recursive partitioning approach, the model performs an exhaustive search across all possible splits and chooses the covariate to split the data on based on an information criterion; this process can lead to overfitting (Mingers 1987). The researcher can address overfitting by pruning, but that procedure introduces variable selection bias (Breiman et al. 1984). Instead, both conditional inference trees and random forests limit overfitting concerns endemic to many other classification approaches and minimize the amount of variable selection bias potentially introduced by the researcher. This is particularly important here, as the goal is to use the supervised algorithms to explore the validity of the human-assigned issue area codes. Overfitting in the training set would necessarily decrease the ability to predict issue areas in the test set, and

[6] A number of preprocessing steps were utilized in order to prepare the texts for analysis; all punctuation and capitalization was stripped from the corpus. I retained only tokens which appear in at least 1% but not more than 50% of documents, thereby removing extremely infrequent and extremely frequent words which impinge topic determination and slow computation.

thus identifying classifiers which minimize overfitting is a priority. Moreover, by choosing classifiers which minimize the amount of pruning (and the attendant biases introduced), I gain a relatively unbiased testing ground.

The two procedures provide a mapping between opinion content operationalizations and human-assigned issue codes. In all, I estimate six models – three conditional inference tree models and three random forests models – of the SCD's issue area codes using the respective measures of opinion content – word counts, weighted word counts, and n-grams – as predictors. The models are trained on a random sample of 80% of the opinions, then evaluated on a held-out test set consisting of the remaining 20% of opinions. To the extent the opinion features map to the human codes, classification accuracy rates should be high. I begin with the accuracy of models based on unweighted word counts; here, the random forests classifier correctly predicts 61.8% of issue area classifications while the conditional inference trees classifier correctly predicts just 42.0%. Weighting by term frequency - inverse document frequency only marginally increases classification accuracy, with correct predictions of 63.5% in the random forests implementation and 46.2% in the conditional inference trees implementation. Finally, classifiers employing both unigrams and bigrams returned classification accuracies of 64.2% for the random forests implementation and 44.7% for the conditional inference trees estimation.

[include Figure 1 about here]

Across models and operationalizations of opinion content, we see little evidence the SCD's opinion content coding scheme provides a reliable representation of opinion text. Consider that the standard benchmark for human intercoder reliability is 70 to 90% (Quinn et al. 2010). Here, the SCD model never exceeds even the minimum benchmark across models. The concern, then, becomes identifying the source of the reliability breakdown. In Figure 1, I plot the proportion of opinions in the test set correctly classified for each of the issue areas against the total number of cases within the test set identified as that issue area. Consistently across implementations, the models are able to recover some issue areas (criminal

procedure, economic activity, and federal taxation) at impressive rates. However, the models also consistently perform poorly in other areas. Judicial power opinions, for instance, are correctly identified less than 60% of the time across implementations, despite being the second most frequent option and – by virtue of that frequency – an area which the classifier is more likely to default to. Likewise, relative to their frequency in the data, cases dealing with civil rights, due process, and federalism are all regularly and disproportionately classified incorrectly. This is particularly egregious in the case of due process and federalism opinions, which never break 10% classification accuracy across any of the six models.

Importantly, where the model struggles to recover topics is entirely sensible based on prior criticisms of the SCD's issue area coding scheme, and specifically the unidimensional assumption. In her work, Shapiro (2009) highlights each of judicial power, federalism, and due process as issue area codes that could – and should – have been applied to opinions more frequently, stating of her findings that "[m]ost striking . . . is the extent to which the Database appears to systematically underreport information about government structure and operations, about judicial power, and about lawyering" (519). Yet because the SCD explicitly attempts to identify a single issue area, these issue areas – commonly discussed in addition to a variety of specific policies – are frequently treated identically – that is, as being absent – to completely irrelevant policy considerations.

Thus, the single issue preference introduces two important measurement problems. First, by treating issue area values as generally mutually exclusive, the coding scheme omits large amounts of important information about an opinion which may have discussed – in part – other issue areas. Likewise, treating issue areas as mutually exclusive simultaneously overestimates the attention devoted to the identified issue area, associating the entirety of the textual content with a single issue area. For the supervised classifiers, this presents an intractable problem. By associating opinion content with only a single issue area, the coding scheme both under-associates and over-associates issue area codes with opinion content. The consequences are clear. No matter how one might operationalize opinion text, these

codes are difficult to recover as a function of their unclear connection to the coding scheme. The problem is magnified in issue areas (like judicial power) present across many opinions classified into other issue areas, thus mitigating the amount of useful information a supervised learner can leverage to identify the cases. With overall classification rates systematically falling short of the minimum 70% – and sometimes falling short by large margins – the benefits of taking an alternative approach are clear.

Topic Models for Opinion Content

In light of the above, one option is manually recoding the opinions, adding such issue areas as necessary. However, this would require massive time and resource costs while still requiring dichotomization in coding choices, treating opinions as either completely or not at all about a particular issue area. Instead, the unsupervised estimation of topical content from text offers an avenue to resolve the underlying problems in a flexible, cost-efficient, and unbiased way while also abandoning the unidimensional assumption. Within political science alone, increasingly sophisticated research has been done on topic models to perform unsupervised topic classification (Quinn et al. 2010; Grimmer 2010) with recent efforts at creating more flexible forms which can adopt additional information (Roberts, Stewart and Airoldi 2016). The aim of this research is to uncover the underlying latent topics of a large corpus of texts. Models uncover patterns of word use across documents and connect those documents which exhibit similar patterns (Blei and Lafferty 2009). Note that, unlike human-coded schemes, topics are explicitly determined by the documents' language. Subsequent analysis within one of these topic areas would thus be able to differentiate other dimensions of interest within an issue area, such as ideological direction (Lauderdale and Clark 2014). Further, no bias is introduced by prior knowledge of the opinion's ideological character.

Multiple algorithms are available for unsupervised topic classification, of which I use latent Dirichlet allocation. In LDA, documents are considered as arising from multiple latent topics, an important improvement over many other clustering algorithms which classified

documents into single topics (Blei, Ng and Jordan 2003). Assuming a certain number of topics associated with a set of documents and having observed the words which make up the documents, through LDA we estimate the proportion of each document which exhibits each topic (Blei and Lafferty 2009). Inference is performed on the posterior probabilities through one of multiple possible methods (Blei, Ng and Jordan 2003), frequently Gibbs sampling (Griffiths and Steyvers 2004).[7] The primary value of the approach lies in the estimation of the relative proportion of each document arising from a topic, or topic proportions. In so doing, the approach allows researchers to move past the unidimensional assumption and the dichotomous treatment of topic attention to more accurate reflections of the Court's attention.

Though LDA requires very few predefined constraints, it does require prior selection of values for α – the Dirichlet hyperparameter for topic proportions – and k – the number of topics – before estimation. For α, I utilize the value of $\frac{50}{k}$ consistent with Steyvers and Griffiths (2007). The number of topics k is a more difficult question; as Blei and Lafferty (2009) state, "[c]hoosing the number of topics is a persistent problem in topic modeling" (11). The choice of k offers another valuable tool for researchers of the Court, as they can adjust the parameter in order to obtain more general or more specific estimates of opinion attention. Rather than settling on a single k here, I instead extract latent topic proportions across multiple specifications, with k set to 25, 50, 75, and 100. Because the choice of hyperparameters may influence results (Wallach, Mimno and McCallum 2009), this variation across values of α and k offers an additional robustness check on the patterns identified in this section and the substantive conclusions of subsequent analyses. Finally, I create topic names by concatenating the five most likely terms by topic.

[include Table 1 about here]

Measurement models are assessed by a number of validity criteria. As evidence of the

[7]The model is implemented in `R` through the `text2vec` package.

face validity of the measure as well as convergent (e.g., does the measure map to other measures which it should map to) and discriminant (e.g., does the measure diverge from other measures in useful ways) validity, I look here to topics identified through the 25 topic LDA. I present in Table 1 the top cases for four topics. Note first the opinions are associated with sensible topics. Moreover, the second column in the table indicates the assigned SCD issue area code. The issue areas assigned by the SCD are generally consistent within topic areas, providing evidence the estimated models generate topics sensible to human readers and coherently associated with concepts with which they should be associated (convergent validity).

Areas where the mapping diverges, however, illustrates the value of moving away from the dichotomous and restrictive SCD coding to alternative approaches. Consider first, the SCD's classification of *Wiggins v. People in Utah* (1876) as disposed on the grounds of Judicial Power, specifically jurisdiction. The entirety of the majority opinion's discussion of jurisdiction is as follows:

> Sec. 3 of the Act of Congress of June 23, 1874, 18 Stat. 254, allows a writ of error from this Court to the Supreme Court of the Territory of Utah, where the defendant has been convicted of bigamy or polygamy or has been sentenced to death for any crime. The present writ is brought under that statute to obtain a review of a sentence of death against plaintiff in error for the murder of John Kramer, commonly called Dutch John, in Salt Lake City.[8]

The remaining six pages of the majority opinion, and the entirety of Justice Clifford's dissenting opinion, focused on a description of the homicide at the heart of the case and the admissibility of witness testimony to the jury. Thus, though the Court certainly considered and was required to address the issue of jurisdiction – as they must in any case – the overwhelming focus of *Wiggins* was on a separate issue that the Court also addressed.

Next, consider prominent cases within the "amendment, school, constitutional, political, members" topic. Here, cases fall in both the SCD's First Amendment and Civil Rights issue areas. Again, the cost of assigning a single topic becomes clear. Here, the topic

[8] 93 U.S. at 466.

captures debate surrounding civil rights and civil liberties issues. While general, recall that the model is restricted to the 25 topics which best describe the distribution of terms for the entire history of the Court. With changes over time in the focus of the Court, operating at this level of generality inevitably leads to compartmentalization. Yet this can be easily relieved for scholars interested in doing so by expanding the size of the model; increasing the numbers of topics permits greater specificity. Moreover, while the SCD classifies each into one of these categories, four of the five cases – *School District of Grand Rapids v. Ball*, *Munro v. Socialist Workers*, *Eu v. San Francisco County Democratic Central Committee*, and *Jenness v. Fortson* – address both First and Fourteenth Amendment arguments.

[include Figure 2 about here]

In all, from this micro-level view of particular cases we see stark evidence of the value of considering latent topic structure as reflected in the words actually used in the opinions. Stepping back, the value of the approach is further evident in analyzing correlated topics. To demonstrate, I plot in Figure 2 a visualization of the correlation structure among the estimated topics. In the plot, darker shades indicate increasingly negatively correlated topics, while lighter shades indicate increasingly positively correlated topics. The topics are ordered according to hierarchical clustering of the resulting correlations, with boxes indicating the identified clusters.[9] Each of the clusters offers evidence of a coherent area of policy attention on the Court; consider, for instance, the cluster clearly pertaining to criminal justice contains"amendment, police, search, warrant, officers", "testimony, i, you, counsel, witness", and "criminal, offense, petitioner, sentence, conviction" topics. That the topic proportions for theoretically connected concepts correlate across the corpus provides yet further evidence of the value in considering opinions as reflective of more than a single, exclusive subject.

[9]Clusters are estimated using the Lance-Williams dissimilarity update formula for agglomerative hierarchical clustering. For more information, see Lance and Williams (1966, 1967) or Murtagh and Legendre (2014).

Topic Concentration and Accuracy

The presence of multiple topics has important implications for research on such fundamental topics as legal development, strategic judicial behavior, and the agenda-setting influence of courts; on these and other subjects, empirical research is systematically excluding large swaths of relevant information. Directly mapping the estimated topic features to the SCD issue area codes provides further evidence of the importance of considering issue attention proportionally, rather than exclusively. As noted earlier, supervised learning offers an avenue by which to assess the validity of issue area codes in capturing opinion content. Therefore, I treat the latent topic proportions for each document as features for the same classification procedures as outlined above, which is equivalent to using LDA as a "fast filtering algorithm for feature selection in text classification" (Blei, Ng and Jordan 2003, 1013). For each model, I utilize the topic proportions to predict the issue area classifications. The results appear in Figure 3.

Once again, classification accuracy falls below standard levels of intercoder reliability, with accuracy rates between 60.7% (25 topic model) and 63.8% (100 topic model). Further, the models again struggle to recover cases classified as Judicial Power at rates commensurate with their size in the corpus, and consistent with prior results the models perform particularly poorly for the Federalism and Due Process issue areas. In all, precisely the same dynamics are observed when utilizing topic proportions as any other operationalization of opinion content in recovering the issues identified by the SCD as singularly prominent in each opinion.

[include Figure 3 about here]

Yet whereas earlier analyses stopped here, further examination of the topic model results provides yet more conclusive evidence of the problems introduced by the unidimensional assumption. Specifically, consider the underlying problem with classification of judicial power cases. As the *Wiggins* example illustrates and as Shapiro has highlighted, the Court must regularly address and decide these considerations in disposing of a case. In this example,

the choice of identifying a case as falling within Judicial Power is particularly onerous. Yet considered more broadly, *any* opinion which addresses multiple topics – rather than being concentrated on a single topic – is likely to be more difficult to classify into a single, exclusive issue area.

To test whether this is the case – that is, whether topic concentration indeed influences classification accuracy – I turn to analyzing the topic proportions across classification successes and failures. As a measure of topic concentration for each opinion, I calculate a normalized Herfindahl - Hirschman index (H-index) of the latent topic proportions recovered from LDA equal to:

$$C_d^* = \frac{(C_d - 1/N)}{1 - 1/N} \tag{1}$$

where $C_d = \sum_{i=1}^{N} \hat{\pi}_i^2$ is the standard H-index, N is the number of topics, and $\hat{\pi}_i$ is the proportion of a topic captured in a given document. Higher values indicate documents more concentrated on a single topic, whereas lower values indicate documents which address many different topics. I then average the document-specific H-indices separately for all correctly classified documents and for all incorrectly classified documents, with classification results pulled from the above random forests classifier.

[include Figure 4 about here]

The fundamental question is whether the misclassified opinions are generally those which address multiple issues, such that identifying a *single* issue as dispositive leads to lower rates of classification accuracy. To that end, I plot the average H-indices of cases correctly predicted (solid line) and incorrectly predicted (dashed line) across values of k, the number of topics, in Fig 4. The results strongly support the hypothesis that cases addressing multiple issues are more difficult to classify. Across different values for the number of topics, Figure 4 demonstrates a consistent difference in the average H-indices of cases correctly and incorrectly classified, with incorrectly classified cases on average always significantly less concentrated

than majority opinions. Whether considered at higher levels of aggregation (25 topics) or in more granular specifications (100 topics), opinions which address multiple topics – that is, opinions which do not fall neatly into singular dimensions as defined by the actual words used in the opinions – are more difficult to classify into an issue area.

[include Figure 5 about here]

Moreover, this dynamic persists across the SCD's assigned issue areas. To see this, in Figure 5, I plot the average H-index for correctly classified (centered at the issue area names) and incorrectly classified (at the arrow point) opinions for the 25 topic model.[10] With the exception of an extremely small increase in the rarely assigned Attorneys issue area, demonstrable decreases are evident in topic concentration when going from correctly classified to incorrectly classified opinions. Again, consider that many of these assigned issue areas are likely cross-cutting but are treated in the overwhelming majority of cases as mutually exclusive; for example, a case classified into the First Amendment issue area inevitably must address issues classified under Judicial Power, as well as potentially issues classified into Due Process, Federalism, and often other areas. Should these topics be only briefly discussed or should the case rest neatly on First Amendment grounds with minimal need for discussing issues under Judicial Power, the case may be appropriately classified. But – as the results throughout this paper make clear – this is likely to rarely be the case.

Discussion and Conclusion

The Supreme Court Database is omnipresent is empirical research on law and courts, and justifiably so. The rich and careful coding of such voluminous primary materials has been transformative in the study of law and politics. The classification of opinion content, however, offers one area where scholars increasingly recognize potential threats to empirical research. In this article, I address the reductivist unidimensional assumption underlying

[10]Results are consistent across specifications.

the classifications. Though the SCD approach was necessary for purposes of creating a tractable human-coding enterprise, I document the difficulty it has in accurately capturing opinion content. Specifically, across a variety of operationalizations of the actual content of opinions – that is, the words used by the justices to describe the issues they are deciding – I am unable to recover the issue area codes at minimally acceptable rates. In it's stead, I introduce a series of measures based on a now well-accepted approach from computational linguistics for capturing textual content. These topic models offer marked improvements over existing measures. First, they overcome the unidimensional assumption by instead providing estimates of the proportion of each document addressing each topic. In so doing, the approach better reflects opinion content in ways consistent with the movement of the field toward the strategic model and more complex assessments of judicial behavior. Moreover, they offer a flexible approach for scholars interested in studying the Court's attention, as they can be simply adjusted and estimated in order to yield more general topical frameworks or highly specific frameworks.

It bears emphasizing the importance of moving away from the current reliance on the SCD's indicators of opinion content. With social scientists and legal scholars increasingly interested in leveraging tools for the computational analysis of text in order to extract increasingly sophisticated dimensions of interest (i.e., ideology, emotional valence, or even personality types), learning from other areas of research is imperative. Indeed, work on scaling ideology from text explicitly notes the importance of restricting the analysis to singular issue frames (Laver, Benoit and Garry 2003). As prior work demonstrates, relying on the SCD's measures of opinion content for these purposes is problematic. Therefore, identifying alternative approaches that are theoretically and empirically valid while also being tractable solutions is imperative.

Moving forward, research might look to build on the approach identified here to construct increasingly sophisticated indicators of the content of legal documents. The approach utilized here should be seen as an important first step in this process, but the effort to identify

and measure opinion content should certainly continue. Legal texts are – as legal scholars have frequently reminded social scientists (Friedman 2006) – much more complex than often presumed in empirical research. Alternative methods for estimating topics from text corpora and minor improvements and adjustments to existing approaches continue to be rapidly introduced across a variety of technical and applied fields of research, and as these tools become established it may become clear they offer yet further improvements on that which I have introduced in this article. I leave this to future research. At present, even the standard, well-established approaches for topic estimation offer vast improvement on current practices.

References

Baird, Vanessa. 2004. "The Effect of Politically Salient Decisions on the U.S. Supreme Court's Agenda." *The Journal of Politics* 66(3):755–772.

Baird, Vanessa. 2007. *Answering the Call of the Court: How Justices and Litigants Set the Supreme Court Agenda*. Charlottesville: University of Virginia Press.

Benesh, Sara and Harold Spaeth. 2007. "The Constraint of Law: A Study of Supreme Court Dissensus." *American Politics Research* 35(5):755–768.

Benoit, Kenneth and Michael Laver. 2006. *Party Policy in Modern Democracies*. London: Routledge.

Black, Ryan, Rachel Schutte and Timothy Johnson. 2013. "Trying to Get What You Want: Heresthetical Maneuvering and U.S. Supreme Court Decision Making." *Political Research Quarterly* 66:819–830.

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

Blei, David and John Lafferty. 2009. Topic Models. In *Text Mining: Classification, Clustering, and Applications*, ed. A Srivastava and M Sahami. Chapman and Hall/ CRC Press.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Breiman, Leo, Jerome Friedman, Charles Stone and R.A. Olshen. 1984. *Classification and Regression Trees*. Florida: Chapman and Hall/ CRC Press.

Caruana, Rich and Alexandru Niculescu-Mizil. 2006. An Empirical Comparison of Supervised Learning Algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*.

Epstein, Lee, Jack Knight and Andrew Martin. 2003. "The Political (Science) Context of Judging." *St. Louis University Law Journal* 47:783–817.

Epstein, Lee and Olga Shvetsova. 2002. "Heresthetical Maneuvering on the U.S. Supreme Court." *Journal of Theoretical Politics* 14(1):93–122.

Evans, Michael, Wayne McIntosh, Jimmy Lin and Cynthia Cates. 2007. "Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research." *Journal of Empirical Legal Studies* 4:1007–1039.

Flemming, Roy, B. Dan Wood and John Bohte. 1999. "Attention to Issues in a System of Separated Powers: The Macro-Dynamics of American Policy Agendas." *Journal of Politics* 61(1):76–108.

Flemming, Roy, John Bohte and B. Dan Wood. 1997. "One Voice Among Many: The Supreme Court's Influence on Attentiveness to Issues in the United States." *American Journal of Political Science* 41(4):1224–1250.

Friedman, Barry. 2006. "Taking Law Seriously." *Perspectives on Politics* 4(2):261–276.

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* 18:1–35.

Grun, Bettina and Kurt Hornik. 2011. "`topicmodels`: An R package for Fitting Topic Models." *Journal of Statistical Software* 40(13).

Harvey, Anna and Michael Woodruff. 2011. "Confirmation Bias in the United States Supreme Court Judicial Database." *The Journal of Law, Economics, and Organization* .

Hillard, Dustin, Stephen Purpura and John Wilkerson. 2008. "Computer Assisted Topic Classification for Mixed Methods Social Science Research." *Journal of Information Technology and Politics* 4:31–46.

Hothorn, Torsten, Kurt Hornik and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15(3):651–674.

Lance, G and W Williams. 1966. "A Generalized Sorting Strategy for Computer Classifications." *Nature* pp. 212–218.

Lance, G and W Williams. 1967. "A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems." *Computer Journal* 9:373–380.

Lauderdale, Benjamin E. and Tom S. Clark. 2014. "Scaling Politically Meaningful Dimensions Using Texts and Votes." *American Journal of Political Science* pp. n/a–n/a.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 47:215–233.

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.

McAtee, Andrea and Kevin McGuire. 2007. "Lawyers, justices, and issue salience: When and how do legal arguments affect the US Supreme Court?" *Law & Society Review* 41:259–278.

McGuire, Kevin and Georg Vanberg. 2005. "Mapping the Policies of the U.S. Supreme Court: Data, Opinions, and Constitutional Law." Prepared for delivery at the Annual Meeting of the American Political Science Association, Washington, D.C., September 1-5.

McGuire, Kevin, Georg Vanberg, Charles Smith and Gregory Caldeira. 2009. "Measuring policy content on the U.S. Supreme Court." *Journal of Politics* 71(4):1305–1321.

Mingers, John. 1987. "Expert Systems – Rule Induction with Statistical Data." *Journal of the Operations Research Society* 38(1):39–47.

Murtagh, Fionn and Pierre Legendre. 2014. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" *Journal of Classification* 31:274–295.

Phan, Xuan-Hieu, Le-Minh Nguyen and Susumu Horiguchi. 2008. Learning to Classify Short and Sparse Text and Web with Hidden Topics from Large-Scale Data Collections. In *17th International World Wide Web Conference*. Beijing, China: pp. 91–100.

Quinn, Kevin, Burt Monroe, Michael Crespin, Michael Colaresi and Dragomir Radev. 2010. "How to Analyze Political Attention With Minimal Assumptions and Costs." *American Journal of Political Science* 54:209–228.

Rice, Douglas. 2014. "The Impact of Supreme Court Activity on the Judicial Agenda." *Law & Society Review* 48(1):63–90.

Rice, Douglas. 2016. "Issue Divisions and U.S. Supreme Court Decision Making." *Journal of Politics* .

Richards, Mark and Herbert Kritzer. 2002. "Jurisprudential Regimes in Supreme Court Decision Making." *American Political Science Review* 96:305–320.

Richards, Mark, Joseph Smith and Herbert Kritzer. 2006. "Does Chevron Matter?" *Law & Policy* 28:444–469.

Roberts, Margaret, Brandon Stewart and Edoardo Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111(515):988–1003.

Segal, Jeffrey. 1984. "Predicting Supreme Court Cases Probabilistically: The Search-and-Seizure Cases, 1962-1981." *American Political Science Review* 78(4):891–900.

Segal, Jeffrey. 1986. "Supreme Court Justices as Human Decision-Makers: An Individual Level Analysis of the Search-and-Seizure Cases." *American Journal of Political Science* 48:938–955.

Shapiro, Carolyn. 2009. "Coding Complexity: Bringing Law to the Empirical Analysis of the Supreme Court." *Hastings Law Journal* 60:477–540.

Steyvers, Mark and Tom Griffiths. 2007. Probabilistic Topic Models. In *Latent Semantic Analysis: A Road to Meaning*, ed. T. Landauer, D. McNamara, S. Dennis and W. Kinstch. Lawrence Erlbaum.

Szmer, John and Barry Edwards. 2011. "Future Directions in Data Analysis and Collection." *Law and Courts Newsletter* 21:16–19.

Wallach, Hanna, David Mimno and Andrew McCallum. 2009. Rethinking LDA: Why Priors Matter. In *NIPS*.

Wedeking, Justin. 2010. "Supreme Court Litigants and Strategic Framing." *American Journal of Political Science* 54(3):617–631.

Topic: "testimony, i, you, counsel, witness"	
Leyra v. Denno (1953)	Criminal Procedure
Johnson Alias Overton v. U.S. (1894)	Criminal Procedure
Stevenson v. U.S. (1895)	Criminal Procedure
Chapman v. California (1966)	Criminal Procedure
Wiggins v. People in Utah (1876)	Judicial Power
Topic: "employees, labor, board, union, employer"	
NLRB v. Wooster Division of Borg-Warner (1957)	Unions
Medo Photo Supply v. NLRB (1943)	Unions
NLRB v. Express Publishing (1940)	Unions
NLRB v. Pacific Greyhound Lines (1937)	Unions
H.J. Heinz v. NLRB (1940)	Unions
Topic: "income, value, taxes, amount, taxation"	
U.S. v. Pleasants (1938)	Federal Taxation
Helvering v. Bliss (1934)	Federal Taxation
Helvering v. Morgan's, Inc. (1934)	Federal Taxation
U.S. v. Anderson (1925)	Federal Taxation
American Hide & Leather v. U.S. (1931)	Federal Taxation
Topic: "amendment, school, constitutional, political, members"	
School District of Grand Rapids v. Ball (1984)	First Amendment
Munro v. Socialist Workers (1986)	Civil Rights
Eu v. San Francisco County Democratic Central Committee (1988)	First Amendment
Maryland Committee for Fair Representation v. Tawes (1963)	Civil Rights
Jenness v. Fortson (1970)	Civil Rights

Table 1: *Top Opinions for Selected Topics from 25 Topic LDA.* Top opinions are those with the highest topic proportion for the identified topic. The right column provides the opinion's SCD issue area code.

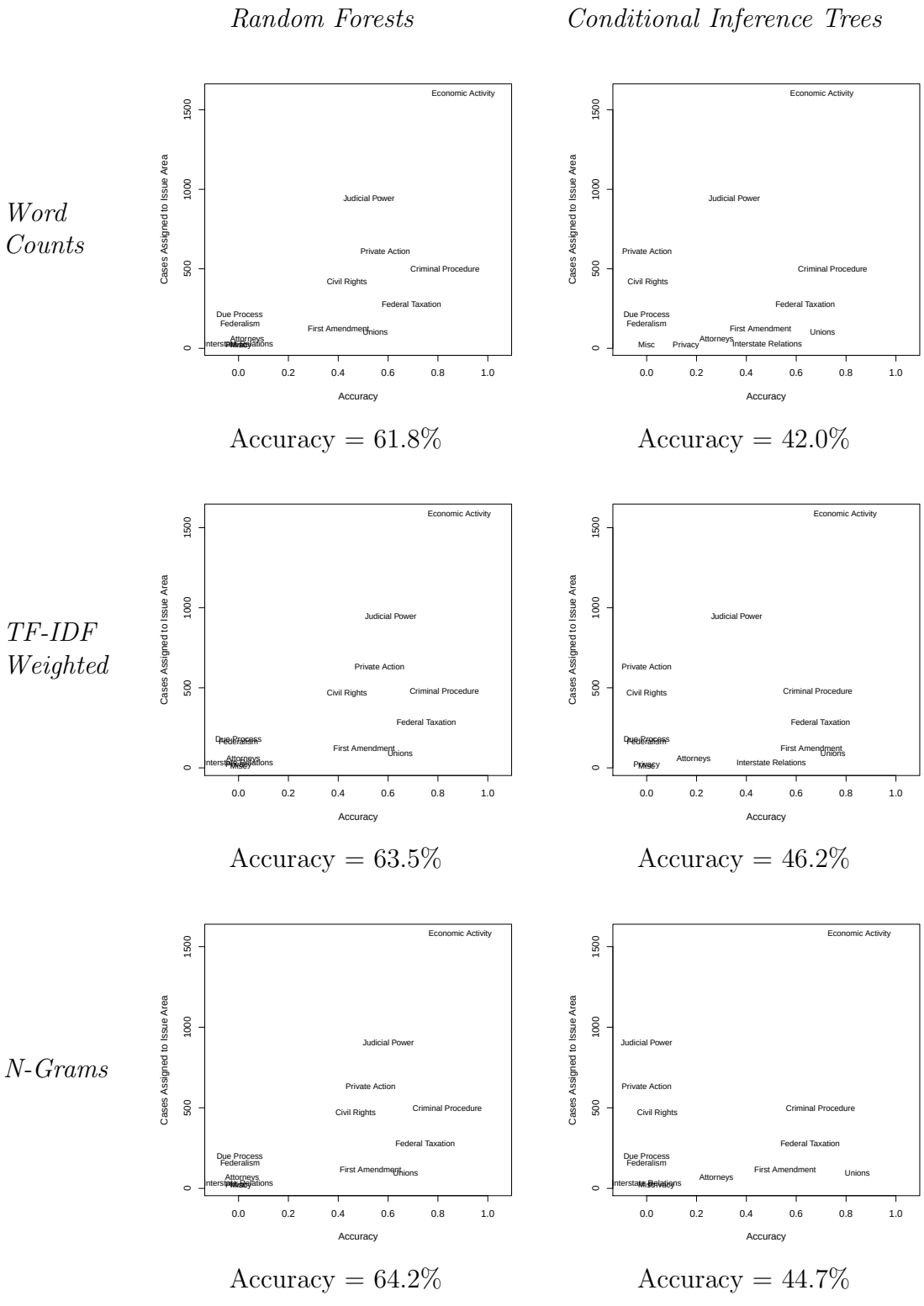

Figure 1: *Scatterplots of classification accuracy and total number of observations within each SCD issue area.* The proportion of cases correctly predicted (x-axis) is plotted against the total number of opinions classified into that issue area (x-axis).

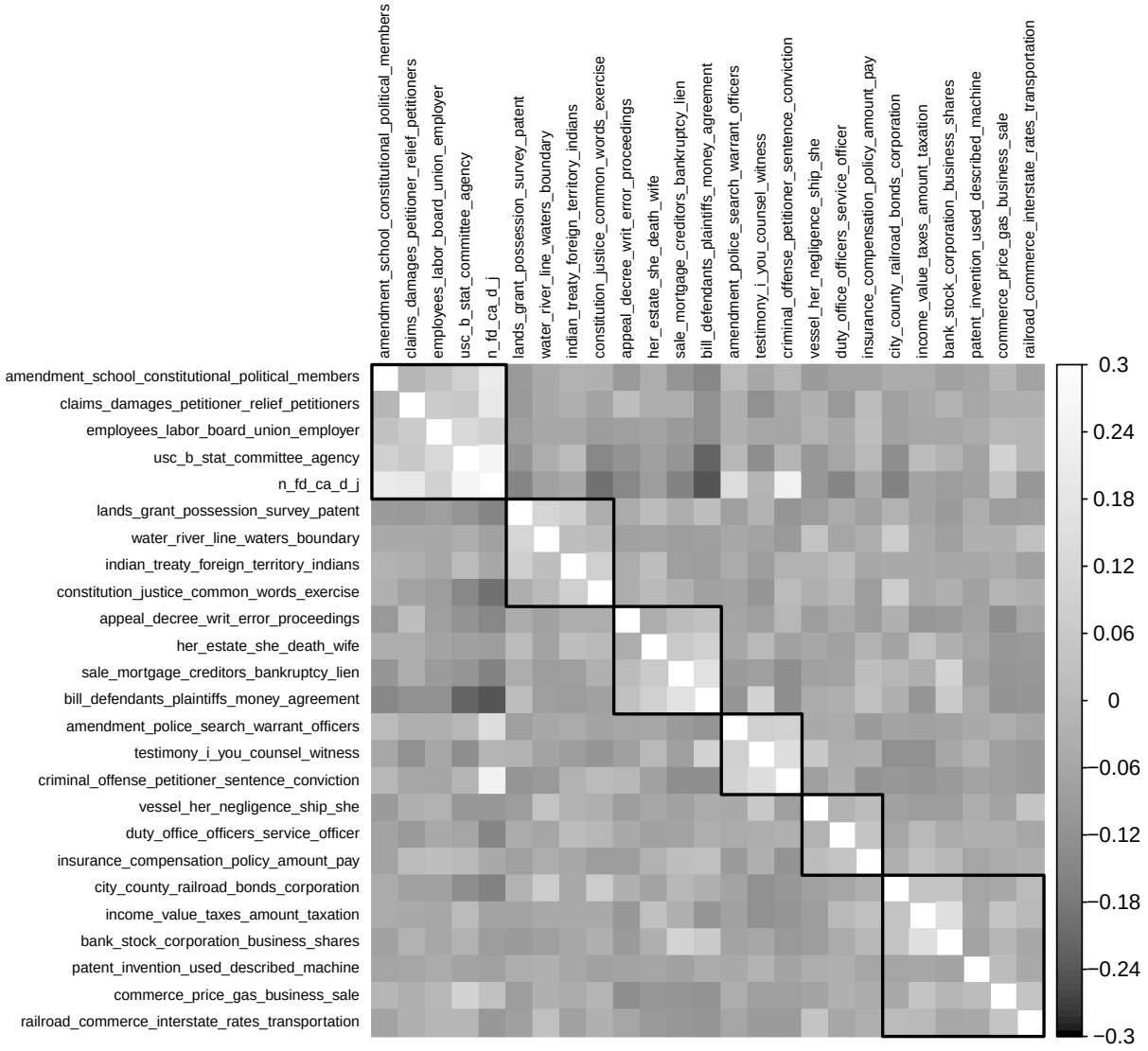

Figure 2: *Heatmap of Correlations by Assigned Document Topic Proportions.* Plot of topic proportion correlations for a 25 topic model, with darker shades indicating negative correlation and lighter shades indicating positive correlation. Boxes indicate clusters as determined by hierarchical clustering of dissimilarities calculated using the Lance-Williams dissimilarity update formula. See text for details.).

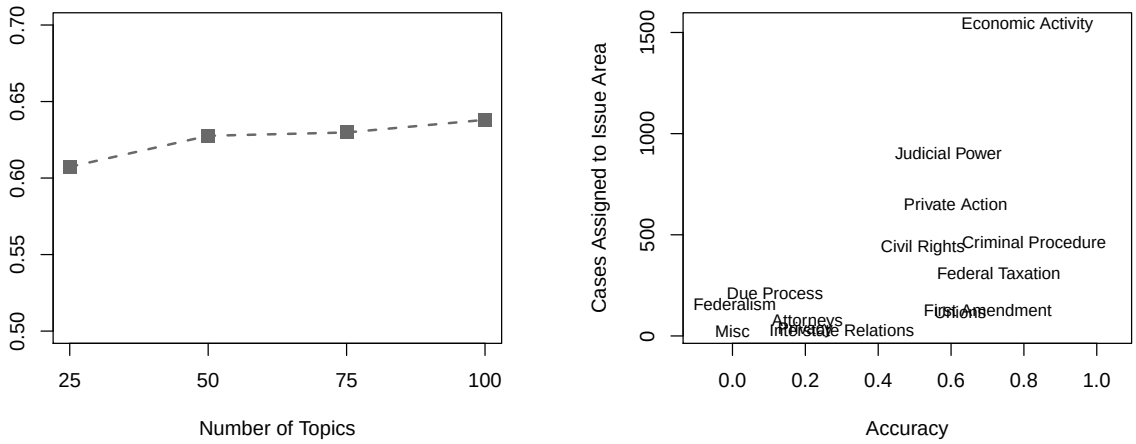

Figure 3: *Latent Topic Proportions as Predictors of SCD Issue Areas.* The left plot indicates the proportion of cases correctly predicted (y-axis) across variations in the number of topics estimated in the topic model. The right plot provides the proportion of cases correctly predicted (x-axis) plotted against the total number of opinions classified into that issue area (y-axis) for the 100 topic model.

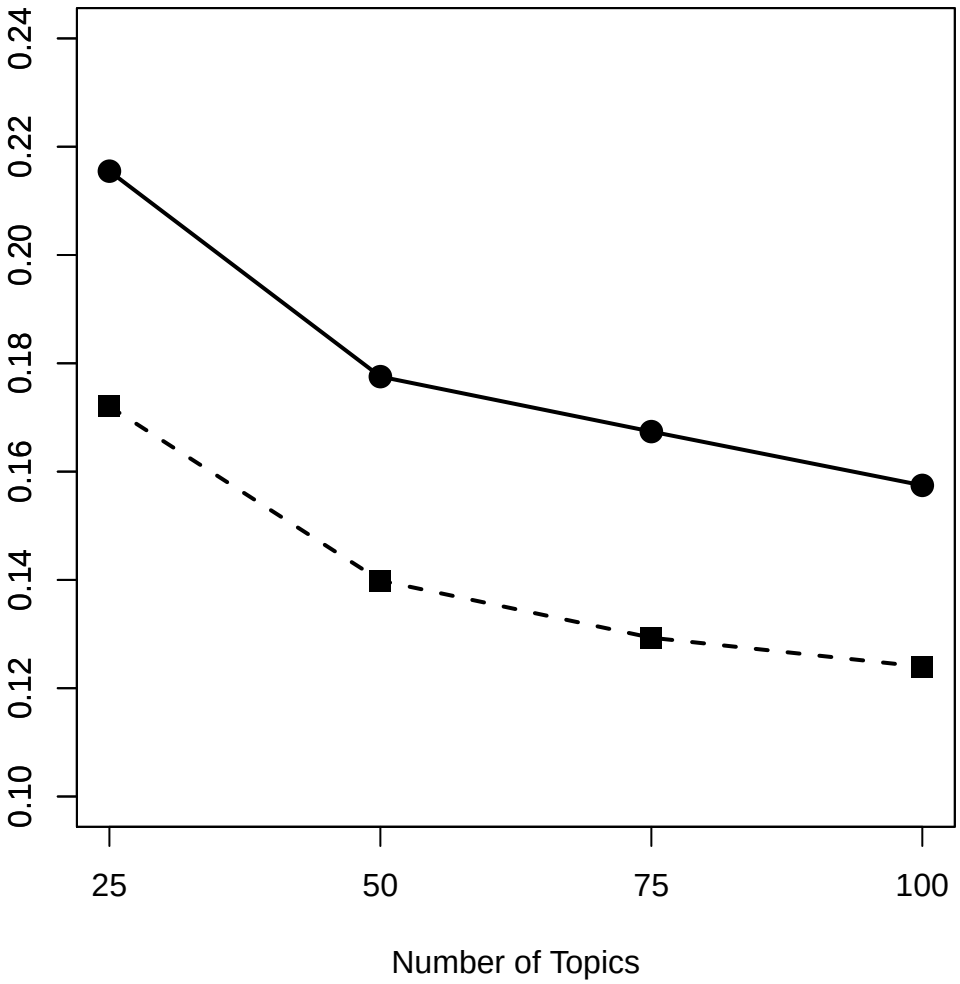

Figure 4: *Plot of Herfindahl Index Over Number of Topics By Prediction Success.* The average Herfindahl index (y-axis) for cases correctly predicted (solid line) is plotted against the average Herfindahl index for cases incorrectly predicted (dashed line) across the number of topics (x-axis).

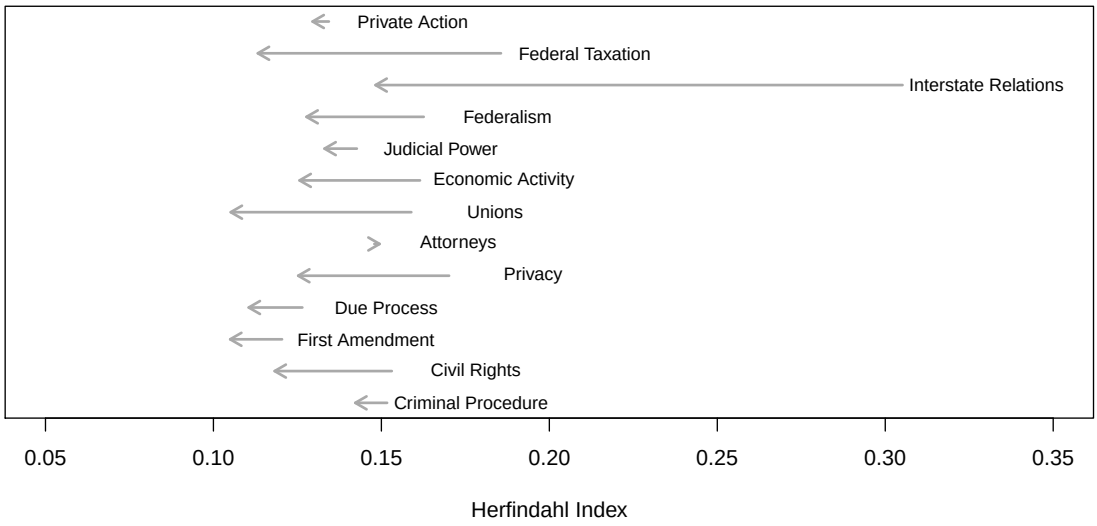

Figure 5: *Plot of Difference in Herfindahl Index Between Correct and Incorrect Predictions by Issue Area.* The average Herfindahl index (x-axis) for cases correctly predicted (indicated by the issue area name) compared to the the average Herfindahl index for cases incorrectly predicted (point of the arrow).

Supporting Information

A Database Overlap

As discussed in the article, human coding carries with it the implicit assumption that classification categories are known ex ante. For purposes of supervised learning, this is a critical assumption; the predefined classification scheme must accurately represent the content of observations, or the learner will struggle to identify features of that text with which to classify observations. One may look to the overlap between similar datasets in order to gauge the validity of this assumption. To the extent the classification schemes are known ex ante, we would expect for similar data collection enterprises to arrive at similar conclusions. Because the PA project and the SCD both prioritize classifying the public policy content of the opinion, there is little reason to expect vast differences in classification. To that end, Table 2 offers a cross-tabulation of opinion classifications for the two datasets.

Yet immediately evident in the table is significant variance across issue areas in classifications. While some variation might be natural, the scope of the variation here is well beyond what might be expected. Take first a point discussed in the main text. For opinions classified by the SCD as dealing with "economic" issues, at least 3 are classified into each of the 19 major policy areas identified by the PA project. Similarly, the PA project topic of "banking & finance" features at least one case in every one of the 13 major issue areas of the SCD classification scheme. Such patterns are manifest throughout the table.

	Supreme Court Database												
Policy Agendas	Criminal Proc.	Civil Rights	First Amdt.	Due Proc.	Privacy	Attys.	Unions	Economic	Judicial Power	Federalism	Interstate Relations	Taxation	Misc
Agriculture	0	2	0	3	0	0	0	20	15	4	0	2	0
Banking & Finance	43	7	6	20	2	8	6	465	117	22	1	73	2
Civil Rights & Liberties	80	337	311	12	53	19	9	10	118	5	0	3	0
Defense	10	61	26	5	3	1	0	11	13	5	0	2	0
Education	2	21	16	8	0	4	2	3	8	5	0	0	0
Energy	0	3	2	6	0	1	0	68	24	16	0	8	0
Environment	3	5	8	6	1	2	0	39	9	11	0	0	0
Foreign Trade	5	0	1	3	0	0	0	33	18	3	0	3	0
Govt Operations	70	96	60	11	6	1	3	43	49	18	2	106	5
Health	6	12	3	11	2	0	2	30	24	10	0	1	0
Housing	3	9	2	6	0	1	0	9	5	0	0	0	0
Intl Affairs	0	4	7	2	0	0	1	3	5	0	0	0	0
Labor, Emp. & Immig.	28	131	23	13	1	6	255	202	91	79	0	7	2
Law and Crime	1126	209	49	119	9	27	6	80	257	26	6	9	2
Macroeconomic	5	2	2	2	0	0	0	56	7	9	0	28	1
Public Lands	3	84	2	19	0	0	0	39	20	14	4	2	1
Science, Tech & Comm.	4	1	25	2	1	0	0	19	10	7	0	1	0
Transportation	14	4	4	13	1	0	2	112	52	19	0	4	1
Welfare	4	59	0	4	0	4	0	3	12	0	0	2	0

Table 2: *Comparison of Coding for Overlapping Opinions in the SCDB and PA Database.*