

# AGENDA FORMATION AND THE U.S. SUPREME COURT: A TOPIC MODEL APPROACH

Michael A. Livermore      Allen B. Riddell      Daniel Rockmore\*

February 29, 2016

Introduction.....	3
I. Agenda formation and text analysis .....	7
A. A simple model of agenda formation.....	7
B. Prior literature .....	11
C. Topic models.....	18
II. Applying the Topic Model .....	23
A. Data .....	23
B. Fitting a topic model to the legal corpus .....	25
C. Testing the topic model .....	27
III. Differences and trends .....	33
A. Certiorari and topic winnowing .....	33
B. Growing distinctiveness of Supreme Court decisions.....	36
IV. “Distant Reading” the Law .....	39
A. The Increasingly “Uncourtlike” Supreme Court.....	39
B. Topic models and empirical legal studies. ....	41
References.....	46
Tables and Figures.....	51

---

\* Michael A. Livermore is an Associate Professor at the University of Virginia School of Law. Allen B. Riddell is a postdoctoral fellow at the Neukom Institute for Computational Science and the Leslie Center for the Humanities at Dartmouth College. Daniel Rockmore is the William H. Neukom 1964 Professor of Computational Science and Professor of Mathematics and Computer Science, Dartmouth College; and Member, External Faculty, The Santa Fe Institute.

## **Abstract**

This paper exploits a relatively new approach to quantitative text analysis—topic modeling—to examine the subject matter of Supreme Court decisions, and in particular to analyze how the semantic content produced by the Court differs from the published decisions of the U.S. Appellate Courts. To conduct this analysis, we fit a topic model to the joint corpus of decisions (Supreme Court plus Appellate Court). The topic model enables a quantitative measurement of differences in semantic content between three corpora: the Supreme Court decisions, the Appellate Court decisions, and Appellate Court cases selected for review. We develop new methods to estimate these differences over time. We reach two findings. First, the Supreme Court has become substantially more semantically idiosyncratic in recent decades, as measured by the use of the topic distribution within a decision as a predictor of the authoring court. We then examine potential causes of this trend, isolating the use of the Court’s case selection power. We find that the topic model based measure of semantic difference between the cases selected for review by the Court does not appear to be increasing over time, indicating instead that the Court has become more distinctive in how it discusses a similarly distinct pool of cases. Normative implications and avenues for future research are discussed. This work demonstrates the utility of topic modeling as a valuable supplement to and/or replacement of hand-coded labels in the study of hierarchically arranged judiciaries. While this case study focuses on the U. S. Courts, extensions and broadening to other national and international judicial corpora can be readily accomplished. More generally, this work opens the door for broader application of topic models within empirical legal studies and related disciplines to study the rich textual corpora generated by legal institutions.

## Introduction

Studying the text of appellate court decisions is a mainstay of traditional legal scholarship. In the mid- and late twentieth centuries, scholars in political science and law added to the methodological toolkit by hand-generating substantial amounts of meta-data concerning the content of these cases, which enabled the use of traditional statistical tools to study judicial decisions (Songer 1998; Spriggs and Hansford 2000). This paper engages in a natural next methodological step of applying statistical techniques directly to the text of court decisions, a move that dovetails usefully with the growing availability of legal materials in machine-readable form and the increasing sophistication of computational text analysis. In particular, we exploit the technique of *topic modeling*, an approach to text analysis that is now a decade old and has spread to a number of academic disciplines including history and political science (Blei 2003; Block and Newman 2011, Quinn et al 2010). The subject of our analysis is the semantic relationship between the U.S. Supreme Court and the U.S. Appellate Courts, how that relationship is affected by the Court's use of its certiorari case selection power, and how that relationship has changed over time.

Our findings are provocative. First, the Supreme Court has become more semantically distinct over the past several decades. Stated simply, our finding is that it is easier today than it was in 1950 to distinguish a Supreme Court opinion from an appellate opinion written in the same year, based on the words in that opinion. We then examine whether the certiorari power helps explain this growing semantic idiosyncrasy. We find no evidence for this proposition, indicating that the Supreme Court today appears to write about cases that are no more distinct (textually) from the body of Appellate Court decisions from which it chooses than the cases that it picked and discussed six decades ago. Together, these two findings can fairly be read to indicate that instead, the Supreme Court has become more distinct from other federal courts in how it analyzes and discusses the legal issues before it—in a way, becoming less “court-like” over time.

These findings raise interesting normative implications, which we discuss, but we spend the bulk of the paper explaining and supporting our analysis. We begin by providing a simplified model for the process that determines the semantic content of U.S. Supreme Court decisions. That model includes the production of U.S. Appellate Court decisions, the certiorari process, and the writing of the decisions of the Court. Because we are primarily interested in consequences, we offer a highly reduced and abstract representation of what is, in reality, a complex process. We then provide a brief summary of the relevant literature, with an emphasis on quantitative analysis of certiorari practices on the Court. We also provide a background discussion of topic models, describing some of the essential features of this new statistical technique and discussing its spread over the past decade within the social sciences and humanities. We then provide some examples to give a sense of the data that is generated when a topic model is applied to a set of judicial decisions.

After providing information on our data and some summary statistics, we describe the results of several analyses to test the validity of our application of a *latent Dirichlet allocation (LDA)* model to the joint corpus of Supreme Court and Appellate Court cases. There are a number of topic models available (and new models are continually being created), but the LDA approach is the most common and there are publically available packages in the coding languages R and Python. Testing topic models for validity remains a difficult exercise because the ideal validation would require manual coding (along with assessment of intercoder reliability) of a vast number of texts. In the absence of such exhaustive expert annotations, a number of ersatz procedures are available. After discussing the options, we take two approaches, one highly informal, and the other more formal. First we use a common informal technique, which is to examine the topic distributions that are generated by the model to determine whether they conform to intuitive subject matter categories. This approach has the advantage of using human natural language processing capacities (which are formidable) and the specialized knowledge of the authors (*caveat emptor*) but it risks confirmation bias and other defects of human rationality. For formal tests, we use two general techniques. First, we examine whether the topic model maintained a reasonable degree of

semantic fidelity with the original documents. We find a loss of accuracy of roughly 5%, a meaningful difference to be sure, but a relatively small reduction in information given the radical reduction in dimensionality when reducing a vocabulary of 21,695 words to 100 topics (a 99.5% reduction). We then test three naïve hypotheses about the relationship between Supreme Court cases and Appellate Court cases using the topic model data, under the assumption that if we could not reject these hypotheses, the topic model was failing to capture important semantic features of our data. We reject all three with a high degree of confidence.

Following these preliminary steps, we proceed to the core of our analysis. First, we undertake a preliminary analysis of the semantic differences between Appellate Court cases and cases selected for review by the Supreme Court. To carry out this analysis, we undertake a standard logistic regression in which the outcome variable encodes whether a case was selected for review, and the independent variables are share of the case's words associated with each topic, case year, and interactions between case topic shares and year. We find that roughly two-thirds of the topics are statistically significant predictors of whether a case will be selected for review. This analysis confirms that the certiorari process favors some topics and disfavors others. Among the topics least associated with being selected for review are *sentencing* and *trusts and estates*. Among the variables most associated with review are the *weight of precedent* and *federal powers*. We also test for time interactions to examine growth or decline in the influence of these variables.

Our second analysis compares semantic content in the Supreme Court and Appellate Courts. To carry out this analysis, we develop a measure of semantic difference based on the predictive accuracy of repeated logistic regression in successfully categorizing cases written by the Supreme Court or Appellate Courts. We construct distributions of predictive accuracy based on many iterations of logistic regression on random samples of Supreme Court and Appellate Court cases and model assessment on held out cases. These distributions are constructed for each year starting in 1951. We then assess whether there has been

an increase in predictive accuracy over time.<sup>1</sup> We find that there has been a marked increase, indicating that the Supreme Court has become more semantically idiosyncratic in recent decades.

We follow up on this observation by testing one potential hypothesis to explain this difference: that the Court is now selecting cases for review that are more distinct than in earlier times. From the preceding analysis, we have already shown that the certiorari process favors cases with certain topics, and disfavors cases with other topics. Perhaps this separating tendency has grown greater with time. To test this proposition, we apply the same procedure of repeated logistic regression that we used to analyze semantic similarity between the Supreme Court and Appellate Court, but this time we create distributions with respect to cases selected for review and all Appellate Court cases. In this analysis, we find no time trend, indicating that, at least as far as our analysis can show, the cases the Court reviews today are no more dissimilar from the total pool of potential cases than they were several decades ago. A fair reading of these two findings is that the Supreme Court has become less “court-like” over the preceding several decades.<sup>2</sup>

The structure of the remainder of the paper is as follows. Part I provides a highly simplified model of agenda formation for the Supreme Court, discusses prior literature that examines how the Court selects cases and issues for review, and provides general background on the topic model approach used for our analysis. Part II discusses the application of the topic model to the corpora of Supreme Court and Appellate Court cases, and discusses several tests to determine whether the topic model was able to represent relevant semantic features of decisions in those corpora. Part III discusses our primary findings regarding how the process of case selection affects the subject matter of the Supreme Court agenda and our analysis of the change in similarity between the Supreme Court and Appellate Courts over time. Analogous studies could easily be undertaken for other hierarchical court systems. Part IV discusses our

---

<sup>1</sup> Pearson product-moment correlation coefficient between year and accuracy is 0.79.

<sup>2</sup> See *infra* Part IV.A.

findings and offers thoughts on the potential for topic models to contribute to the broader enterprise of empirical legal studies.

## **I. Agenda formation and text analysis**

### *A. A simple model of agenda formation*

The hierarchal judicial structure generates a vast winnowing of cases and issues before they reach the pages of the U.S. Reports. Each year, roughly 30 million cases are filed in state courts, and one million cases are filed in federal courts.<sup>3</sup> In reporting year 2012, there were 35,302 federal appeals terminated on the merits, disposing of a number of cases roughly equivalent to 10% of the non-bankruptcy filings in the federal courts.<sup>4</sup> The vast majority of these appellate dispositions were not accompanied by a published opinion: in the same period, the courts issued 4,165 published decisions and 31,137 unpublished decisions. Finally, the Supreme Court received several thousand petitions for a writ of certiorari and granted 111.<sup>5</sup> Based on these numbers, it is roughly 25 times more likely that a randomly chosen person living in the United States will be struck by lightning in their lifetime than a case filed somewhere in the United States will land in the U.S. Supreme Court.<sup>6</sup>

Since the Judiciary Act of 1925, the U.S. Supreme Court has exercised nearly absolute power over whether or not to hear cases on the merits.<sup>7</sup> And, given that the Court receives thousands of petitions for review each term and hears around 1% of them, the Court certainly has some latitude to decide how it

---

<sup>3</sup> Note that 70% of the federal cases are bankruptcy filings.

<sup>4</sup> Administrative Office of the United States Courts, Annual Report of the Director: Judicial Business of the United States Courts (2013) tbl. B-5, available at <http://www.uscourts.gov/uscourts/Statistics/JudicialBusiness/2013/appendices/B05Sep13.pdf>. This figure does not include the U.S. Court of Appeals for the Federal Circuit.

<sup>5</sup> *Id.* at tble B-2, available at <http://www.uscourts.gov/uscourts/Statistics/JudicialBusiness/2013/appendices/B02Sep13.pdf>.

<sup>6</sup> National Oceanic and Atmospheric Administration, How Dangerous is Lightning? <http://www.lightningsafety.noaa.gov/odds.htm>.

<sup>7</sup> Judiciary Act of 1925; Supreme Court Case Selections Act of 1988. Sup. Ct. R. 10 (criteria for cert grants). There are very few obligatory cases (Cordray and Cordray 2001).

will intervene in the fates of litigants and the interpretation of the nation's laws.<sup>8</sup> But unlike Zeus with his lightning bolts, the U.S. Supreme Court cannot so easily reach out to select the cases of its choice; nor do its decisions appear to follow inscrutable divine whim. Instead, the Court is limited in the cases that it may reach, and its behavior appears to conform to consistent patterns and systematically respond to the actions of external actors. Most obviously, cases are only eligible for Supreme Court review when one of the litigants petitions the Court for certiorari (Mak et al. 2013), and typically only after a final disposition has been entered in a federal or state court of last resort.<sup>9</sup> The Court is also much less likely to review cases when the petition for certiorari proceeds *in forma pauperis* (Provine 1980). The Court is more likely to grant certiorari when there are amicus briefs in favor of (and opposed to) the grant (Caldeira and Wright 1988) and when the United States is a petitioner (Caldeira et al. 1999). Whether a case reaches the Supreme Court, then, is based on a number of decisions emanating from a number of factors, including lower courts, litigants, and amicus parties and the Solicitor General.

The Court also follows its own internal procedures for deliberating over certiorari petitions and allocating power between the Justices. In 1972, Justice Powell suggested the idea of a certiorari pool in which clerks share responsibility for reviewing, summarizing, and making recommendations on certiorari petitions. The recommendations of the “cert pool” clerk, which appear to bear weight (Black et al. 2014), are then passed along and the Justices create a “discuss list” of cases: those for which at least one Justice decides that it merits further consideration (Caldeira and Wright 1990). Following convention, the Justices then convene and follow the “rule of 4” in which less than a majority is needed to grant a petition for certiorari (Revesz and Karlan 1998).

While the process is relatively formalized, if opaque, the Court places few substantive constraints on itself with respect to decisions on certiorari petitions. Supreme Court Rule 10 notes that review “is not

---

<sup>8</sup> See Thompson and Wachtell (2009).

<sup>9</sup> There is an exception in the Court's rules that allows the Court to grant certiorari before a final judgment is entered, but only if the case “is of such imperative public importance as to justify deviation from normal appellate practice and to require immediate determination in this Court.” Sup. Ct. Rule 11.

a matter of right, but of judicial discretion” and offers “considerations” that are “neither controlling nor fully measuring the Court’s discretion.” Those considerations include conflicts between appellate courts of last resort (states and federal circuits) on “important federal question[s]” as well as extreme “depart[ure] from the accepted and usual course of judicial proceedings” or a decision that “conflicts with relevant decisions of this Court.” Finally, the Court may identify “important questions of federal law that . . . should be [] settled by this Court.” The Court also notes that it is unlikely to take up petitions when the “asserted error” was a misapplication of the law or erroneous factual finding.<sup>10</sup>

The Court’s agenda setting power extends beyond the question of whether to grant a petition for certiorari. Ulmer (1982) refers to the phenomenon of “issue fluidity” in which the Court either “discover[s] and decides[s] an issue not raised by the petitioner . . . [or] suppresses and does not decide an issue posed by the petition[er]” (Ulmer 1982, 322). While the degree of issue *expansion* on the Court is the subject of controversy (McGuire and Palmer 1995; Epstein et al. 1996; McGuire and Palmer 1996), issue *suppression* is common and explicit: when the Court grants certiorari, it frequently limits review to specific questions. The Court also exercises considerable latitude in the relative emphasis that it places on the various legal questions presented in a case.

After certiorari has been granted, the Court follows a more formal and transparent process (Codray and Codray, 2004). In addition to briefing, the Court holds oral argument, in which the parties have an opportunity to reiterate points and respond to questions, and most of the Justices engage in a form of public deliberation with each other. There are also substantive and interpretive norms that may constrain the Justices’ discretion on the merits. But Justices have wide discretion to determine the relative weight to place on the potential legal questions that are presented in a case, and each holds the authority to issue a concurring or dissenting opinion. Furthermore, the author of the majority opinion enjoys some discretion over emphasis in drafting (Bonneau et al. 2007). During the drafting phase, versions of the

---

<sup>10</sup> Sup. Ct. Rule 10.

majority, concurring, and dissenting opinions are circulated within the Court, spurring additional deliberations, occasional vote-shifting, and redrafting and editing. All of these internal operations are governed by both formal rules and entrenched conventions (Perry 1994).

These two mechanisms—case selection and issue selection and emphasis—determine the extent to which the Supreme Court’s decisions bear on potential legal questions that it could address. The process in which these mechanisms are embedded is complex and includes cert petitioning (or lack thereof) and supporting briefing (Yates et al. 2013; Caldeira and Wright 1988); the certiorari pool (Black and Boyd 2012); and, of course, merits briefing and oral argument (Sorenson and Johnson 2014) followed by drafting and editing (McGuire and Palmer 1995). The Court is an actor in the formation of its agenda, with an outsized role, but ultimately its agenda is the result of decisions made by a large and diffuse set of actors within the legal system.

For purposes of our analysis, while we look forward to more explicit modeling of agenda formation, we currently treat the details of this process as a black box. Our analysis focuses on the consequences of this process, as measured by semantic differences between the full pool of published U.S. Appellate Court decisions, the pool of decisions that are selected for review, and the pool of decisions written by the Supreme Court. We can therefore reduce and simplify the complex process for forming the Supreme Court’s agenda just discussed.

We begin by defining the *agenda* of the Supreme Court as *the subject matter* of Supreme Court *decisions*. Decisions have similar subject matters when they contain the same or similar words. Words are counted as similar if they reliably co-occur—i.e., if they appear, they tend to appear together in documents in the corpus. As will be discussed in more detail below, the topic model approach formalizes our definition of subject matter. Notice that this definition is specifically oriented to the decisions of the Court, and not the pool of cases selected for review. As a consequence, this definition allows for both the process of case selection and the process of issue selection to determine the Court’s agenda. In addition,

this definition captures the sense in which the agenda of the Court is what the Court ultimately decides *about*, but not the substantive resolution of those issues.

We imagine the subject matter of Supreme Court decisions (i.e., the Court's agenda) as being constructed through a three-step process as follows:

**Step 1.** Generative process  $G$  produces  $AC$ , which is the set of all Appellate Court decisions.

**Step 2.** Selection process  $S$  creates subset  $C \subseteq AC$ , which is all decisions in  $S$  selected for review by the Court (i.e., granted certiorari).

**Step 3.** Transformation  $T$  operates on  $C$  to create  $SC$ , which is the set of all Supreme Court decisions.

Each of the decisions in  $AC$  and  $SC$  comprise a group of subjects of significance. The distribution of these subjects discussed in all of the decisions in  $SC$  is the Court's agenda. We are interested in the relationship between the subjects in  $AC$  and those in  $SC$  and the importance of the intermediary processes  $S$  and  $T$  in affecting the Court's agenda.

### *B. Prior literature*

Given the consequences of the Court's certiorari jurisdiction, it is not surprising that it has long been a subject of conflict on the Court, as well as a subject of study by social scientists and academic lawyers (Harper and Rosenthal 1950; Schubert 1958). Some of the resulting literature is explicitly normative and focuses on the kinds of legal questions that the Court ought to direct its attention toward (e.g., Estreicher and Sexton 1984). For example, one might debate whether, given the limited number of cases that the Court is likely to take in a given year, it should focus on areas of disunity within federal law. Or perhaps it would be better to allow some degree of percolation of issues in the lower courts, to inform later Supreme Court deliberation over the issues (Clark and Kestellec 2012). It has also been suggested that instead of focusing on bringing the circuits into greater harmony, perhaps the Court ought

to focus on areas of the law where its intervention would advance particularly important social goals (Brennan 1973).

Among positive scholars of agenda formation for the Court, most have focused on questions related to certiorari jurisdiction, with a smaller number focused on agenda formation at later stages. There are roughly three theories of how Justices decide which cases to select for review. One theory, advanced by Ulmer (1972) and others, is that Justices seek to identify cases that were wrongly decided in the appellate courts—so that a vote of “yes” by a Justice on certiorari is an indication by the Justice that he/she has found an error that required correction. The “error correction hypothesis” has spawned a number of empirical studies (e.g., Boucher and Segal 1995). A second theory, advanced by Schubert (1958), is that Justices recognize the two-stage game structure of certiorari decision, and vote to grant certiorari when they believe that they would be successful at the merits stage. Scholars have tested various versions of this strategic conception of certiorari decision-making (Brenner 2000; Grant et al. 2012). Another notable theory, advanced by Tanenhaus et al. (1963), is that the Court’s certiorari decisions are based on “cues,” including whether there was a dissenting judge in the case below and whether the United States is a petitioner. This cue (or “signaling”) theory allows for the interaction of the Court with other strategic actors in a mutual game that ultimately shapes the Court’s docket (Cameron et al. 2000). Justices looking for cues may seek to identify cases that will promote their policy goals or may apply a more “jurisprudential” analysis oriented toward error correction or uniformity (Perry 1991; Cordray and Cordray 2004; Black and Owens 2009). Black and Boyd (2012) look beyond the Justices’ preferences to examine the role of court procedures, and in particular the influence of the clerk charged with making an initial recommendation as part of the cert pool. They find that the assignment of the clerk during the pool process has a significant effect on the ultimate certiorari decisions.

Scholars have also studied how the Court decides how to allocate its attention within the potential subjects presented by a case selected for review. Petitioners to the Court often raise a large number of objections to the lower court opinion that is being appealed, and the Court frequently only certifies a

limited number of them for review, effectively refusing certiorari with respect to the remaining questions. The Court, however, is not bound by that decision, and sometime issues are resuscitated during oral argument (Sorenson and Johnson 2014). McGuire and Palmer (1995) examine the Court's 1998 term to determine the frequency of "issue fluidity" in which some issues present in the grant of certiorari are suppressed while others that were not present are discovered. They find that fluidity of some form is common. Epstein et al. (1996) critique the data and methods used by McGuire and Palmer (1995), ultimately concluding that norms against raising issues *sua sponte* remain strong on the Court. McGuire and Palmer (1996) and Palmer (1999) respond to the critiques raised in Epstein et al. (1996) and expand their analysis, again concluding that the issues covered in the final decision do not exactly track those that are raised in the cases granted certiorari.

A related line of research examines the respective roles of opinion drafter and median Justice on the bench in influencing the content of final majority opinions (Carrubba et al. 2012). Under a basic model of majority decision-making, the median voter determines outcomes (Downs 1957) and can be expected to exercise considerable pressure on case outcomes in the Supreme Court (Epstein and Lee 1998). Maltzman and Wahlbeck (1996) find that Justices do indeed switch their votes for strategic purposes. But there are reasons to believe that majority authors also have some ability to shape the outcome of decisions (Lax and Cameron 2007): researchers have found that majority authors (Lauderdale and Clark 2013; Bonneau et al. 2007) and coalition members (Carrubba et al. 2012) influence opinion content. The post-argument phase, in which authoring roles are assigned, opinions are drafted and circulated, and bargaining ensues, then becomes an important stage in the process of determining the content of final decisions.

The influence of actors outside of the Court also has been subject to scrutiny. Most obviously, litigants take on a substantial role in shaping the issues that are presented to the Court. It is costly to litigate cases before the Supreme Court, a fact that Zorn (2002) finds affects even the Solicitor General's

decision about whether to seek certiorari.<sup>11</sup> Even when the government is not a party to the case, the Solicitor General is frequently consulted by the Court over certiorari decisions and appears to have substantial influence (Black and Owens 2012; Thompson and Wachtell 2009). Mak et al. (2013) and Yates et al. (2013) examine the interaction between litigants' choices over whether to appeal and the ideological dispositions of Justices currently on the Court. In turn, Baird (2004) examines the degree to which the Court can signal to litigants the kind of cases that it would like to hear. Caldeira and Wright (1988) examine the influence of amicus parties, finding that the presence of amicus supporters and opponents increases the likelihood of a grant of certiorari. McGuire and Caldeira (1993) examine the effects of the organized bar on agenda selection, finding a significant effect associated with participation of the "obscenity bar" in relevant cases. Epstein et al. (2002) examine the role that Congress plays in influencing the decisions that are taken up by the Court.

Ultimately, Yates et al. (2013) argue that "[a] wealth of judicial politics literature suggests that Justices have an interest in taking on cases that are salient, resolve important legal conflicts, and, in fact, do map well onto Justices' distinct ideological preferences."<sup>12</sup> In Black and Boyd (2012), the authors combine this literature into an overall model that determines the "certworthiness" of a petition. The factors identified by these authors include whether the case presented an issue of legal conflict, such as a Circuit split; whether the Solicitor General supported the grant; whether the decision below was published and included a dissenting opinion; whether the litigant hired counsel to file the petition; and whether it presented constitutional claims, civil liberties claims, and was opposed or supported by amicus parties. While only some of these variables were found to be significant in logistic regression of the Court's decisions in the study period, collectively they added substantial predictive value.

---

<sup>11</sup> It bears noting that the United States has a more expansive conception of costs than that of the average individual. Zorn (2002) estimates costs to the government in terms of potential for civil liability or invalidation of prior government action.

<sup>12</sup> Yates et al. (2013: 852).

Scholars have also used extensive and thorough coding schemes to examine how the issue agenda of the Supreme Court has evolved over time. Pacelle (1991) examines the “apportionment of agenda space” among fourteen policy areas and how that apportionment has shifted over time, identifying rough eras between 1933 and 1988 during which, in turn, economics, federalism, regulation, and civil liberties dominated the Court’s agenda. Hurwitz (2006) takes advantage of the similarity in coding protocol between Spaeth’s Supreme Court Database (Spaeth et al. 2014) and Songer’s Database of Appellate Court Decisions (Songer 1998) to compare how the two institutions interact, finding that the Supreme Court’s agenda choices affect litigant behavior in such a way as to influence the composition of the cases in the Appellate Courts (Songer 1998). Rice (2014) takes a computational text analysis approach to examine whether Court intervention in an issue area affects the composition of cases in the lower court. For that analysis, Rice uses a supervised learning algorithm to classify cases in the Appellate Courts according to their Policy Agendas Project issue area and then analyzes the effects, over time, of a Court decision in one of those areas. The finding of that analysis is that the Court simultaneously decreases litigation and increases publication rates when it issues a decision in an area.

Our study builds on the prior literature through a principled application of topic model analysis to the corpus of federal and Supreme Court cases. We restrict our analysis to the period of agenda formation that begins with the choice by an Appellate Court panel to publish a decision in the Federal Reporter and ends with a case in the U.S. Reports. Topic modeling is used to estimate the relative weight in a decision that is paid to a distribution of reliably co-occurring and typically semantically related words (referred to as “topics”). This analysis is applied to three sets of cases: all cases published in the Federal Reporters during the study period; the subset of those cases that were selected by the Supreme Court for review; and the subset of those cases that were published in the U.S. Report (i.e., Supreme Court decisions). The first set is used as a baseline to determine the relative prevalence of different topics<sup>13</sup> within the total body of

---

<sup>13</sup> Henceforth, we reserve the word “topic” as referring to its technical meaning within the statistical learning methodology of topic modeling, explained *infra* Part I.C.

federal court cases. We can then compare the differences in subject matter between the federal court cases baseline and the two subsets, with a null hypothesis that both subsets represent a random draw and should therefore reflect the mix of topics that are represented in the total population. Similarly, the two subsets can be compared to determine the degree of agenda formation that occurs through case selection as well as after selection of cases for review.

There are several advantages to the topic model approach used here, as well as important limitations that should be noted. Topic modeling allows use of all published federal court cases as a baseline for the relative presence of topics that could form the subject of Supreme Court decisions.<sup>14</sup> We can therefore examine together the effects of litigants' choices over whether to petition and the Court's decision to grant certiorari, sidestepping the difficult question about the relative influence of petitioners and the Court in setting the Court's docket.<sup>15</sup> Similarly, rather than seeking to identify the case features or strategic techniques that the Court or individual Justices use to select cases, we create a framework wherein we can measure the effect of whatever signaling or strategic factors are used over the subject matter of the Court's decisions. In addition, our analysis is able to generate a finer grained estimate of the subject matter of decisions than both the Songer database and the Policy Agenda Project categories deployed by Rice (2014). With 100 topics, we have a larger number of potential subjects than the Policy Agenda Project. Perhaps more importantly, our analysis decomposes cases into weighted distributions, rather than lists, of the topics that they contain. This enables a more detailed quantitative case comparison that is more closely linked to the actual text, because the relative weight that cases afford to a subject can

---

<sup>14</sup> For purposes of convenience, we use a large random sample of appellate court cases to represent the entire population. There is no genuine practical limitation on applying the topic model to the entire corpus—it is simply a matter of waiting for the computation (a process that can take several weeks). In the final published version of this paper, we will likely fit the topic model to the entire corpus for the sake of completeness.

<sup>15</sup> In a future analysis, we plan to separate the cases that were petitioned for certiorari in order to test the degree of agenda formation that occurs at the petitioning stage. Even that additional analysis, however, will not be able to determine the degree to which litigants' behavior is conditioned by beliefs about the type of cases that the Court is likely to accept.

be estimated. This characteristic of topic models gives them substantial power to represent the subject matter of a corpus of texts in ways that are helpful for analysis.

We do not attempt to show that the topic categories that we identify in this study contribute directly to the Court's, or to any other actor's judgment. Rather, we measure the overall effect of whatever set considerations bear on the wide range of actors that play a role during the studied stages of agenda formation. An example can help illustrate this study's contributions as well as its limitations. Ulmer (1984), among others, has shown that the presence of a circuit split is a predictive variable for a grant of certiorari. If this is the case, and some subjects are more likely than others to create a circuit split, then those subjects should be more highly represented in the cases selected for review than the Federal Appellate corpus as a whole. Our analysis is able to estimate whether there is a difference in subject matter in the cases selected for review and Supreme Court decisions. But we do not attempt to identify or disaggregate the causal mechanisms that generate those differences: it may be that differences in subject matter are an incidental effect of the Court's focus on circuit splits, or it may be that in the process of selecting cases with the desired subject matter, cases with circuit splits are selected at a higher rate. Our analysis does not attempt to answer that question, or account for the range of non-topic-related factors that might predict whether a case is taken up for review.

What our study does produce is a useful methodology (topic modeling) to estimate a quantitative variable—subject matter difference—directly from text corpora of interest. Besides the kinds of analysis presented in this paper, such measures can be used in predictive and causal analysis of agenda formation. Herein we present this as a methodology for tractably measuring semantic content and to show that there are significant differences in the semantic content of Supreme Court decisions, cases selected for review, and the mine-run of published decisions in the Appellate Courts. It is likely that this content interacts in complex ways with the causal processes that determine the Court's agenda. In Part IV.B, we present some ideas for ways in which topic modeling can provide a useful means of examining these causal processes, but we leave that task to future analysis.

### *C. Topic models*

Representations of Supreme Court and Appellate Court decisions in terms of word use (vocabulary and frequency) are ubiquitous in the lives of legal scholars and practitioners because they enable the keyword searchers that serve as a bedrock of contemporary legal research. Recent advances in text analysis have built on word frequencies as representations of textual documents to construct sophisticated statistical models capable of drawing out relevant features of large collections of documents (“corpora”). These new approaches include a class of technique referred to as “topic models,” which have proven especially useful to scholars in both the social sciences and humanities interested in the analysis of large textual corpora. This following discussion provides general background on topic models and their growing use in text analysis.

Raw word frequencies can provide an efficient, if crude, measurement of “similarity” between two texts that can be used to suggest related documents and rank search results. This information can then be subjected to additional manipulation to draw out relevant characteristics within texts. For example, a simple weighted variation of word frequency calculates the *term frequency-inverse document frequency* (referred to as “tf-idf”), which places relatively less weight on words that are very common within the corpus.<sup>16</sup>The tf-idf measure, and variants thereof, are used in search engines as a way to estimate the relevance of a document for a user query.

Simple word frequencies suffer from a number of drawbacks. Representing texts in terms of a high-dimensional feature vector of word frequencies (defined, canonically, with respect to the vocabulary of the entire corpus) fails to reliably capture semantic features of those texts. For example, they do a poor job of disambiguating multiple meanings (e.g., “flag” as an emblem and “flag” as meaning to

---

<sup>16</sup> The tf-idf weight is made up of the normalized term frequency (tf), which is the number of times a word appears in a document divided by the total number of words in that document multiplied by the inverse document frequency (idf), which is the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

“distinguish”).<sup>17</sup> Importantly, taken alone they do not provide useful information about a text’s membership in broader, more familiar categories (e.g., “cases concerning the First Amendment”). The pervasive use of a variety of (costly) human headnote-type annotations provides a measure of evidence for the value of these additional higher-level categorizations for legal practitioners. These higher-level categories are also more useful to those interested studying the relative degrees of attention given to different areas of law.

A topic model offers one means of recovering accurate approximations of these higher-level categorizations without the expense (in time and resources) of hand-coding decisions. The “topics” within these models are word frequency-based; however, by their construction they provide a good deal more granularity. To be a bit more precise, a “topic” (in the technical topic modeling sense) is a probability distribution over a vocabulary, where each word in the vocabulary is assigned a positive weight (such that the collection of all weights sums to one). The most highly weighted words within a topic provides a sense of the subject matter that the distribution represents. For example, in Topic 14 generated by our model, the words “estate, trust, death, property, decedent, wife, interest” are weighted highly, which led us to hand label that topic as “*trusts and estates*”.

These distributions are produced as the best fit to an underlying generative probabilistic model for the observed count data (i.e., word frequencies). The canonical topic model is a latent Dirichlet allocation (LDA) mixed-membership model.<sup>18</sup> The LDA model posits some number of topics (distributions over the vocabulary) that account for all words observed in a corpus according to the following generative story: for each document in the corpus, a set of topic proportions (or “shares”) is drawn from a global

---

<sup>17</sup> Similarly, word frequencies alone would fail to distinguish between the use of the word “minor” in “minor parties” and “minor offense” from the use of “minor” in, for example, *Bethel School District v. Fraser* (478 U.S. 675), a decision concerning free speech in public schools. To add further nuance to this discussion, it should be noted that the plural form of the noun, “minors,” has a comparatively modest range of interpretations.

<sup>18</sup> The Dirichlet distribution is a family of continuous multivariate probability distributions commonly used as prior distributions in Bayesian statistics, included in mixture models. Mixed-membership models are a statistical framework in which groups are modeled as mixture over a shared overarching structure (Zhang & Paisley 2015; Airoldi et al. 2008). For our purposes, documents are the groups (of words), with the ‘topics’ as the shared overarching structure.

probability distribution; then, each word in the document is drawn from a topic distribution in which the topic distribution in question is selected according to the previously mentioned document-specific set of proportions.<sup>19</sup> Drawing from Blei (2012), Equation 1 provides the joint distribution of the hidden and observed variables for this generative process.

$$(1) p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

In this equation, the hidden variables are as follows:

Topic distributions (distributions over the vocabulary):  $\beta_{1:K}$  where  $\beta_k$  is topic  $k$ .

Topic proportions (distributions of topics within documents):  $\theta_d$  for the  $d$ th document where  $\theta_{d,k}$  is the topic proportion for topic  $k$  in document  $d$ .

Topic assignments (topics associated with words):  $z_d$  for the  $d$ th document where  $z_{d,n}$  is the topic assignment for the  $n$ th word in document  $d$ .

The observed variable is the words that are located within documents in the corpus:  $w_d$ , where  $w_{d,n}$  is the  $n$ th word in document  $d$ . The posterior distribution (i.e., the conditional distribution given the observed data) to be approximated is given in Equation 2.

$$(2) p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

Here, the denominator is the marginal probability of the observations, which can be understood as the sum of the joint distribution over every possible version of the hidden topic structure. This value cannot be directly calculated because of the extremely large scale of possible topic structures. Instead, it is approximated through algorithms that respond to this computational problem.

The algorithms used to estimate topic model parameters bear a family resemblance to the techniques used in linear regression. Topic models are often fit using an iterative algorithm (known as

---

<sup>19</sup>Each topic distribution is typically given a minimally informative uniform Dirichlet prior.

variational approximation) or by using a Markov Chain Monte Carlo (MCMC) approach. In the case of the topic model, the parameters of interest are typically restricted to the topic-word distributions describing the association between topics and words, and the document-topic distributions which describe for each document, the probability of finding words associated with each topic. The particular variant of MCMC used in this case is called a Gibbs sampler, which makes drawing samples from a high-dimensional posterior distribution tractable by considering a sequence of conditional distributions of smaller subsets of the parameters (Casella and George, 1992).

A useful reference point for understanding the inferential machinery of topic models is the ideal point model of Supreme Court justices' votes developed by Martin and Quinn (2002). Topic models and ideal point models are Bayesian models which use the strategy of latent variable augmentation to facilitate inference and make use of MCMC to estimate model parameters. The ideal point model confronts the challenge of translating from a continuous space (the embedding of the ideal points) to a discrete space of observed votes. In the case of the topic model the underlying continuous space is the probability structure and the discrete space is the observed space of word frequencies.

A primary advantage of the LDA topic model over earlier word frequency approaches is that it accounts for probability of word co-occurrence, which intuitively helps to uncover characteristics of context (e.g., multiple co-occurrences of "flag" with "waving" versus "flag" with "manuscript" would begin to distinguish two of the definitions and usages of "flag"). This characteristic also allows for the radical reduction in the dimensionality needed to describe documents within a corpus. Any given group of documents may contain as many as 50,000 unique words. To represent documents as frequency distributions over those words requires vectors with an equal number of components. The number of topics generated by a topic model is much smaller, typically fewer than 500. In the present case, we fit the corpus with 100 topics. The parameters in the topic model that are of interest are the document-specific distributions over topics (i.e., the topic proportions) and the topic distributions themselves. The document-specific distributions have an interpretation as the proportion of words in a document that were

“generated” by each topic. The topic distributions are distributions over words and are typically described in terms of words with the highest likelihood of having been sampled from the relevant distribution.

The tendency of the LDA model to group together semantically connected words (such as “child” and relevant uses of “minor”) earned it the moniker “topic model” (Blei, Ng, and Jordan 2003). These models have been shown in practice to succeed where the naïve word frequency or other “bag-of-words” representations (i.e., approaches that treat a document as a simple list of word frequencies) fall short (D. M. Blei, Ng, and Jordan 2003; Pritchard, Stephens, and Donnelly 2000).<sup>20</sup>

While numerous incremental improvements to LDA topic modeling have emerged in the intervening years, the essence of the original model remains and the LDA topic model remains a general “industry standard” for text analysis and serves, with minor variations, as a building block in more elaborate models of text data.<sup>21</sup> More than a decade after the model's introduction, researchers using topic models and closely related models may be found in almost every field where machine-readable text data is abundant. Topic models are now a familiar part of the methodological landscape in the human and social sciences, in fields such as women's history, political science, history of science, archeology, and classical studies (Block and Newman 2011; Grimmer 2010; Hall et al. 2008; Mimno 2011; Mimno 2012; Quinn et al. 2010; Riddell 2014).

Demonstrating that topic models succeed in capturing semantic features of texts is itself a continuing area of research, in particular because there is no shared definition of “semantic features.” Traditional statistical measures—e.g., the ability to predict words in new or unseen texts from the same corpus—provide a means to evaluate a topic model’s performance compared to previous approaches. But the capacity of a topic model to decompose a corpus into semantic themes or topics relies heavily on

---

<sup>20</sup> Other early methods for dimensionality reduction directed at recovering semantic representations of texts include (probabilistic) Latent Semantic Indexing, Principal Components Analysis, and Mixture of Unigrams (Nigam et al. 1999; Hofmann 1999).

<sup>21</sup> For example, the two Dirichlet distributions involved in latent Dirichlet allocation may be replaced with a single non-parametric hierarchical stochastic process such as the hierarchical Dirichlet Process or a hierarchical Pitman-Yor Process (Teh et al. 2006; Teh 2006; Doyle and Elkan 2009).

anecdotal testimony of the kind Justice Potter Stewart would appreciate: many people observe that the inferred probability distributions over words (topics) tend to look like lists of semantically connected words. In addition to anecdotal evidence, a variety of (indirect) empirical tests involving human readers have been conducted. In these tests, readers' classifications of texts are compared with the representation provided by the topic model, or the topics themselves are directly evaluated (D. M. Blei, Ng, and Jordan 2003; Chang et al. 2009). The widespread use of topic models in research and industrial settings also offers some validation of their ability to produce parsimonious representations of large text corpora (Rice 2012; Quinn et al. 2010; Ahmed et al. 2012). In Part II.C we discuss our approaches to validating the topic model described here.

## **II. Applying the Topic Model**

### *A. Data*

In the past, computational analysis of legal texts has been hampered by difficulties accessing the relevant data (Arewa 2006). Although judicial opinions are not copyrightable, the commercial databases that provide ready digital access to these opinions are protected by terms of use agreements. Limits on machine reading may be necessary to protect the proprietary content that has been produced by these publishers, but they can also inhibit academic research and access to the non-copyrighted government documentary information included within these resources.

Public.Resource.Org, a private not-for-profit corporation, has created a digital version of the Supreme Court and federal Appellate Court corpus, based on the non-copyrightable information within the Westlaw database, and published that information online at "bulk.resource.org." The bulk resource data has been used in prior n-gram studies of text usage in the federal courts and Supreme Courts,<sup>22</sup> and provides the public with access to a digital version of the nation's judicial opinions. However, the bulk resource data in its raw form has some important limitations, including a lack of readily available meta-

---

<sup>22</sup> See Katz et al. (2011).

data. CourtListener, an effort within the Free Law Project has augmented the information contained in the bulk resource data and created a user friendly interface, and provides public access to its data. We relied on CourtListener as the source for all the texts for the Supreme Court and Appellate Court decisions.

The set of Supreme Court decisions used in this study includes all formally decided full opinion cases as defined in the Supreme Court Database.<sup>23</sup> It covers the period between 1951 and 2007, inclusive. There are 7,528 documents in this set. The median decision length for this set is 3,339 words. The full set of Appellate Court decisions issued between 1951 and 2007 includes 289,550 documents. To reduce the computational burden of fitting the topic model, we randomly selected 25,000 decisions from within the Appellate Court set. As these decisions were selected at random, they are representative in a well-understood sense. The number of randomly sampled decisions is large enough that we believe any parameter estimates discussed would be essentially unchanged were the number expanded. In addition to the 25,000 randomly selected Appellate Court decisions, 4,180 Appellate Court decisions which were referenced by Supreme Court decisions in the manner described below are also included. In total there are 29,180 Appellate Court decisions.<sup>24</sup> The median decision length is 2,781 words.

To identify the set of Appellate Court decisions that were selected for review by the Court, we gathered information from Lexis/Nexis, which provides “prior history” and “disposition” fields for Supreme Court decisions.<sup>25</sup> Citations to Federal Appellate court decisions were identified by their Federal Reporter citation (either “F. 2d” or “F.3d”). We verified the accuracy of the citations provided by

---

<sup>23</sup> Specifically, using the Supreme Court Database coding, we include decision types 1 (opinion of the court, orally argued), 2 (per curiam, no oral argument), 5 (equally divided vote), 6 (per curiam, orally argued), and 7 (judgment of the Court, orally argued). We exclude decrees and seriatim opinions.

<sup>24</sup> While the resulting corpus of Appellate Court decisions is biased in that it has more Appellate Court decisions which were referenced by Supreme Court decisions than is actually the case, this imbalance is unlikely to influence the estimation of the topic proportions associated with each case. In the unlikely event that these cases (decisions subsequently referenced by Supreme Court decisions) systematically used different language than other Federal Appellate decisions, their influence on the composition of the individual topics would be small as these cases are a small percentage of the total cases in the corpus.

<sup>25</sup> The accuracy of the citations provided by Lexis/Nexis was assessed by collecting a random sample of 100 Supreme Court cases and comparing the citations with those found by a manual check. All citations were accounted for.

Lexis/Nexis by randomly sampling 50 decisions and manually checking the lower court citation. There are 724 decisions within our random sample of Appellate Court decisions that are associated with decisions in the Supreme Court corpus (i.e., appeals which were granted certiorari).

The vocabulary associated with the corpus is comprised of those words occurring at least 20 times in the entire Supreme Court corpus. There are 21,695 total words in the vocabulary. Some topic model applications exclude function words (sometimes referred to as “stop words”) from the analyzed vocabulary.<sup>26</sup> Our analysis retained all words in the vocabulary, subject to the minimum appearance constraint. As would be expected from Zipf’s law, word frequencies within the vocabulary are strongly right-skewed, as is apparent in Figure 1.<sup>27</sup>

### *B. Fitting a topic model to the legal corpus*

The LDA topic model was introduced in over decade ago (Blei 2003) and a substantial amount of subsequent work has improved on the basic framework. The more advanced variant that we use is a non-parametric topic model using the Pitman-Yor Process in place of the traditional Dirichlet distributions (Buntine and Mishra 2014). This non-parametric version preserves the core heuristic of LDA while performing considerably better on held-out prediction tasks. The software we use “hca” is authored by Buntine and is open source (Buntine 2014).

One of the modeling choices common to topic models is the number of topics selected by the modeler. More recently developed topic models attempt to be invariant to (a sufficiently large) number of topics, such that, for a given corpus and a given probabilistic model there is, at least in theory, a “natural” number of topics for a given probabilistic topic model. Under such an assumption, specifying an

---

<sup>26</sup> Function words include: “is”, “as”, “of”, “though”, “around”. These words are frequently used in stylistic analysis of texts. *See generally* Carlson, Livermore, and Rockmore 2016.

<sup>27</sup> Zipf’s law, named after linguist George Kingsley Zipf, conjectures that in natural language use, the frequency of a word is inversely proportional to its rank. If true for a given corpus, then the second most common word will be used half as frequency as the first and the third most common word will be used half as much as the second (and a quarter as much as the first) and so on.

excessively large number of topics will simply result in a large number of “empty” topics which are associated with a negligible percentage of words in the corpus. In the current analysis, we select 100 topics, which is large enough to capture a great deal of the semantic variability of the corpus and small enough to make fitting the topic model computationally tractable.<sup>28</sup>

Table 1 (appendix) provides a list of the topics generated by the model, the top words in each topic (i.e. the most mostly highly weighted words in the distribution) and the hand-coded topic labels.<sup>29</sup> The top words for the first three topics generated (the order in which topics appear is not meaningful) were:

- 1: “union board labor employees employer nlrp company bargaining relations national local act unfair order strike respondent practice cir work unit unions”
- 2: “ms mrs did told husband time testified asked sexual home stated fact received mother daughter claim record office day appellant later return woman given leave left”
- 3: “election political party candidates candidate campaign parties primary elections contributions ballot office expenditures voters state vote buckley amendment”

We labeled these topics “labor,” “family” and “elections” respectively, which map nicely to intuitive subjects within judicial opinions.

Figure 2 provides an illustration of how the “topics” generated by the model map to the words within the text of an opinion, using *NCAA v. Board of Regents of the University of Oklahoma*, 468 U.S. 85 (1984) as an example. In our 100-topic fit of the Supreme and Appellate Court opinions, the words of *NCAA v. Board of Regents* derive in large part from topics 23 and 95. Table 1 shows that the following

---

<sup>28</sup> In subsequent versions, we plan to expand the number of topics for the sake of completeness.

<sup>29</sup> Most of the topics admitted relatively straightforward labeling, although a few were a bit more difficult and some defined labeling. One potential test, not performed by the authors (yet), would be to provide law student research assistants with lists of topics, drawn from a human headnote-style coding system, and ask them to match headnotes to topics based on the top words.

words are assigned high probability in topic 95: “school,” “university,” “education,” and “students.” As this constellation of words does not characterize any of the other 100 topics, we are tempted to call this the “education” topic. High probability words under topic 23 include “antitrust,” “market,” “competitive,” and “anticompetitive.” Again, this fits with what we know about the case (the issue at hand is how laws concerning the regulation of anticompetitive behavior should be applied to an association of university athletic programs).

### *C. Testing the topic model*

In this section we discuss the results of our initial test of how well the topic model captures relevant semantic characteristics of agenda formation on the Supreme Court. For this analysis, we first examine the general semantic fidelity of the topic model by examining how well a supervised classifier categorized cases according to the issue labels provided in the Supreme Court Database (Spaeth et al. 2014) with, on the one hand, total word frequency counts, and on the other, only the topic information for each case. Second, we examine three naïve and intuitive hypotheses concerning the subject matter distribution within Supreme Court decisions. These tests are meant to determine whether the topic model discussed above has captured semantic features with sufficient salience to enable additional useful analysis.

#### **Test 1. Performance of supervised classifier**

One method of testing or validating that a topic model (or any dimensionality-reducing technique) preserves information about the substantive issues found in opinions is by calculating what information about known classifications is lost when reducing a rich, high-dimensional representation of a corpus (word frequencies) to a low-dimensional summary (topic model). A common method for calculating this loss considers how well a supervised classifier performs at the task of placing documents in appropriate categories when provided with either (1) the low-dimensional summary provided by the topic model or (2) the high-dimensional data of word frequency counts. (Blei (2003) makes use of this method.) This method of validation is only available in situations where exhaustive expert labels are

provided. Fortunately, the Supreme Court Database provides precisely such an exhaustive set of labels for the Supreme Court decisions. Using the 14 issue area labels used in the SCDB dataset (variable “issueArea”) as the reference categories and multi-class logistic regression as the classification algorithm, we find that the loss in classification accuracy is minimal (Figure 3). Classification accuracy is assessed using cross-validation; by turns 5% of the 7,503 Supreme Court opinions are "held out" and the algorithm, given access to the remaining documents with the Spaeth labels, is asked to predict the labels.

The classifier should, in theory, have an easier task at predicting the SCDB issue area label using the original feature set (word frequencies of 21,695 unique words) than using a smaller set of topic shares. That the classifier is able to perform as well as it does indicates that the topic model is capturing information that is highly relevant (in the sense of being predictively useful) to the expert issue area classifications. This validation is rudimentary and does not assess the validity of the claim that the topic model more faithfully captures information than a simple discrete labeling. Fully evaluating topic models would require expert labels of roughly the same granularity as the topic model (e.g., 90% of the words in the case relate to privacy and 10% relate to federal taxation).

## **Test 2. Naïve hypotheses**

*Hypothesis A. The topic distributions in Supreme Court decisions resemble the topic distributions in the matched opinion selected for review.*

Our first naïve hypothesis tests whether the topic model is able to identify semantic similarities between cases selected for review and the matched Supreme Court opinions. The null hypothesis is that there is no difference in the semantic similarity between: (1) matched Supreme Court cases and cases selected for review; versus (2) Supreme Court cases and randomly selected lower court cases. Semantic similarity is represented by the distribution of topics in each case. This test is performed for each Supreme Court opinion independently: we compare the dissimilarity between a Supreme Court case and the appellate opinion under review with the (distribution of) dissimilarity we would see if the appellate opinion was selected at random from all appellate opinions.

For our measure of “resemblance” we rely on a symmetrized form of the Kullback-Leibler (KL) divergence, a standard measure of the difference between two probability distributions.<sup>30</sup> Since every decision in the corpus can be represented in terms of a vector of probabilities over the 100 topics, the symmetrized KL divergence is a natural method for estimating similarity (i.e. dissimilarity) between topic shares. Following convention, we use the symmetrized version of KL divergence that is the average of the KL divergences of A with respect to B and B with respect to A (see Equation (3)).<sup>31</sup>

$$(3) D_{KL}(A, B) = \frac{1}{2} \sum_{w \in W} A(w) \log \left( \frac{A(w)}{B(w)} \right) + B(w) \log \left( \frac{B(w)}{A(w)} \right).$$

The procedure that we used for this first null hypothesis test is as follows. For a given Supreme Court decision ( $SC_j$ ) we identified the matching Appellate Court decision ( $AC_j$ ). We then calculated the KL divergence between the topic shares of these cases. For that same Supreme Court decision, we then randomly drew 10,000 Appellate Court decisions  $\{AC_1, AC_2 \dots AC_{10000}\}$  with replacement and calculated the KL divergence between those cases and the Supreme Court decision.

$$(4) D(SC_j) = \{D_{KL}(SC_j, AC_1), D_{KL}(SC_j, AC_2) \dots D_{KL}(SC_j, AC_{10000})\}$$

We then calculated the mean and standard deviation for  $D(SC_j)$  and construct a p-value based on  $D_{KL}(SC_j, AC_j)$  and those values.

Figure 4 shows the distribution of p-values derived by performing this test for each Supreme Court decision with its matching Appellate Court decision. The mean p-value is 0.03; we can reject the hypothesis that the topic shares associated with Supreme Court decisions are no more similar to the Appellate Court decisions being reviewed than to randomly selected decisions. As is apparent from Figure 4, the distribution is right skewed, with some p-values that are quite high. This is unsurprising given that, for each Supreme Court decision, there is only one observation. Although there are individual Supreme Court decisions for which we cannot reject the null hypothesis, taken as a whole, we observe

---

<sup>30</sup> See Hughes et al. (2012); Carlson Livermore and Rockmore (2016).

<sup>31</sup> Hereinafter, when we refer to using the KL divergence we implicitly mean in this symmetrized form.

greater similarity between Supreme Court decisions and their matched Appellate Court decisions than would be expected by random chance. There is therefore a high degree of confidence that the topic model is successful at capturing semantic similarities that we assume exists between a case selected for review and the accompanying Supreme Court decision.

*Hypothesis B. Supreme Court decisions and Federal appellate decisions are distinguishable in terms of their topics.*

We now examine whether the topic model is able to identify aggregate level differences in the semantic content of Supreme Court decisions and federal court decisions. To undertake this analysis, we rely on the topic share in each decision, interpretable as the proportion of words in the decision associated with a given topic. For each topic, we can examine the distribution of document-specific proportions across all Supreme Court cases and all Appellate Court cases and ask whether or not the distributions of these proportions are different using the Kolmogorov-Smirnov (KS) test.<sup>32</sup>

The first step is to construct the 200 distributions corresponding to each of the (100) topics in each of the (2) corpora. The observations in these distributions are the topic shares in each of the decisions, so that  $t_{d,k}$  is the share of topic  $k$  in decision  $d$ . For each topic, the observed distribution is made up by the number of decisions at each point along a continuous zero to one axis that corresponds to the frequency of that topic within those decisions. Each of the topic distributions for the Supreme Court cases can then be compared to the corresponding distribution for the Appellate Court cases.

Two probability distributions may be said to be different in a variety of ways. In some cases, assessing the difference between two distributions' means may suffice (such as in the Student's t-test). In other cases, it may be important to also consider the difference between the distributions' variances. Here the comparison relies on an assessment of the difference between two distribution using the KS test, which is the non-parametric analog to the Student's t-test and thus does not assume normality. Rather

---

<sup>32</sup> The KS test is a non-parametric analog to Student's t-test (which only tests for a difference in the population mean).

than comparing two distributions via their means, the KS test considers both location and scale by using the empirical distribution function. We are able to use this more precise test (compared to the KL-divergence approach used in *Hypothesis A*) because we have, in essence, more information about the distributions to be compared. For *Hypothesis A*, there is only a single observation of a given Supreme Court-to-Appellate Court match. We thus use a less precise measure (KL-divergence) to compare any two distributions, but we are able to do that comparison many times to construct the distribution in Figure 4. On the other hand, comparing the two corpora allows us construct a more reliable distribution for each topic, making the KS test more appropriate.

With the present 100-topic fit, the distributions of shares in each of the two populations are all unambiguously distinct according to the KS test for all topics. Figure 5 shows the distribution of topic shares for four topics where the expected difference in topic shares between the Supreme Court and the Federal Appellate courts is large. Words associated with topic 91 tend to make up a greater percentage of words in Federal Appellate decisions than Supreme Court decisions. Topic 91 includes a variety of words connected to contract law, including “contract,” “agreement,” “parties,” “breach,” and “terms.” Words associated with topic 28 are associated with constitutional law and appear more frequently in Supreme Court decisions. Words associated with topic 28 include “amendment,” “constitutional,” “power,” and “rights.” In the remaining ten topics with the largest differences in shares, the Supreme Court tends to have a higher share of words associated with statutory interpretation and constitutional rights, while the pool of federal cases had a higher share of words associated with time and dates, testimony and evidence, proof, and appeals.

As with *Hypothesis A*, we can safely reject the null hypothesis that the Supreme Court is, in essence, indistinguishable from just another Appellate Court. Failure to reject this hypothesis would have cast doubt on the ability of our topic model to represent semantically meaningful categories. Since there are strong intuitive reasons to believe that there are dissimilarities between these two corpora, the ability

of the topic model to pick up on (at least some of) those dissimilarities provides additional reason to have some level of confidence in this representation of the data.

*Hypothesis C. The corpus of Supreme Court decisions is topically distinguishable from the corpus of Federal Appellate decisions selected for review.*

Given the clear differences between the topics characterizing Federal Appellate decisions as a whole and Supreme Court decisions, it is natural to ask how the subset of Federal Appellate decisions that are selected for review compares with the body of Supreme Court decisions. If the Supreme Court selects cases at random, then there would be no greater similarity between those cases and the Supreme Court cases, because the (randomly) selected cases would simply mirror the broader population of Appellate Court cases. There are strong intuitive reasons to believe that this is not the case, and we anticipate that the body of Federal Appellate decisions selected for review should more closely resemble the Supreme Court decisions than the body of all case. Failure to reject the null hypothesis in this case would again indicate that the topic model is not performing well at capturing relevant semantic features of cases.

This analysis is carried out in the same fashion as in *Hypothesis B*, substituting the set of Appellate Court cases selected for review for the set of all Appellate Court cases. A set of 200 distributions is created using the same method, and the KS test is applied. In this case, the pairs of distributions for only 29 of 100 of the topics are different at traditional significance thresholds ( $p < 0.001$ ). (Recall that in testing Hypothesis 2, the pairs of distributions were clearly different for all topics.) In order of expected difference, the 29 topics are 26, 59, 15, 41, 21, 35, 38, 63, 18, 80, 67, 99, 94, 20, 12, 78, 58, 87, 98, 17, 10, 57, 75, 40, 1, 33, 96, 7, 71.

Figure 6 shows the distribution of topic shares within each corpus for three of the 29 topics. Words associated with topics 21 and 28 are overrepresented in this subset of the Supreme Court corpus and words associated with topic 59 are more likely to be found among Appellate Court cases selected for review. Some of the different language appears to reflect the distinct procedures and actors associated with the different courts because several topics capture language specific to one court. But other

differences are more substantive. For example, words associated with constitutional law are overrepresented relative to the lower court decisions that was selected for review. The topic model is thus able to detect further agenda refinement by the Court after the selection of cases for review. It is not surprising that the cases selected for review occupy an intermediary status—more similar to the Supreme Court cases than the pool of all Appellate Court cases, but not yet indistinguishable from the Supreme Court cases. The ability of the topic model to locate the cases selected for review in this intermediary status provides additional confidence.

### **III. Differences and trends**

#### *A. Certiorari and topic winnowing*

Having established some basic validity of our topic model, we now turn to analyzing what it can tell us about the Step 2 in the process of constructing the Court’s agenda discussed above—the selection of cases for review. To undertake this analysis, we used a simple logistic regression model to predict certiorari. Observations in the model are the randomly selected set of Appellate Court decisions, coded “1” if they are included in the set of cases that that were selected for review. To examine whether certain subjects are associated with a greater or lower likelihood of being located within the set of Appellate Court decisions granted certiorari (and, by extension, the Supreme Court docket) we generate a group of variables made up of the feature vectors of topic concentrations for each decision, represented in terms of log odds (rather than probabilities) and further standardized to have a mean of zero and a standard deviation of 0.5. We add a set of indicator variables to control for each year. It is also natural to ask if the effects that we identify change over time, and so we add an interaction term between the year of the Appellate Court decision and the topic concentrations. We run three models (year effects; topics and year effects; topics, interactions, and year effects). The full results of the three models run with these variables are reported in Table 2.

Figure 7 reports the statistically significant variables from the full model ( $p < 0.05$ ), along with confidence intervals. With year interactions, there are six configurations of potential effects that we can

discover: topics can (1) increase likelihood or (2) decrease likelihood, and these tendencies can be (a) static, (b) increasing, or (c) decreasing. Roughly half (53 out of 100) of the topics are significant, either alone or in interaction with year. Of those, seventeen variables have year interactions of some sort. This ratio (17:53) indicates that the topics in the Court's docket are, to some degree, a moving target, with a substantial number of topics that are subject to temporal trends, but also that there is some stability to the selection process.

Topics that stably increase likelihood of being in the pool of selected cases (configuration 1a) during our study period include certain constitutional issues, such as speech (topic 76) and education (topic 95), topics that relate to interbranch relations, such as Congress (topic 21) and statutory interpretation (topic 35), and some substantive areas that appear to be of durable interest, such as antitrust (topic 23) and Native American affairs (topic 17). Two clusters of topics that stand out for having relatively stable reduced likelihood (configuration 2a) are state law related topics and criminal law related topics. The former includes family law (topic 1), trusts and estates (topic 34), negligence (topic 51), and contracts (topic 91). The criminal law family of topics includes narcotics (topic 3), surveillance (topic 47), conspiracy (topic 70), and perhaps to some degree automobiles (topic 16).

We identify two topics that tend to be selected for review more often than one would otherwise expect with an increase in this tendency over time (configuration 1b). These are the death penalty (topic 27) and state-federal relations (topic 40). For the death penalty topic, in a 1979 decision an increase in the proportion of words associated with the death penalty from one standard deviation below the mean to one standard deviation above the mean is associated with a 1.9% increase in the probability that the decision in question is selected for review. In 1995, by contrast, such an increase in the proportion of words associated with the topic is associated with a 2.6% increase in the probability that the decision is selected for review.

Two topics that have declining tendencies to increase likelihood (configuration 1c) are due process (topic 28) and discrimination (topic 31). For the discrimination topic, in a 1979 decision an

increase in the proportion of words associated with discrimination from one standard deviation below the mean to one standard deviation above the mean is associated with a 1.4% increase in the probability that the decision in question is selected for review. In 1995 such a shift in the proportion of words associated with discrimination does not increase the probability of review appreciably.

Two topics that decrease likelihood, with an increasing tendency in this direction (configuration 2b) are insurance (topic 9), sentencing (topic 19) and energy(topic 66). In the case of insurance, in a 1979 decision an increase in the proportion of words associated with insurance from one standard deviation below the mean to one standard deviation above the mean is associated with a 1.6% decrease in the probability that the decision in question is selected for review. In 1995 such a shift is associated with a 1.8% decrease in the probability that the decision in question is selected for review (a meaningful decrease given that the baseline odds against a case being selected have increased considerably between 1979 and 1995 due to a shrinking Supreme Court docket). Under-represented topics headed in the opposite direction (configuration 2c) include patents and invention (topic 61).

The preceding discussion clarifies that Step 2 in our simple model of agenda formation (in which a group of cases is selected from the pool of all Appellate Court cases) plays an important and dynamic role in forming the agenda of the Supreme Court. There is a winnowing process in which the subject matter of cases is a significant predictor of the likelihood that a case will end up being heard and decided by the Supreme Court. In this process, there are certain legal areas that are consistently favored, others that are consistently disfavored, and many in which predictive power has shifted over time. Some of the time trends have reinforced over-representation (such as the for death penalty cases) or under-representation (such as for insurance cases) while the predictive power of others topics has gradually declined (such as for discrimination).

The next section turns from Step 2 to Step3 in our simple model of agenda formation in which the Supreme Court transforms the cases before it into its own corpus of decisions. In particular, we will

examine the semantic distinctness of the Court over time in relation to the Appellate Courts in general, as well as with respect to the cases selected for review.

### *B. Growing distinctiveness of Supreme Court decisions*

A signal advantage of the topic model approach is that a topic model permits systematic comparison of arbitrary collections of decisions without the need for expensive time-consuming manual labeling. Even if consistent expert labels were available for all Appellate Court and Supreme Court decisions, the labels themselves may not be comparable.<sup>33</sup> And even if the expert labels used are commensurable, as is the case with the Songer U.S. Appeals Courts Database (which adopts a labeling scheme similar to the SCDB's "issue area" labeling), the concern arises that, as the labeling has been performed by different individuals in different years, it would be challenging to reproduce independently. A topic model of the entire body of decisions, by contrast, provides an unsupervised decomposition of each decision into a small number of topic weights (technically, these are mixture weights for the set of component probability distributions). While the decomposition requires different interpretation than a set of discrete labels, the ease with which the decomposition may be obtained and its ability to be independently reproduced count in its favor.

The ability to compare arbitrary sub-collections within the corpus of Appellate Court and Supreme Court decisions makes it possible to pose exceedingly general questions about the relationship between the language in the decisions in the two levels of the judiciary. Given an exhaustive, consistent "labeling" of Appellate Court and Supreme Court decisions in terms of topics, it is possible, for example, to ask whether or not the distribution of topics associated with Supreme Court decisions and/or Appellate Court decisions is changing in such way that it becomes easier to tell the two types of cases apart. An

---

<sup>33</sup> The problem of incommensurable labels confronted Rice (2014). One set of labels, the Policy Agendas (PA) labels, use 19 major issue categories. Rice provides as an example the PA labels for *Lawrence v. Texas*: "Civil Rights," "Minority Issues" and "Civil Liberties." The PA scheme differs substantially from the Spaeth/Supreme Court Database (SCDB) issue and issue area labels. The PA labels and the SCDB labels are the most widely used labeling schemes and it is not possible to assess the similarity of cases based on labels when the labeling scheme differs.

affirmative answer would be consistent with either growing specialization in non-substantive language of the courts (e.g., language describing the jurisdiction-specific procedures and customs) and/or a relative narrowing or expanding of the range of legal subjects in one of the sets of decisions over time. One way to test this claim is to ask how difficult, given only the topic weights for each decision, it is to tell a Supreme Court decision from an Appellate Court decision in a given year. If the ease of prediction increases, we can be confident that one (or both) of the collections of topics is becoming more distinct with respect to the other. One way to quantify "predictability" is (for each year) to hold out fifty percent of the Appellate Court and Supreme Court decisions, train a classification model (here, logistic regression) using the remaining decisions and then evaluate the model on the held-out fifty percent. This task is repeated many times, each time randomly sampling the fifty percent of cases which are held out.<sup>34</sup> If the model has an easier time predicting the (known) category of the decisions in the held-out set, we know that the distinctiveness of one or both of the different types of decisions is growing.

Figure 6, 7, and 8 display the results of this analysis. We first find that, logistic regression as a classification model performs reasonably well in predicting the difference between an Appellate Court decision and a Supreme Court decision—this result is in keeping with the findings in *Hypothesis B* above. More interestingly, we find that this result is growing over time. Accuracy is calculated as the sum of true positives and true negatives, divided by total cases, and precision is calculated as true positive over the sum of true positives and false positives. The center of the distribution of the accuracy of held out prediction starts at roughly 80% in the 1950s, but over time increases to nearly 100% in the 2000s. Precision undergoes a similar transition, with a rate centered near 80% in the 1950s increasing to above 95% by the 2000s. This result quite clearly indicates that language in the respective bodies of decisions is growing more distinctive, although the specific distinctive language in question might be different from year to year.

---

<sup>34</sup> Because the number of decisions in each year varies considerably—e.g., there are far more Federal Appellate decisions in 2000 than in 1960—we randomly sample year-specific corpora of equal sizes before performing the prediction task.

In our simple model of the agenda formation process introduced in Part I.A above, there were two processes of interest:  $S$ , which is the selection of a subset of cases from the total set of Appellate Court cases, and  $T$ , which “transformed” the subset of selected cases into a different set of Supreme Court opinions. Part III.A above discussed the first process ( $S$ ) in which cases are selected, identifying topics that are associated with increased or reduced likelihood of a case appearing in the set of cases selected for review (and time trends). Given the results reported in Figure 8, it is worth considering whether this process has become more important over time, in the sense of playing a greater role in distinguishing the agenda of the Supreme Court for the overall subject matter of the Appellate Courts. To investigate this question, we carry out the same logistic regression classifier analysis on two different corpora: the set of all Appellate Court cases, and the set of Appellate Court cases selected for review. If the Supreme Court is using its certiorari power more “aggressively” than in the past (in the sense of selecting cases that are more distinct from the pool of all Appellate Court cases), we should find that the performance of the classifier would increase over time. In fact, as reported in Figure 9, we are not able to generate that finding: the accuracy and precision of the classifier in the 2000s is essentially the same as during the 1950s, in stark contrast to the growth in both of those measures when comparing the Supreme Court and Appellate Court corpora.

This finding leads naturally to the conclusion that the decisions within the Supreme Court corpus must be becoming more distinct from the Appellate Court cases selected for review, which we confirm in the analysis presented in Figure 10. We find that for both measures, the performance of the classifier is improving over time, starting out with accuracy and precision centered at roughly 70% in the 1950s and increasing well over 90% by the 2000s. The lesson from this analysis is that, although the cases selected for review in recent years are no more distinct from the pool of all Appellate Court cases, the way that the Supreme Court analyzes and discusses the legal issues presented in those cases has grown increasingly idiosyncratic over time.

#### IV. “Distant Reading” the Law

Literary scholar and leading digital humanist Franco Moretti has coined the phrase “distant reading” to characterize the now growing practice of applying quantitative tools to the study of literary texts (Moretti 2013). As might be expected, this development has not been uncontroversial, and a lively debate has broken out in the field about the relative merits and demerits of quantitative ‘distant’ reading and qualitative ‘close’ reading. This discourse is reminiscent of the debates within the legal academy that broke out as quantitative tools from the social sciences began to work their way into law scholarship. Over the subsequent several decades legal academics from both camps largely have been able to reconcile themselves to, and find the value in, alternative methodologies.

Prior quantitative approaches to empirical legal studies continued to rely on ‘close’ human reading of texts—the goal has primarily been to standardize such readings, record them at scale, and then apply statistical tools to examine the results (Hall and Wright 2008). This paper takes the natural further step of applying new statistical tools to recently available electronic resources to directly treat text as data. The result is a new vantage point, even more ‘distant’ than prior empirical legal scholarship, that allows for the quantitative estimation of macro-level trends that are impossible to perceive for even the most dedicated group of human readers. This Part examines one such trend—the increasing idiosyncratic Supreme Court—and discusses the greater potential for topic models to contribute to empirical legal studies.

##### *A. The Increasingly “Uncourtlike” Supreme Court*

The analysis reported in Part III.B leads to the provocative conclusion that the Supreme Court has become more distinct from other federal courts, and specifically more distinct when controlling for the fact that it reviews a self-selected set of cases. Our interpretation of this finding is that the Supreme Court these days does not find itself with a set of legal issues that are any more different from the issues in the Appellate Court than in the past. However, the Supreme Court appears to be analyzing and discussing those issues in increasingly idiosyncratic ways. If the Appellate Courts are taken as the paradigmatic

baseline of how courts analyze and discuss cases (a controversial but defensible proposition), then the Court has become less “court-like” over time.

It is perhaps worth restating our empirical conclusion in more intuitive terms. Imagine a hypothetical law student, walking the halls of law library in the year 1954. This law student notices on the floor a few pages torn out a Federal Reporter, perhaps 1,000 words in all, but lacking direct information identifying the authoring court. This student sets himself or herself the task of trying to guess whether the opinion was written by the Supreme Court or an Appellate Court. If that student performed as well as our classifier, he or she would have a roughly 70% chance of being right. Transport our hypothetical student to the present day and he or she would almost certainly be able to identify the case, even if the page lying on the floor was from a case that had been selected for review by the Supreme Court.

It is not obvious that this is a normative problem, but it is not obvious that this is *not* a normative problem either. It certainly seems like something that is worth discussing and arguing over. The place of the Supreme Court and its relationship to the courts it supervises and the political branches is an endless source of discourse and controversy. It would require further analysis—empirical, political, and philosophical— to assess the normative weight of the results reported above. Certainly it would be helpful to have an explicit normative conception of the Supreme Court’s judicial role, and how that role relates to the way in which the Justices analyze and discuss the cases before it. Additional empirical estimates that provide a more fine grained analysis of how the Supreme Court differs would also be useful. There may be certain relatively harmless semantic quirks that the Supreme Court has developed, which do not indicate a major divergence in how Justices and judges approach cases. Alternatively, it may be entirely appropriate for the Supreme Court to depart from the Appellate Courts, given the its special status.

The contribution of the topic model approach, and distant reading more generally, to this inquiry is in the ability to formalize and identify trends and relationships that would be extremely difficult (or perhaps even impossible) to estimate through human reading. To return to the hypothetical law student, it

would be possible to hire research assistants and ask them to distinguish between randomly selected snippets of Supreme Court and Appellate Court cases, to determine whether they are better at classifying more recent cases than earlier cases. But it would be impossible to return to 1954 and ask a law student at that time to carry out this task, so the effect associated with the greater familiarity of a contemporary law student with contemporary law could not be eliminated. And, if trends are noisy, it might be infeasible to hire the necessary number of researchers to collect the amount of data necessary to identify them. The quantitative approach discussed above avoids this problem because it does not rely on human readers. More to the point, distant reading can help draw attention to and clarify questions that might not otherwise be asked.

### *B. Topic models and empirical legal studies.*

Taken together, the preceding analyses generate several useful insights. First, the process of agenda formation has measurable consequences for the subject matter of the Supreme Court corpus. Measured against a baseline of all Appellate Court cases, the opinions of the Supreme Court do not simply reflect the business of the federal courts. Rather, the Court's writings are concentrated on a smaller world of topics, with some topics receiving a great deal of attention in the Appellate Courts and relatively short shrift in the Supreme Court, and other topics receiving greater focus in the Supreme Court than they do below. These differences result in a corpus in the U.S. Reports that has a markedly different subject matter.

We also find that case selection is not the only mechanism of agenda formation for the Court. As noted above, there is some controversy in the empirical literature about the degree to which the Court engages in issue suppression or issue expansion after cases have been selected for review. The preceding analysis highlights the fact that the subject matter in the Court's opinions does not mirror the subject matter of the cases selected for review. After certiorari has been granted, there is a further process of winnowing in which the Court refines the subject matter of its decisions. The pool of selected cases is an

intermediary corpus—more closely aligned with the Court’s than all federal cases, but still including residual affinities from the larger pool from which it is drawn.

These first two insights are, to some extent, obvious, although there is value in formalizing and quantifying these effects. But, because they also confirm the ability of the topic model to identify meaningful semantic features of cases, we are able to undertake additional analysis of agenda formation described in Part I.A as involving two major processes, the selection of cases for review (*S*) and the transformation of the issues in those cases into Supreme Court decisions (*T*). First examining process *S*, we find a number of interesting trends in the topics that are associated with increased or decreased likelihood of being selected. We find that some topics have relatively static effects—consistently increasing or decreasing likelihood during our study period—while others have dynamic effects that change over time.

We also examine process *T* in which the Justices take the cases that are selected for review and turn them into Supreme Court opinions. We find that the effect of this process on the overall process is growing over time. Given an equally distinct set of cases selected for review, the Supreme Court now writes about those cases in an increasingly distinctive and idiosyncratic fashion. This distinctiveness is sufficiently great that it is easily identifiable, even with the radical reduction in dimensionality associated with a topic model, not to mention the loss of data associated with term frequency approaches that treat documents as bags of words (in which the *order* of the words in an opinion does not matter). This finding certainly raises important normative questions that are worthy of future exploration.

Finally, and perhaps most important, the preceding analysis validates topic modeling as a useful tool for understanding agenda formation on the Court, and engaging in empirical legal studies more generally. Quantitative analysis of law has traditionally been hampered by the lack of attractive mechanisms for estimating legal features of cases (Clark and Lauderdale 2010). Topic modeling provides a promising avenue to estimate legal variables because it is oriented toward the semantic content of cases, which is exactly where the law would be expected to show up. It also avoids some of the pitfalls of

human readers, including error, bias, and most important, time and attentional limits. By naïvely characterizing the semantic content of cases based on a very general set of assumptions about the structure of the corpus, topic models provide a quantitative and computationally tractable method to represent the text of the law. The corpus of the law—both the published case law in the state and federal reporters and other legal texts as well—is an enormous and rich dataset, and topic models provide an effective means of capturing important characteristics of that data that can be subjected to analysis. There is of course also nothing special about the U.S. Court system. At the very least, the topic modeling methodology as presented here could be immediately deployed on other English language legal text corpora. The British legal system, the EU’s European Court of Justice, or the decisions of the Dutch Supreme Court are natural next places to investigate with these techniques, each presenting their own particular issues worthy of study.

As for the U.S. system there are a number of potential research projects that topic modeling of the corpus of legal decisions could help facilitate. One straightforward question that we do not address is whether the pool of federal cases for which certiorari is petitioned differs from the pool of all federal cases—examining this question would identify the role that selection for petitioning plays in influencing the Court’s agenda. Dynamic topic models (Blei and Lafferty 2006) could be deployed to track the evolution of subject matter in the Court over time and relate that evolution to internal factors, such as the composition of the Court. Studies, such as Rice (2013) and Hurwitz (2006), that examine the Court’s influence on litigation, could be expanded to include more detailed decomposition of subject matter into a larger number of topics. In addition, the relationship between the Court’s agenda and external factors, including the congressional agenda, mass media political discussions, and economic variables, could be examined (cf., Pacelle 1991). More detailed analysis of the role of median, majority median, and author in setting the Court’s agenda could also be empowered, along the lines used in Lauderdale and Clark (2013). That study used citation as a means of identifying the influence of different Justices on the outcome of individual decisions—if the tendency to emphasize some subject matters at the expense of others could

similarly be mapped onto judicial preferences, then it could also be used as a measure of influence.

Further afield, topic models could potentially identify particular styles of judicial reasoning—such as use of legislative history—and then be used to examine trends in judicial approach and causal influences on those trends.

These projects fit into the broader schema of a “standardized package,” a framework that has proved effective in facilitating cross-disciplinary research and in this case, one in which an integration of computational studies, humanities, and social sciences could facilitate a productive line of empirical legal research. In this context, we reference the considerable body of literature devoted to studying the origins of successful scientific and industrial research programs (Knorr-Cetina 1999; Kuhn 1962; Saxenian 1994). In particular, in exploring the emergence of cancer research in the 1960s and 1970s, Fujimura attributed cancer research’s success at allowing researchers from distinct research communities to communicate and collaborate on concrete “do-able” problems to the existence of a “standardized package” that consisted of a scientific approach (molecular biology) and “a standardized set of technologies adopted by many members of multiple social worlds” (e.g., recombinant DNA technologies) (Fujimura 1992, 169; Fujimura 1988). The existence of a common theoretical reference point, along with a transportable set of technologies, allowed researchers from a range of disciplines to agree on a (temporarily) stable definition of cancer and construct a vibrant line of research.

Quantitative text analysis alongside the specific, transportable interface of topic modeling the corpus of Appellate Court and Supreme Court decisions, has the potential to make up a standardized package that can support an important line of empirical legal studies. More broadly, this standardized package would enable collaboration across a range of disciplines where communication is attenuated or infrequent, including computer science, computational linguistics, intellectual history, and social science. The importance of having a set of reproducible experiments in the development of research programs and scientific disciplines has been emphasized by many (Fujimura 1992; Kuhn 1962). A standardized package would contribute to the “stabilization of facts” vital for research across disciplines: a common corpus of

publicly available legal texts alongside standardized tools for decomposing those texts into measurable features that can be subjected to quantitative analysis. Topic models and computational text analysis more generally would allow researchers to “define a conceptual and technical work space which is less abstract, less ill-structured, less ambiguous, and less amorphous” (Fujimura 1992, 169).

## References

- Airoldi, Edoardo M., David M. Blei, Stephen E. Fienberg, and Eric P. Xing. 2008. "Mixed Membership Stochastic Blockmodels." *Journal of Machine Learning Research*. 9: 1981–2014.
- Ahmed, Amr, Mohamed Aly, Joseph Gonzalez, Shraavan Narayanamurthy, and Alexander J. Smola. 2012. "Scalable Inference in Latent Variable Models." In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, 123–132. WSDM '12. New York, NY, USA: ACM. doi:10.1145/2124295.2124312.
- Arewa, Olufunmilayo. 2012. "Open Access in a Closed Universe: Lexis, Westlaw, Law Schools and the Legal Information Market," *Lewis and Clark Law Review* 10: 797.
- Black, Ryan C. and Christina L. Boyd. 2012. "The Role of Law Clerks in the U.S. Supreme Court's Agenda-Setting Process." *American Politics Research*. 40 (1): 147–173.
- Black, Ryan C. and Ryan J. Owens. 2012. *The Solicitor General and the United States Supreme Court: Executive Branch Influence and Judicial Decisions*. Cambridge University Press.
- Black, Ryan C. and Ryan J. Owens. 2009. "Agenda Setting in the Supreme Court: The Collision of Policy and Jurisprudence." *The Journal of Politics*. 71 (3): 1062–1075.
- Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the Association for Computing Machinery*. 55(4): 77–84.
- Blei, David M. and John D. Lafferty. 2006. "Dynamic Topic Models." In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*. 3: 993–1022.
- Block, Sharon, and David Newman. "What, Where, When, and Sometimes Why: Data Mining Two Decades of Women's History Abstracts." *Journal of Women's History* 23 (1):81–109.
- Bonneau, Chris W., Thomas H. Hammond, Forrest Maltzman and Paul J. Wahlbeck. 2007. "Agenda Control, the Median Justice, and the Majority Opinion on the U.S. Supreme Court." *American Journal of Political Science*. 51 (4): 890–905.
- Boucher, Robert L., and Jeffrey A. Segal. 1995. "Supreme Court Justices as Strategic Decision Makers: Aggressive Grants and Defensive Denials." *The Journal of Politics*. 57 (3): 824.
- Brennan, William J. 1973. "The National Court of Appeals: Another Dissent." *University of Chicago Law Review*. 40: 473.
- Brenner, Saul. 2000. "Granting Certiorari by the United States Supreme Court: An Overview of the Social Science Studies." *Law Library Journal*. 92 (2): 193.
- Buntine, Wray. 2014. "hca 0.61." *Machine Learning Open Source Software*. Accessed September 20. <http://mloss.org/software/view/527/>
- Buntine, Wray L., and Swapnil Mishra. 2014. "Experiments with Non- Parametric Topic Models." In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 881– 890. KDD '14. New York, NY, USA: ACM. doi:10.1145/2623330.2623691.
- Caldeira, Gregory A., John R. Wright and Christopher J. W. Zorn. 1999. "Sophisticated Voting and Gate-Keeping in the Supreme Court." *Journal of Law, Economics & Organization*. 15 ( 3): 549-572.
- Caldeira Gregory A. and John R. Wright. 1990. "The Discuss List: Agenda Building in the Supreme Court." *Law & Society Review*. 24 ( 3): 807-836.

- Caldeira and Wright. 1988. "Organized Interests and Agenda Setting in the U.S. Supreme Court." *American Political Science Review*. 82 (4) 1109-1127.
- Cameron, Charles M., Jeffrey A. Segal and Donald Songer. 2000. "Strategic Auditing in a Political Hierarchy: An Informational Model of the Supreme Court's Certiorari Decisions." *The American Political Science Review*. 94 (1): 101-116.
- Carlson, Keith, Michael Livermore, and Daniel N. Rockmore. 2016. "A Quantitative Analysis of Writing Style on the U.S. Supreme Court" *Washington University Law Review*. 93(6): (forthcoming).
- Carrubba, Cliff, Barry Friedman, Andrew D. Martin, and Georg Vanberg. 2012. "Who Controls the Content of Supreme Court Opinions?" *American Journal of Political Science*. 56 (2): 400-412.
- Casella, George, and Edward I. George. "Explaining the Gibbs sampler." *The American Statistician* 46, no. 3 (1992): 167-174.
- Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information Processing Systems 22*, ed. Y. Bengio.
- Clark, Tom S and Jonathan P. Kastellec. 2013. "The Supreme Court and Percolation in the Lower Courts: An Optimal Stopping Model." *The Journal of Politics*. 75 (1): 150–168.
- Clark, Tom S. and Benjamin Lauderdale. 2010. "Locating Supreme Court Opinions in Doctrine Space." *American Journal of Political Science*. 54 (4): 871.
- Cordray, Margaret, and Richard Cordray. 2004. "The Philosophy of Certiorari: Jurisprudential Considerations in Supreme Court Case Selection." *Washington University Law Quarterly*. 82 (2): 389.
- Cordray, Margaret, and Richard Cordray. 2001. "The Supreme Court's Plenary Docket." *Washington & Lee Law Review*. 58: 737.
- Downs, Anthony. 1957. "An Economic Theory of Political Action in a Democracy." *Journal of Political Economy*. 65 (2): 135-150.
- Doyle, Gabriel, and Charles Elkan. 2009. "Accounting for Burstiness in Topic Models." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 281–288. ICML '09. New York, NY, USA: ACM. doi:10.1145/1553374.1553410.
- Epstein, Lee and Jack Knight. 1998. *The Choices Justices Make*. Washington, DC: CQ Press.
- Epstein, Lee, Jeffrey A. Segal and Timothy Johnson. 1996. "The Claim of Issue Creation on the U.S. Supreme Court." *The American Political Science Review*. 90 (4): 845-852
- Estreicher, Samuel, and John E. Sexton. 1984. "A Managerial Theory of the Supreme Court's Responsibilities: An Empirical Study." *New York University Law Review* 59: 681.
- Fujimura, Joan H. 1988. "The Molecular Biological Bandwagon in Cancer Research: Where Social Worlds Meet." *Social Problems* 35 (3) : 261–283. doi:10.2307/800622.
- Fujimura, Joan H. 1992. "Crafting Science: Standardized Packages, Boundary Objects, and Translation." In *Science as Practice and Culture*, ed. Andrew Pickering, 168–211.
- Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis*. 18 (1): 1–35.
- Grant, Emily, Scott A. Hendrickson and Michael A. Lynch. 2012. "The Ideological Divide: Conflict and the Supreme Court's Certiorari Decision." *Cleveland State Law Review*. 60:559.

- Harper, Fowler V. and Alan S. Rosenthal. 1950. "What the Supreme Court Did Not Do in the 1949 Term. An Appraisal of Certiorari." *University of Pennsylvania Law Review*. 99 (3): 293-325.
- Hall, David, Daniel Jurafsky, and Christopher D. Manning. 2008. "Studying the History of Ideas Using Topic Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: 363-71.
- Hall, Mark A. and Ronald F. Wright. 2008. "Systematic Content Analysis of Judicial Opinions." *California Law Review*. 96: 63-122.
- Hofmann, Thomas. 1999. "Probabilistic Latent Semantic Indexing." In *Proceedings of the Twenty-Second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-57. New York: ACM. doi:10.1145/312624.312649.
- Hurwitz, Mark S. 2006. "Institutional Arrangements and the Dynamics of Agenda Formation in the U.S. Supreme Court and Courts of Appeals." *Law & Policy*. 28 (3): 321-344.
- Hughes, James M., Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore. 2012. "Quantitative Patterns of Stylistic Influence in the Evolution of Literature." *Proceedings of the National Academy of Sciences*. 109(20): 7682.
- Katz, Daniel M., Michael J. Bammarito, Julie Seaman, Adam Candeub and Eugene Agichtein. 2011. "Legal N-Grams? A Simple Approach to Track the 'Evolution' of Legal Language." *Proceedings of JURIX 2011: The 24th International Conference on Legal Knowledge and Information Systems*, Vienna, 2011, <http://ssrn.com/abstract=1971953>.
- Knorr-Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge: Harvard University Press.
- Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.
- Lauderdale, Benjamin E. and Tom S. Clark. 2013. "Who Controls Opinion Content? Testing Theories of Authorship Using Case-Specific Preference Estimates for the US Supreme Court."
- Lax, Jeffrey R., and Charles M. Cameron. 2007. "Bargaining and Opinion Assignment on the U.S. Supreme Court." *Journal of Law, Economics & Organization*. 23(2): 276-302.
- Mak, Maxwell, Andrew H. Sidman, and Udi Sommer. 2013. "Is Certiorari Contingent on Litigant Behavior? Petitioners' Role in Strategic Auditing." *Journal of Empirical Legal Studies*. 10 (1) 54-75.
- Maltzman, Forrest, and Paul J. Wahlbeck. 1996. "Strategic Policy Considerations and Voting Fluidity on the Burger Court." *American Political Science Review*. 90(3): 581-92.
- Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. New York: Cambridge University Press.
- Martin, Andrew D. and Kevin M. Quinn, 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999." *Political Analysis*, 10: 134-153.
- Mimno, David. 2011. "Reconstructing Pompeian Households." In *Uncertainty in Artificial Intelligence*. Barcelona, Spain.
- Mimno, David. 2012. "Computational Historiography: Data Mining in a Century of Classics Journals." *ACM Journal of Computing in Cultural Heritage*. 5 (1): 3:1-3:19.
- McGuire, Kevin T. and Gregory A. Caldeira. 1993. "Lawyers, Organized Interests, and the Law of Obscenity: Agenda Setting in the Supreme Court." *American Political Science Review*. 87 (3): 717-726

- McGuire, Kevin T. and Barbara Palmer. 1995. "Issue Fluidity on the U.S. Supreme Court." *The American Political Science Review*. 89 (3): 691-702.
- McGuire, Kevin T. and Barbara Palmer. 1996. "Issues, Agendas, and Decision Making on the Supreme Court." *The American Political Science Review*. 90(4): 853-865.
- Moretti, Franco. 2013. *Distant Reading*. Brooklyn: Verso.
- Nigam, Kamal, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. 1999. "Text Classification from Labeled and Unlabeled Documents Using EM." In *Machine Learning*, 103–134.
- Pacelle, Richard L. Jr. 1991. *The Transformation of the Supreme Court's Agenda: From the New Deal to the Regan Administration*. Boulder: Westview Press.
- Palmer, Barbara. 1999. "Issue Fluidity and Agenda Setting on the Warren Court." *Political Research Quarterly*. 52:39. doi: 10.1177/106591299905200102.
- Perry, H.W. Jr. 1991. *Deciding to Decide: Agenda Setting in the United States Supreme Court*. Harvard University Press.
- Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2): 945–959.
- Provine, Doris Marie. 1980. *Case Selection in the United States Supreme Court*. Chicago: University of Chicago Press.
- Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209– 228. doi:10.1111/j.1540-5907.2009.00427.x.
- Rice, Douglas. 2014. "The Impact of Supreme Court Activity on the Judicial Agenda." *Law & Society Review* 48 (1): 63–90. doi:10.1111/lasr.12056.
- Revesz, Richard L. and Pamela S. Karlan. 1998. "Nonmajority Rules and the Supreme Court." *University of Pennsylvania Law Review*. 36: 1067.
- Rice, Douglas. 2014. "The Impact of Supreme Court Activity on the Judicial Agenda." *Law & Society Review*. 48 (1): 63-90.
- Rice, Douglas. 2012. "Measuring the Issue Content of Supreme Court Opinions through Probabilistic Topic Models." Paper presented at the 2012 Midwest Political Science Association Conference, April 13, 2012. <http://qssi.psu.edu/files/NF5Rice.pdf>.
- Riddell, Allen. 2014. "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models." In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, ed. Matt Erlin and Lynne Tatlock. Rochester, New York: Camden House: 91–114.
- Saxenian, AnnaLee. 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge: Harvard University Press.
- Schubert, Glendon A. 1958. "The Study of Judicial Decision-Making as an Aspect of Political Behavior." *American Political Science Review*. 52(4): 1007.
- Songer, Donald R. 1998. *United States Courts of Appeals Database Phase 1, 1925–1988*. ICPSR02086-V1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].
- Sorenson, Maron W., and Timothy R. Johnson. 2014. "Issue Resuscitation at Oral Argument on the US Supreme Court." Unpublished manuscript, March 23, 2014. [https://apw.polisci.wisc.edu/APW\\_Papers/Johnson.pdf](https://apw.polisci.wisc.edu/APW_Papers/Johnson.pdf). Tanenhaus, Joseph, Marvin Schick, Matthew Muraskin, and Daniel

- Rosn. 1963. "The Supreme Court's Jurisdiction: Cue Theory." In *Judicial Decision-Making*, ed. Glendon Schubert. New York: Free Press.
- Spaeth, Harold, J. Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh (2014). *Supreme Court Database, Version 2014 Release 01*. URL: <http://Supremecourtdatabase.org>.
- Teh, Yee Whye. 2006. "A Hierarchical Bayesian Language Model Based on Pitman-Yor Processes." In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 985–992. ACL-44. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1220175.1220299.
- Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101 (476): 1566–1581.
- Thompson, David C., and Melanie F. Wachtell. 2009. "An Empirical Analysis of Supreme Court Certiorari Petition Procedures: The Call for Response and the Call for the Views of the Solicitor General." *George Mason Law Review*. 16 (2): 237.
- Ulmer, Sidney S. 1984. "The Supreme Court's Certiorari Decisions: Conflict as a Predictive Variable." *American Political Science Review*. 78 (4): 901.
- Ulmer, Sidney S. 1982. "Issue Fluidity in the United States Supreme Court." In *Supreme Court Activism and Restraint*, eds. Charles M. Lamb and Stephen C. Halpern.
- Ulmer, Sidney S., William Hintze and Louise Kirklosky. 1972. "The Decision to Grant or Deny Certiorari: Further Consideration of Cue Theory." *Law & Society Review*. 6 (4): 637-644.
- Yates, Jeff, Damon M. Cann, and Brent D. Boyea. 2013. "Judicial Ideology and the Selection of Disputes for U.S. Supreme Court Adjudication." *Journal of Empirical Legal Studies*. 10 (4): 847–865.
- Zorn, Christopher J. W. "U.S. Government Litigation Strategies in the Federal Appellate Courts." *Political Research Quarterly*. 55 (1): 145-166.
- Zhang, Aonan and John Paisley. 2015. "Markov Mixed Membership Models." *Proceedings of The 32nd International Conference on Machine Learning*, 475–483.

## Tables

**Table 1**

Topic	Label	Characteristic Words
0	Labor	union board labor employees employer nlr company bargaining relations national local act unfair order strike respondent practice cir work unit unions workers
1	Family 1	ms mrs did told husband time testified asked sexual home stated fact received mother daughter claim record office day appellant later return woman given leave left
2	Elections	election political party candidates candidate campaign parties primary elections contributions ballot office expenditures voters state vote buckley amendment
3	Narcotics	united drug cocaine government cir defendant conspiracy evidence drugs marijuana possession distribute district agent narcotics intent heroin agents
4	Immunity	immunity officers officer official police law county qualified officials city conduct rights liability constitutional state force sheriff actions established
5	Prison-conditions	prison inmates inmate prisoner prisoners officials confinement conditions security jail amendment district rights correctional disciplinary eighth
6	Procedure 1	motion district judgment appeal rule order filed summary party appeals judge final fed notice rules motions cir denied procedure granted trial discretion dismiss
7	Medical	dr medical hospital mental treatment health care patient drug expert patients physician condition doctor blood opinion mentally testing drugs disease
8	Crim pro 1	trial defendant plea guilty indictment united jeopardy criminal double prosecution government charges conviction charge right judge speedy motion second clause
9	Insurance	insurance policy company insured coverage insurer life ins policies liability loss judgment mutual business american accident insurers companies contract
10	Standard of proof	evidence record fact finding findings burden standard proof support facts reasonable trial did testimony issue sufficient substantial district conclusion
11	Crim law	offense crime criminal conviction united felony convicted defendant intent firearm person offenses crimes possession convictions law committed conduct charged

Topic	Label	Characteristic Words
12	6th 7th cir.	ohio illinois ill michigan chicago state oklahoma indiana tennessee kentucky wisconsin colorado mich martin defendant law taylor sixth tenn county supreme
13	Procedure 2	district order injunction relief decree preliminary orders injunctive pending courts judge proceedings action stay merits jurisdiction remedy parties appeal
14	Tax	tax income commissioner taxpayer revenue internal irs taxes year code paid return taxable deduction taxpayers years section payment united expenses business
15	Precedent 1	cir district id did issue circuit does omitted review claim ct appeal decision argument argues quoting analysis supreme claims second ii holding based held conclude
16	Automobile	car vehicle truck driver automobile ford stop motor vehicles highway traffic road driving cars drivers defendant license passenger trailer border miles van
17	Evidence	evidence testimony trial defendant witness rule witnesses cross defense united statement statements examination jury testified prior cir error testify
18	Precedent 2	ct ed supra cases denied held united cert id cf supreme question decisions opinion issue decision does dissenting note judgment view recognized supp course rev
19	Sentencing	sentencing sentence district defendant guidelines offense united level guideline departure range criminal conduct enhancement months judge cir government
20	Federal-powers	power constitution congress law authority judicial president constitutional united powers clause article executive courts art government laws general exercise
21	Congress	act congress cong legislative rep sess history committee section stat house senate congressional report provisions federal statute intended provision
22	Local govt	city ordinance housing public property park county area building council municipal town zoning use local project facilities home hud construction district
23	Antitrust	antitrust market price competition act sherman trade prices business united laws competitive clayton conspiracy violation corp power competitors effect
24	Confessions	confession police miranda right defendant statements rights interrogation statement amendment voluntary accused waiver arrest questioning united suspect

Topic	Label	Characteristic Words
25	Indian	indian tribe lands united water reservation indians tribal land river rights tribes state treaty jurisdiction act waters secretary interior federal non alaska
26	Briefs	respondent petitioner appeals id petitioners respondents ante opinion ibid supra pp app judgment brief question did does decision held dissenting certiorari
27	Death-penalty	death penalty capital sentencing murder sentence state mitigating evidence circumstances punishment jury aggravating defendant life supreme eighth phase crime
28	Due-process	process state amendment right constitutional protection clause rights fourteenth equal constitution statute liberty persons does person unconstitutional law
29	Family 2	child children abortion parents state family juvenile parent minor statute mother father woman foster pregnancy doe age custody health life roe parental decision
30	Disability	disability alj claimant benefits secretary work evidence disabled medical ada impairment security decision social substantial dr claim pain perform condition
31	Discrimination	discrimination racial race white black discriminatory equal women minority district title impact seniority based blacks vii basis applicants sex selection group
32	Environment	epa environmental water agency permit waste air project corps site forest administrator nuclear pollution impact control clean proposed nepa national quality act
33	Aviation	appellants air convention aircraft airport flight airlines international united faa airline aviation american article travel plane contend transportation
34	Trusts-estates	estate trust death wife decedent spouse husband property divorce law beneficiary marriage trustee deceased community mrs marital support children gift trusts
35	Stat interp 1	statute language congress interpretation statutory meaning does provision term section construction definition intended united act intent reading used phrase
36	Military	military army war navy service united force security defense national forces armed martial civilian guard air duty personnel base discharge jurisdiction officer
37	Banks	bank loan credit banks national account funds loans deposit savings note money check bonds financial trust banking accounts fdic bond payment security checks

Topic	Label	Characteristic Words
38	Stat interp 2	sec section shall act provides person provisions cir pursuant subsection united provided provision title district circuit sections does statute ct judge
39	Mining	land property lease coal use oil taking compensation mining owner lands mineral value public leases right condemnation acres gas just owners takings production
40	State federal	state federal law california cal statute laws supreme code courts pennsylvania ann stat statutes rev arizona commonwealth county common question local supp pa
41	Precedent 1	cases does majority supra fact course particular make need example opinion courts simply view point way reason important approach likely different analysis id
42	Fees	fees attorney fee award attorneys costs district litigation party prevailing reasonable expenses awarded legal counsel services sanctions discretion plaintiff
43	Securities	securities exchange fraud rule act stock investment sec sale purchase offer security market commission information trading investors duty transaction price
44	Communications	commission fcc cable communications television service public access broadcast broadcasting station order telephone license radio local telecommunications
45	Valuation	cost value costs rate total market percent method year increase rates based price pay basis used reasonable average time fair use paid million years excess formula
46	Search warrant	search warrant fourth united officers amendment arrest probable police cause officer seizure evidence reasonable privacy searches seized agents circumstances
47	Surveillance	king telephone gambling united conversations order electronic surveillance wiretap club tape tapes use application communications conversation government
48	Juries	jury trial verdict defendant instruction error judge instructions evidence jurors juror reasonable defense charge doubt prosecutor instructed harmless law
49	Self incrimi- nation	grand privilege criminal united jury contempt witness amendment fifth prosecution testimony government subpoena self attorney incrimination immunity evidence
50	Jurisdiction	jurisdiction federal action district claims courts claim suit forum jurisdictional defendant subject matter law parties judgment plaintiff party complaint

Topic	Label	Characteristic Words
51	Negligence	negligence plaintiff liability defendant duty accident injury cause negligent law caused company injuries care judgment liable cir safety warranty use condition
52	Sentencing	sentence parole release probation imprisonment term imposed years time sentences punishment criminal united fine sentenced revocation prison year bail
53	Ad law boards	board review hearing administrative decision judicial agency proceedings action process procedures order determination final proceeding procedure appeals
54	Deadlines	time period notice date limitations filed days statute year day filing years delay months december action october prior september january file june august april
55	???	said question shall opinion mr ed cases supra think page ct united fact present time course act make right held view general effect involved matter reason trial taken
56	Employ discrim	discrimination employment title vii plaintiff eoc employer employee age position employees claim harassment discharge adea judgment district retaliation job
57	Ad law regs	regulations secretary agency regulation authority safety act standard standards requirements fed health reg regulatory administrative requirement general
58	Police test	police evidence testified car identification house robbery officer time man gun testimony did saw told officers door men trial room left apartment said victim
59	Appeals	appellant united district cir circuit judge appellee appeals atty appeal judges denied washington asst brief affirmed john defendant plaintiff judgment america
60	Mails	service board postal mail united local veterans letter classification selective registrant office induction claim appeal form registration conscientious
61	Patent-invention	patent claim claims invention infringement art patents prior use said fed application device process corp means patented district claimed patentee machine cir
62	Corporations	corporation stock company business corporate corporations shares partnership control assets capital directors shareholders owned merger management companies
63	Procedure filings	defendants plaintiffs plaintiff complaint claim district claims defendant appellees cir action alleged appellants judgment summary allegations motion

Topic	Label	Characteristic Words
64	Press	public press publication news newspaper times amendment statements article media false publicity new libel published newspapers malice sullivan defamation
65	Tax 2	tax state taxes exemption taxation exempt property income business clause revenue imposed use taxing assessment sales collection burden does pay exemptions
66	Gas-power-oil	commission gas power rate order natural ferc public company rates federal price service energy utility electric fpc pipeline oil act utilities customers sales new
67	Habeas	petitioner state habeas petition federal corpus claim writ relief conviction claims district review appeal supreme courts constitutional denied appeals
68	Employ	employees employee work employment employer compensation pay workers act time labor job wages wage employed hours working employers leave benefits paid position
69	Erisa	plan benefits erisa fund pension benefit retirement plans employee trust fiduciary employer employees trustees contributions funds health participants terms
70	Conspiracy	united conspiracy count government counts indictment defendant fraud cir evidence false money charged defendants scheme acts criminal violation conviction jury
71	Ineffective assist	counsel trial attorney defendant right defense assistance hearing judge lawyer representation criminal client prejudice appointed ineffective sixth williams
72	Rail	commission carrier railroad carriers transportation icc service rates interstate commerce line rail rate railroads act freight pacific public motor railway
73	Maritime	vessel maritime ship admiralty act cargo vessels jones marine seaman port waters owner law sea barge shipowner shipping crew boat coast liability longshoremen
74	Sales	sales sale products goods business sold product price trade food milk order retail sell selling market customers company tobacco purchase marketing seller
75	8th cir.	district missouri johnson kansas iowa st arkansas mo miller minnesota eighth nebraska louis circuit city dakota state ark minn hall judge submitted nelson morris
76	Speech	speech amendment public government free content expression regulation protected commercial rights statute forum conduct advertising activity does

Topic	Label	Characteristic Words
77	Federal-powers 2	commerce state interstate pre clause congress regulation federal act power carolina local regulate statute preemption laws emption north activity liquor effect
78	5th cir.	texas florida louisiana tex georgia virginia state alabama brown la mississippi fla fifth ga west county va law houston supreme ala clark smith code southern jackson
79	Obscene-material	sexual davis obscene material obscenity adult conduct young statute film materials pornography amendment films child sexually sex books standards community
80	Testimony	did time smith stated fact told received judge testified record later denied asked cir evidence appellant given facts day march failure knew said claim january
81	Labor arbitration	arbitration agreement union bargaining collective contract labor dispute parties arbitrator agreements grievance award disputes local duty clause employees
82	Damages	damages award liability law punitive injury claim plaintiff recovery tort action damage loss claims recover liable conduct compensatory awarded judgment
83	Political orgs	members committee association member organization party activities membership bar communist organizations group american local action associations political
84	Property	property lien sale transfer forfeiture title judgment real law right rights possession mortgage owner notice liens proceeds assignment value ownership subject
85	Immigration	united foreign alien immigration deportation aliens ins country bia attorney general petitioner naturalization citizenship persecution asylum application
86	Political voting	district voting county districts state vote plan population political election voters minority elections majority general rights census legislature black
87	1st 2d cir.	new york city jersey judge state massachusetts mass law second boston connecticut supp island counsel docket brief decided southern division mckinney rhode
88	Bankruptcy	bankruptcy debtor trustee creditors claim creditor chapter debt code claims plan estate petition secured bankrupt reorganization discharge filed order debtors
89	Entitlements	program services benefits funds secretary assistance federal programs title social department care payments health reimbursement medicaid medicare security

Topic	Label	Characteristic Words
90	Testimony 2	mr judge did said know yes don record asked going think say told counsel time just testimony statement question testified want right questions people like didn sir
91	Contracts	contract agreement parties contracts party terms breach agreements contractor agreed settlement payment law clause construction contractual letter
92	Foia-disclosure	information disclosure documents records report agency public investigation discovery department access reports request confidential exemption material
93	Copyright	copyright use mark trademark trade work infringement confusion product works used act plaintiff protection corp law marks registration rights lanham products
94	Precedent	rule law majority decision supreme opinion cases judge application new rules courts circuit applied issue panel apply banc decisions question doctrine effect en
95	Education	school religious schools education board students district public student religion university state educational church establishment children teachers clause
96	Federal claims	government united federal act claims claim private district public general department governmental exception islands tucker ftca cl agency virgin customs supp
97	Class-action	class standing action injury members plaintiffs settlement certification claims interests litigation individual parties party suit rule challenge plaintiff
98	Sovereign immunity	action federal private rights suit civil act immunity congress right state amendment remedy sovereign relief suits actions remedies statute eleventh courts cause
99	Sup. ct.	opinion cite id dissenting tion ing united ante stevens supra breyer syllabus concurring ment ibid thomas circuit appeals scalia does kennedy certiorari judgment

Table 1: Most probable words in the 100-topic model of the Supreme and Appellate Court decisions. Extremely common words (“stopwords”) are not displayed. Categorizations are based on the authors’ subjective judgements, for illustrative purposes. Not all topics were easily identifiable with standard legal categories.

**Table 2:** Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision’s ‘prior history’). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
year_intercept1951	-1.18* (0.17)	-1.04* (0.19)	-1.40* (0.24)
year_intercept1952	-1.05* (0.15)	-0.82* (0.17)	-1.19* (0.22)
year_intercept1953	-1.06* (0.15)	-0.93* (0.17)	-1.18* (0.21)
year_intercept1954	-1.22* (0.15)	-1.09* (0.17)	-1.34* (0.21)
year_intercept1955	-1.02* (0.13)	-0.83* (0.16)	-1.03* (0.19)
year_intercept1956	-0.90* (0.13)	-0.83* (0.15)	-1.13* (0.18)
year_intercept1957	-0.97* (0.13)	-0.80* (0.15)	-0.98* (0.19)
year_intercept1958	-0.87* (0.12)	-0.83* (0.14)	-1.02* (0.17)
year_intercept1959	-0.93* (0.13)	-0.76* (0.15)	-0.95* (0.17)
year_intercept1960	-1.14* (0.13)	-1.11* (0.14)	-1.37* (0.17)
year_intercept1961	-1.26* (0.14)	-1.37* (0.16)	-1.62* (0.19)
year_intercept1962	-1.28* (0.14)	-1.33* (0.16)	-1.54* (0.17)
year_intercept1963	-1.46* (0.14)	-1.52* (0.16)	-1.68* (0.18)
year_intercept1964	-1.29* (0.14)	-1.46* (0.16)	-1.57* (0.17)
year_intercept1965	-1.38* (0.13)	-1.38* (0.15)	-1.54* (0.16)
year_intercept1966	-1.58* (0.14)	-1.62* (0.15)	-1.73* (0.16)
year_intercept1967	-1.28* (0.12)	-1.25* (0.14)	-1.33* (0.15)

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
year_intercept1968	-1.41* (0.12)	-1.49* (0.14)	-1.56* (0.15)
year_intercept1969	-1.70* (0.13)	-1.92* (0.14)	-2.01* (0.15)
year_intercept1970	-1.85* (0.13)	-2.05* (0.14)	-2.12* (0.15)
year_intercept1971	-1.77* (0.12)	-2.05* (0.14)	-2.11* (0.14)
year_intercept1972	-1.82* (0.12)	-2.25* (0.14)	-2.30* (0.14)
year_intercept1973	-1.25* (0.11)	-1.70* (0.13)	-1.72* (0.13)
year_intercept1974	-1.24* (0.11)	-1.78* (0.12)	-1.79* (0.12)
year_intercept1975	-1.50* (0.12)	-2.03* (0.13)	-2.06* (0.13)
year_intercept1976	-1.44* (0.12)	-1.92* (0.13)	-1.93* (0.13)
year_intercept1977	-1.42* (0.11)	-2.09* (0.13)	-2.08* (0.13)
year_intercept1978	-1.45* (0.11)	-2.01* (0.13)	-2.00* (0.13)
year_intercept1979	-1.53* (0.11)	-2.17* (0.12)	-2.15* (0.12)
year_intercept1980	-1.42* (0.11)	-2.12* (0.12)	-2.07* (0.12)
year_intercept1981	-1.15* (0.09)	-1.84* (0.11)	-1.81* (0.11)
year_intercept1982	-1.66* (0.10)	-2.36* (0.12)	-2.33* (0.12)
year_intercept1983	-1.56* (0.10)	-2.23* (0.12)	-2.21* (0.12)
year_intercept1984	-1.62* (0.10)	-2.24* (0.12)	-2.21* (0.12)
year_intercept1985	-1.59* (0.10)	-2.24* (0.11)	-2.21* (0.11)
year_intercept1986	-1.71* (0.10)	-2.35* (0.11)	-2.33* (0.11)

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
year_intercept1987	(0.10) -1.86*	(0.12) -2.58*	(0.12) -2.57*
year_intercept1988	(0.11) -1.85*	(0.13) -2.52*	(0.13) -2.52*
year_intercept1989	(0.11) -1.98*	(0.12) -2.46*	(0.12) -2.48*
year_intercept1990	(0.11) -1.82*	(0.12) -2.36*	(0.12) -2.39*
year_intercept1991	(0.11) -2.01*	(0.12) -2.34*	(0.12) -2.41*
year_intercept1992	(0.11) -2.17*	(0.13) -2.61*	(0.13) -2.69*
year_intercept1993	(0.12) -2.46*	(0.14) -2.77*	(0.14) -2.86*
year_intercept1994	(0.14) -2.30*	(0.15) -2.72*	(0.15) -2.79*
year_intercept1995	(0.13) -2.02*	(0.14) -2.52*	(0.15) -2.59*
year_intercept1996	(0.11) -2.13*	(0.13) -2.55*	(0.14) -2.69*
year_intercept1997	(0.12) -1.98*	(0.14) -2.30*	(0.14) -2.45*
year_intercept1998	(0.12) -2.34*	(0.14) -2.91*	(0.15) -3.14*
year_intercept1999	(0.14) -2.13*	(0.16) -2.43*	(0.17) -2.65*
year_intercept2000	(0.14) -2.08*	(0.16) -2.19*	(0.17) -2.54*
year_intercept2001	(0.13) -2.27*	(0.15) -2.28*	(0.17) -2.66*
year_intercept2002	(0.14) -2.14*	(0.16) -2.57*	(0.18) -2.83*
year_intercept2003	(0.14) -2.20*	(0.16) -2.83*	(0.17) -3.13*
year_intercept2004	(0.14) -2.02*	(0.16) -2.57*	(0.18) -2.87*
	(0.13)	(0.15)	(0.17)

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
year_intercept2005	-2.50*	-2.90*	-3.23*
	(0.16)	(0.18)	(0.20)
year_intercept2006	-2.16*	-2.57*	-2.89*
	(0.15)	(0.17)	(0.19)
Topic000.labor		0.01	-0.05
		(0.04)	(0.04)
Topic001.family_1		-0.11*	-0.10*
		(0.04)	(0.04)
Topic002.elections		0.04	0.06
		(0.04)	(0.04)
Topic003.narcotics		-0.20*	-0.16*
		(0.05)	(0.05)
Topic004.immunity		-0.03	-0.02
		(0.04)	(0.04)
Topic005.prison.conditions		-0.06	-0.04
		(0.04)	(0.04)
Topic006.procedure_1		-0.36*	-0.37*
		(0.04)	(0.04)
Topic007.medical		0.08*	0.09*
		(0.04)	(0.04)
Topic008.crim_pro_1		-0.08	-0.06
		(0.04)	(0.04)
Topic009.insurance		-0.18*	-0.19*
		(0.04)	(0.04)
Topic010.standard_of_proof		-0.26*	-0.26*
		(0.04)	(0.04)
Topic011.crim_law		-0.11*	-0.09*
		(0.04)	(0.04)
Topic012.6th_7th_Cir.		0.03	0.04
		(0.04)	(0.04)
Topic013.procedure_2		-0.01	-0.02
		(0.04)	(0.04)
Topic014.tax		-0.03	0.01
		(0.04)	(0.04)
Topic015.precedent_1		-0.07	0.05
		(0.05)	(0.06)
Topic016.automobile		-0.13*	-0.09*

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
		(0.04)	(0.04)
Topic017.evidence		-0.21*	-0.21*
		(0.04)	(0.05)
Topic018.precedent_2		0.67*	0.62*
		(0.05)	(0.05)
Topic019.sentencing		-0.32*	-0.22*
		(0.05)	(0.06)
Topic020.federal.powers		0.13*	0.15*
		(0.03)	(0.03)
Topic021.congress		0.41*	0.40*
		(0.04)	(0.04)
Topic022.local_govt		-0.09*	-0.07
		(0.04)	(0.04)
Topic023.antitrust		0.15*	0.16*
		(0.04)	(0.04)
Topic024.confessions		-0.00	0.02
		(0.04)	(0.04)
Topic025.indian		0.08*	0.09*
		(0.04)	(0.04)
Topic026.briefs		0.03	0.04
		(0.04)	(0.04)
Topic027.death.penalty		0.22*	0.20*
		(0.04)	(0.04)
Topic028.due.process		0.14*	0.14*
		(0.04)	(0.04)
Topic029.family_2		-0.02	-0.01
		(0.04)	(0.04)
Topic030.disability		-0.13*	-0.09
		(0.05)	(0.05)
Topic031.discrimination		0.13*	0.14*
		(0.04)	(0.04)
Topic032.environment		-0.05	-0.03
		(0.04)	(0.04)
Topic033.aviation		-0.03	-0.02
		(0.04)	(0.04)
Topic034.trusts.estates		-0.15*	-0.09*
		(0.04)	(0.04)

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
Topic035.stat_interp_1		0.21*	0.22*
		(0.04)	(0.04)
Topic036.military		0.03	-0.00
		(0.04)	(0.04)
Topic037.banks		-0.08	-0.05
		(0.04)	(0.04)
Topic038.stat_interp_2		0.05	0.04
		(0.04)	(0.04)
Topic039.mining		-0.07	-0.05
		(0.04)	(0.04)
Topic040.state_federal		0.17*	0.18*
		(0.04)	(0.04)
Topic041.precedent_1		0.27*	0.31*
		(0.04)	(0.04)
Topic042.fees		-0.01	0.01
		(0.04)	(0.04)
Topic043.securities		-0.04	-0.03
		(0.04)	(0.04)
Topic044.communications		-0.09*	-0.07
		(0.04)	(0.04)
Topic045.valuation		0.05	0.05
		(0.04)	(0.04)
Topic046.search_warrant		0.01	0.01
		(0.04)	(0.04)
Topic047.surveillance		-0.12*	-0.11*
		(0.04)	(0.04)
Topic048.juries		-0.03	-0.01
		(0.04)	(0.04)
Topic049.self.incrimination		0.10*	0.06
		(0.04)	(0.04)
Topic050.jurisdiction		-0.17*	-0.16*
		(0.04)	(0.04)
Topic051.negligence		-0.26*	-0.25*
		(0.04)	(0.05)
Topic052.sentencing		-0.09*	-0.08
		(0.04)	(0.04)
Topic053.ad_law_boards		-0.14*	-0.15*

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
		(0.04)	(0.04)
Topic054.deadlines		-0.14*	-0.14*
		(0.04)	(0.04)
Topic055		0.02	0.02
		(0.05)	(0.05)
Topic056.employ_discrim		-0.14*	-0.09
		(0.04)	(0.05)
Topic057.ad_law_regs		-0.03	-0.00
		(0.04)	(0.04)
Topic058.police_test		-0.11*	-0.09*
		(0.04)	(0.05)
Topic059.appeals		0.00	-0.01
		(0.05)	(0.05)
Topic060.mails		-0.10*	-0.12*
		(0.04)	(0.04)
Topic061.patent.invention		-0.15*	-0.10*
		(0.04)	(0.04)
Topic062.corporations		0.01	0.04
		(0.04)	(0.04)
Topic063.procedure_filings		-0.06	-0.03
		(0.04)	(0.04)
Topic064.press		0.02	0.03
		(0.04)	(0.04)
Topic065.tax_2		0.07	0.08*
		(0.04)	(0.04)
Topic066.gas.power.oil		-0.08*	-0.13*
		(0.04)	(0.04)
Topic067.habeas		0.05	0.05
		(0.04)	(0.04)
Topic068.employ		0.02	0.02
		(0.04)	(0.04)
Topic069.ERISA		-0.02	0.00
		(0.04)	(0.04)
Topic070.conspiracy		-0.09*	-0.10*
		(0.04)	(0.04)
Topic071.ineffective_assist		-0.04	-0.03
		(0.04)	(0.04)

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
Topic072.rail		-0.06 (0.04)	-0.09* (0.04)
Topic073.maritime		0.00 (0.04)	-0.05 (0.05)
Topic074.sales		-0.08* (0.04)	-0.07 (0.04)
Topic075.8th_Cir.		-0.16* (0.04)	-0.15* (0.04)
Topic076.speech		0.09* (0.04)	0.11* (0.04)
Topic077.federal.powers_2		0.18* (0.04)	0.19* (0.04)
Topic078.5th_Cir.		-0.08* (0.04)	-0.06 (0.04)
Topic079.obscene.material		-0.02 (0.04)	-0.00 (0.04)
Topic080.testimony		-0.28* (0.04)	-0.30* (0.04)
Topic081.labor_arbitration		-0.00 (0.04)	0.01 (0.04)
Topic082.damages		0.08* (0.04)	0.08* (0.04)
Topic083.political_orgs		0.10* (0.04)	0.07 (0.04)
Topic084.property		-0.06 (0.04)	-0.03 (0.04)
Topic085.immigration		-0.01 (0.04)	0.00 (0.04)
Topic086.political_voting		-0.08* (0.04)	-0.07 (0.04)
Topic087.1st_2d_Cir.		0.08* (0.04)	0.05 (0.04)
Topic088.bankruptcy		-0.13* (0.04)	-0.13* (0.05)
Topic089.entitlements		-0.01 (0.04)	0.00 (0.04)
Topic090.testimony_2		-0.10* (0.04)	-0.09* (0.04)

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
		(0.04)	(0.04)
Topic091.contracts		-0.27*	-0.26*
		(0.04)	(0.04)
Topic092.FOIA.disclosure		0.03	0.03
		(0.04)	(0.04)
Topic093.copyright		-0.16*	-0.16*
		(0.05)	(0.05)
Topic094.precedent		0.61*	0.63*
		(0.04)	(0.04)
Topic095.church.state.schools		0.11*	0.12*
		(0.04)	(0.04)
Topic096.federal.claims		0.13*	0.14*
		(0.04)	(0.04)
Topic097.class.action		-0.02	-0.00
		(0.04)	(0.04)
Topic098.sovereign.immunity		0.14*	0.14*
		(0.04)	(0.04)
Topic099.Sup._Ct.		-0.14*	-0.30*
		(0.06)	(0.07)
Topic000.labor:year			-0.15
			(0.09)
Topic001.family_1:year			0.14
			(0.09)
Topic002.elections:year			0.02
			(0.09)
Topic003.narcotics:year			0.05
			(0.10)
Topic004.immunity:year			0.06
			(0.09)
Topic005.prison.conditions:year			0.05
			(0.10)
Topic006.procedure_1:year			-0.15
			(0.08)
Topic007.medical:year			-0.01
			(0.08)
Topic008.crim_pro_1:year			-0.05
			(0.09)

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
Topic009.insurance:year			-0.25* (0.09)
Topic010.standard_of_proof:year			0.14 (0.08)
Topic011.crim_law:year			-0.06 (0.09)
Topic012.6th_7th_Cir.:year			-0.18* (0.08)
Topic013.procedure_2:year			0.02 (0.08)
Topic014.tax:year			0.08 (0.08)
Topic015.precedent_1:year			0.17 (0.13)
Topic016.automobile:year			0.14 (0.08)
Topic017.evidence:year			0.07 (0.09)
Topic018.precedent_2:year			-0.23* (0.09)
Topic019.sentencing:year			-0.29* (0.12)
Topic020.federal.powers:year			0.08 (0.07)
Topic021.congress:year			0.03 (0.07)
Topic022.local_govt:year			0.11 (0.08)
Topic023.antitrust:year			-0.12 (0.08)
Topic024.confessions:year			-0.02 (0.09)
Topic025.indian:year			0.13 (0.08)
Topic026.briefs:year			-0.04 (0.08)
Topic027.death.penalty:year			0.32*

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
			(0.09)
Topic028.due.process:year			-0.16*
			(0.08)
Topic029.family_2:year			0.02
			(0.09)
Topic030.disability:year			0.04
			(0.10)
Topic031.discrimination:year			-0.21*
			(0.09)
Topic032.environment:year			0.06
			(0.09)
Topic033.aviation:year			0.05
			(0.08)
Topic034.trusts.estates:year			0.11
			(0.08)
Topic035.stat_interp_1:year			0.07
			(0.08)
Topic036.military:year			-0.17*
			(0.08)
Topic037.banks:year			0.17
			(0.09)
Topic038.stat_interp_2:year			-0.11
			(0.08)
Topic039.mining:year			-0.09
			(0.08)
Topic040.state_federal:year			0.30*
			(0.08)
Topic041.precedent_1:year			0.11
			(0.08)
Topic042.fees:year			0.05
			(0.09)
Topic043.securities:year			0.11
			(0.09)
Topic044.communications:year			-0.02
			(0.09)
Topic045.valuation:year			0.01
			(0.08)

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
Topic046.search_warrant:year			-0.03 (0.09)
Topic047.surveillance:year			-0.17 (0.09)
Topic048.juries:year			0.10 (0.09)
Topic049.self.incrimination:year			-0.22* (0.08)
Topic050.jurisdiction:year			0.08 (0.08)
Topic051.negligence:year			0.12 (0.09)
Topic052.sentencing:year			0.08 (0.09)
Topic053.ad_law_boards:year			-0.16 (0.08)
Topic054.deadlines:year			0.10 (0.08)
Topic055:year			-0.33* (0.12)
Topic056.employ_discrim:year			-0.05 (0.10)
Topic057.ad_law_regs:year			0.00 (0.08)
Topic058.police_test:year			0.09 (0.09)
Topic059.appeals:year			0.09 (0.11)
Topic060.mails:year			-0.09 (0.08)
Topic061.patent.invention:year			0.19* (0.08)
Topic062.corporations:year			0.07 (0.08)
Topic063.procedure_filings:year			0.27* (0.09)
Topic064.press:year			-0.06

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
			(0.08)
Topic065.tax_2:year			0.11
			(0.08)
Topic066.gas.power.oil:year			-0.23*
			(0.08)
Topic067.habeas:year			0.02
			(0.08)
Topic068.employ:year			-0.06
			(0.08)
Topic069.ERISA:year			-0.04
			(0.09)
Topic070.conspiracy:year			-0.13
			(0.09)
Topic071.ineffective_assist:year			0.04
			(0.09)
Topic072.rail:year			-0.13
			(0.08)
Topic073.maritime:year			-0.30*
			(0.08)
Topic074.sales:year			0.10
			(0.08)
Topic075.8th_Cir.:year			0.18*
			(0.09)
Topic076.speech:year			0.06
			(0.09)
Topic077.federal.powers_2:year			0.03
			(0.07)
Topic078.5th_Cir.:year			0.15
			(0.08)
Topic079.obscene.material:year			-0.06
			(0.09)
Topic080.testimony:year			-0.11
			(0.08)
Topic081.labor_arbitration:year			-0.16
			(0.09)
Topic082.damages:year			-0.00
			(0.08)

Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
Topic083.political_orgs:year			-0.10 (0.07)
Topic084.property:year			0.09 (0.08)
Topic085.immigration:year			-0.05 (0.08)
Topic086.political_voting:year			-0.02 (0.09)
Topic087.1st_2d_Cir.:year			-0.27* (0.08)
Topic088.bankruptcy:year			0.15 (0.09)
Topic089.entitlements:year			0.09 (0.09)
Topic090.testimony_2:year			-0.04 (0.08)
Topic091.contracts:year			-0.01 (0.08)
Topic092.FOIA.disclosure:year			-0.10 (0.08)
Topic093.copyright:year			0.13 (0.09)
Topic094.precedent:year			0.08 (0.08)
Topic095.church.state.schools:year			0.01 (0.09)
Topic096.federal.claims:year			-0.04 (0.08)
Topic097.class.action:year			0.00 (0.08)
Topic098.sovereign.immunity:year			0.05 (0.08)
Topic099.Sup._Ct.:year			0.61* (0.13)
<i>N</i>	28825	28825	28825
AIC	24586.91	20527.88	20389.52
BIC	26439.17	25687.73	28856.97

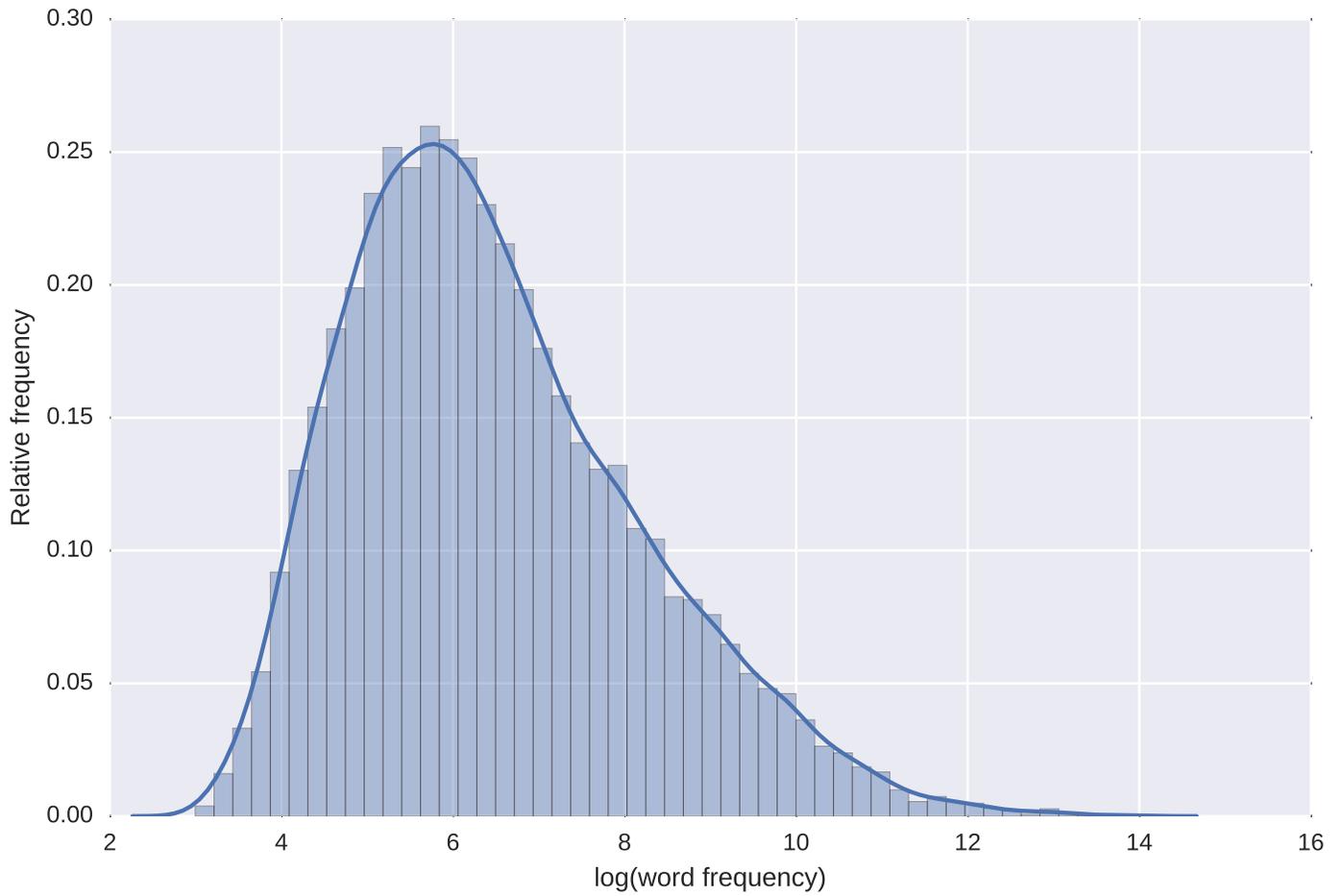
Table 2: Regression results (predicting Federal Appellate decision being referenced in a Supreme Court decision's 'prior history'). Each model includes indicator variables for each year. Model 1 uses just these dependent variables. Model 2 uses topic proportions (expressed as log odds) as well. Model 3 interacts topic proportions with the year. Input variables have been standardized to have mean zero and standard deviation 0.5.

	Model 1	Model 2	Model 3
$\log L$	-12069.46	-9639.94	-9170.76

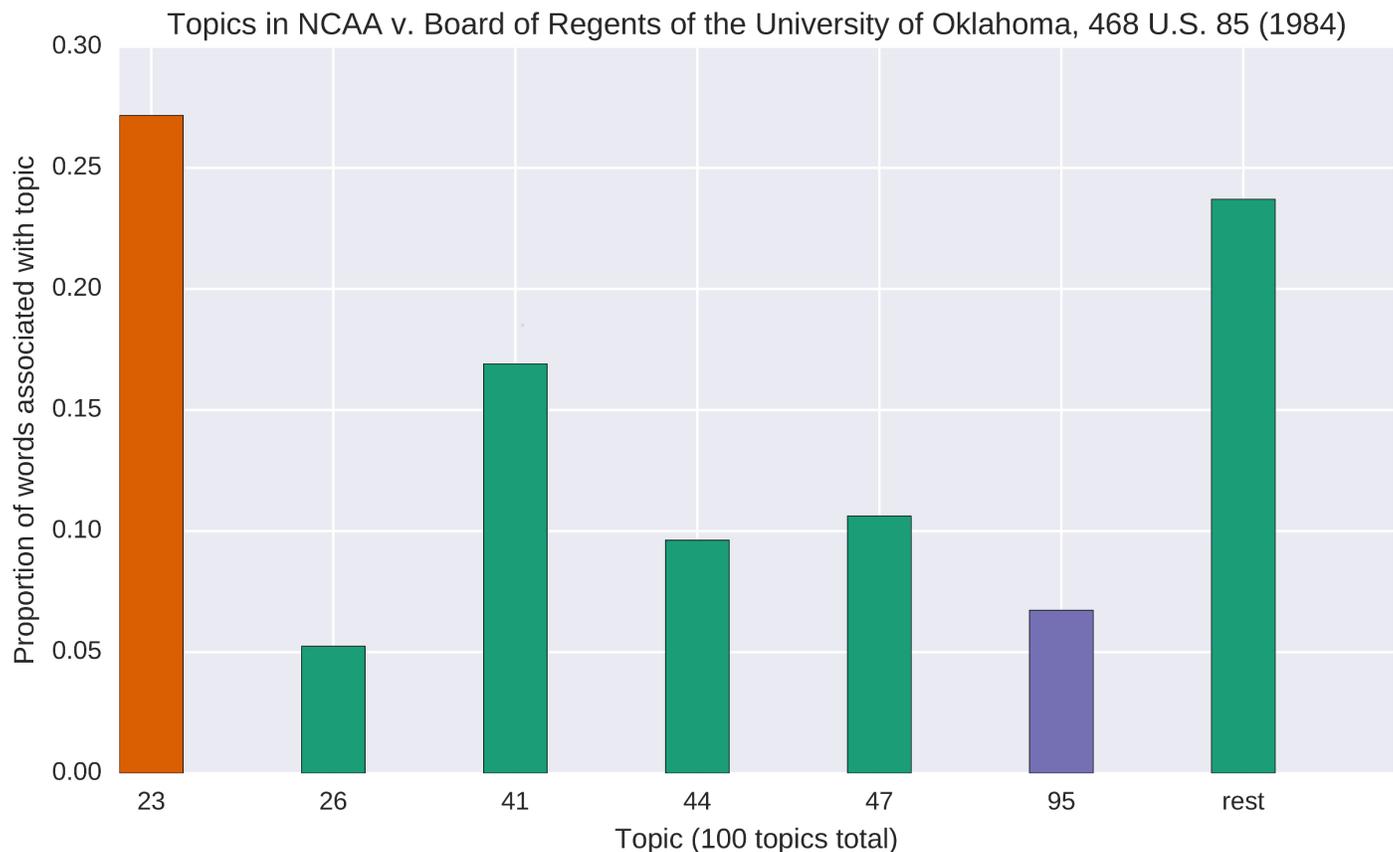
Standard errors in parentheses

\* indicates significance at  $p < 0.05$

## Figures



**Figure 1:** The distribution of word frequencies across all words occurring in the corpus. The median frequency in the approximately ~10 million words in the corpus is ~500. The maximum frequency, for the word “the”, is ~1.1 million. Given the constraint that excludes words occurring in fewer than 20 documents, the minimum word frequency is 20.



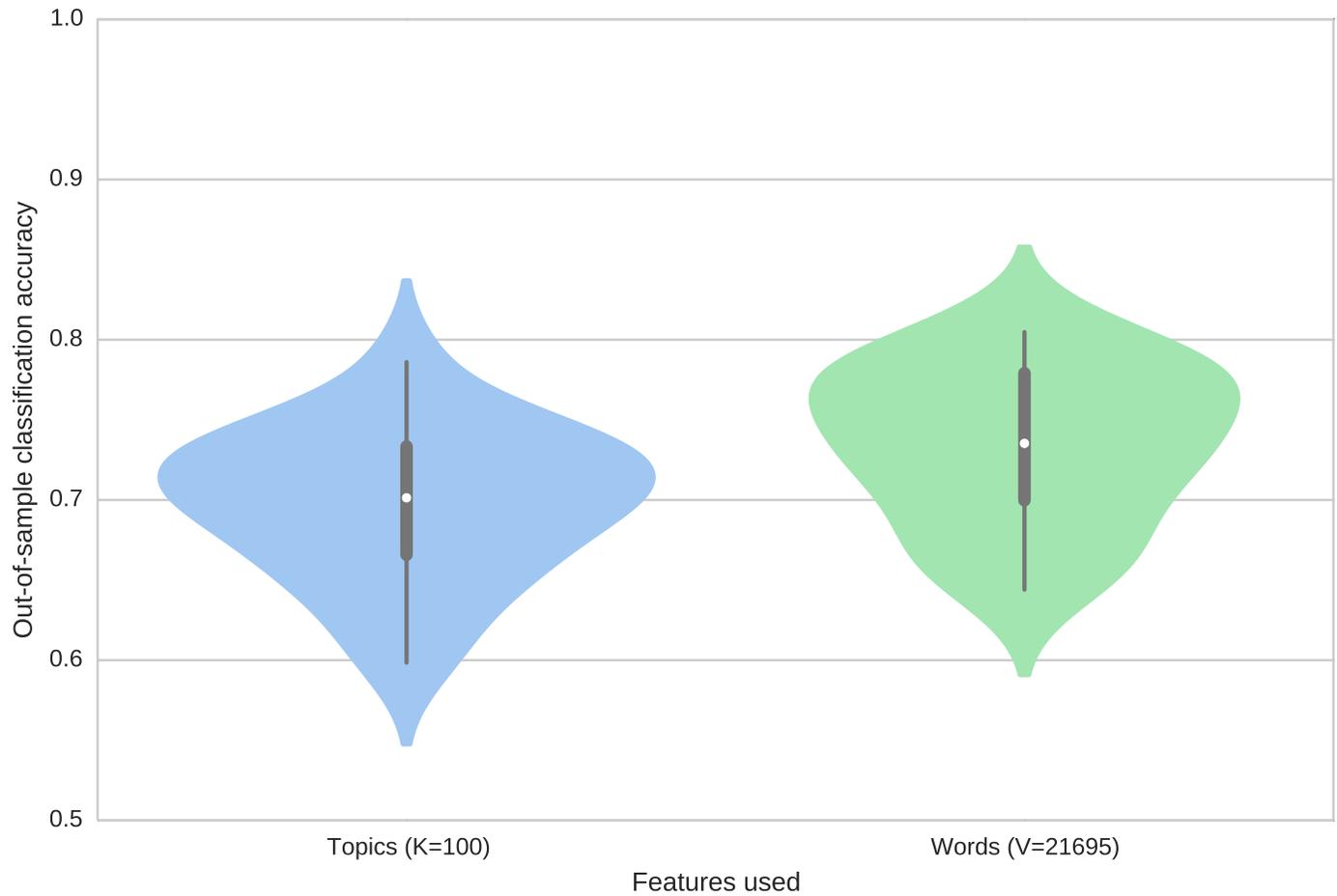
## NCAA v. Board of Regents of the University of Oklahoma, 468 U.S. 85 (1984)

Some years ago, five **major** conferences **together** with **major** football-playing **independent institutions** organized the **College** Football Association (CFA). The original **purpose** of the CFA was to **promote** the interests of **major** football-playing **schools** **within** the **NCAA** **structure**. The **Universities** of Oklahoma and **Georgia**, respondents in this Court, are members of the CFA.

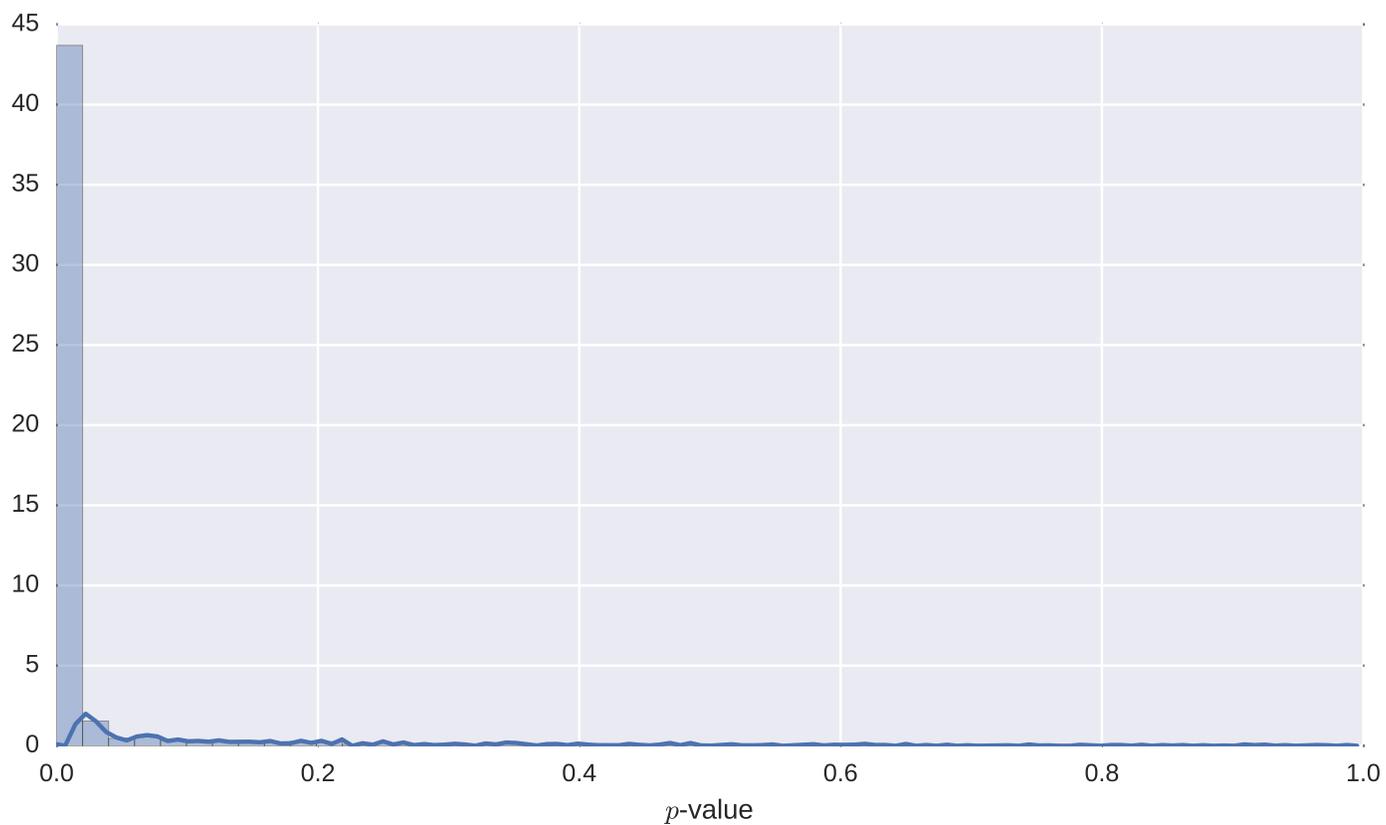
...

As Judge Bork has noted: "[S]ome activities can only be carried out **jointly**. Perhaps the **leading** example is league sports. When a league of professional lacrosse teams is **formed**, it would be pointless to **declare** their **cooperation illegal** on the ground that there are no other professional lacrosse teams." R. Bork, *The Antitrust Paradox* 278 (1978). What the NCAA and its member **institutions** **market** in this case is **competition** itself — contests between **competing institutions**. Of course, this would be completely ineffective if there were no rules on which the **competitors** **agreed** to **create** and **define** the **competition** to be **marketed**. A myriad of rules **affecting** such matters as the **size** of the **field**, the number of players on a team, and the extent to which physical violence is to be **encouraged** or **proscribed**, all must be **agreed** upon, and all **restrain** the manner in which **institutions** **compete**. Moreover, the NCAA seeks to **market** a particular **brand** of football — **college** football. ...

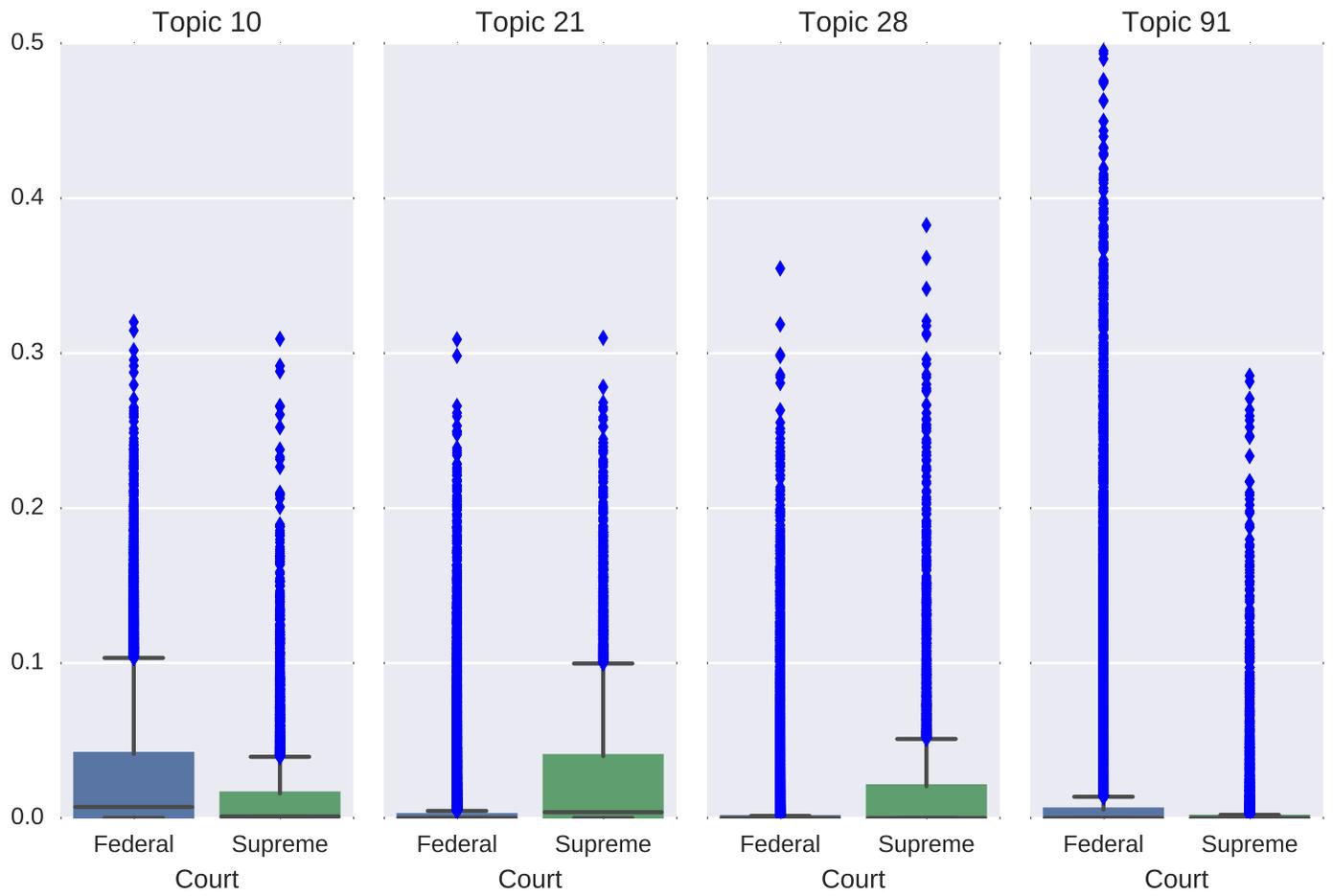
**Figure 2:** A topic model posits the existence of a number of latent topics (distributions over words) from which all words in the corpus are (probabilistically) derived. Each document is associated with a distribution *over topics*; each individual word may be thought of as having, in the most common formulation of the model, a latent topic assignment. This figure shows words with one of two topic assignments where the posterior assignment is clear. The topics and topic assignments in this figure are derived from the topic model fit used in this paper.



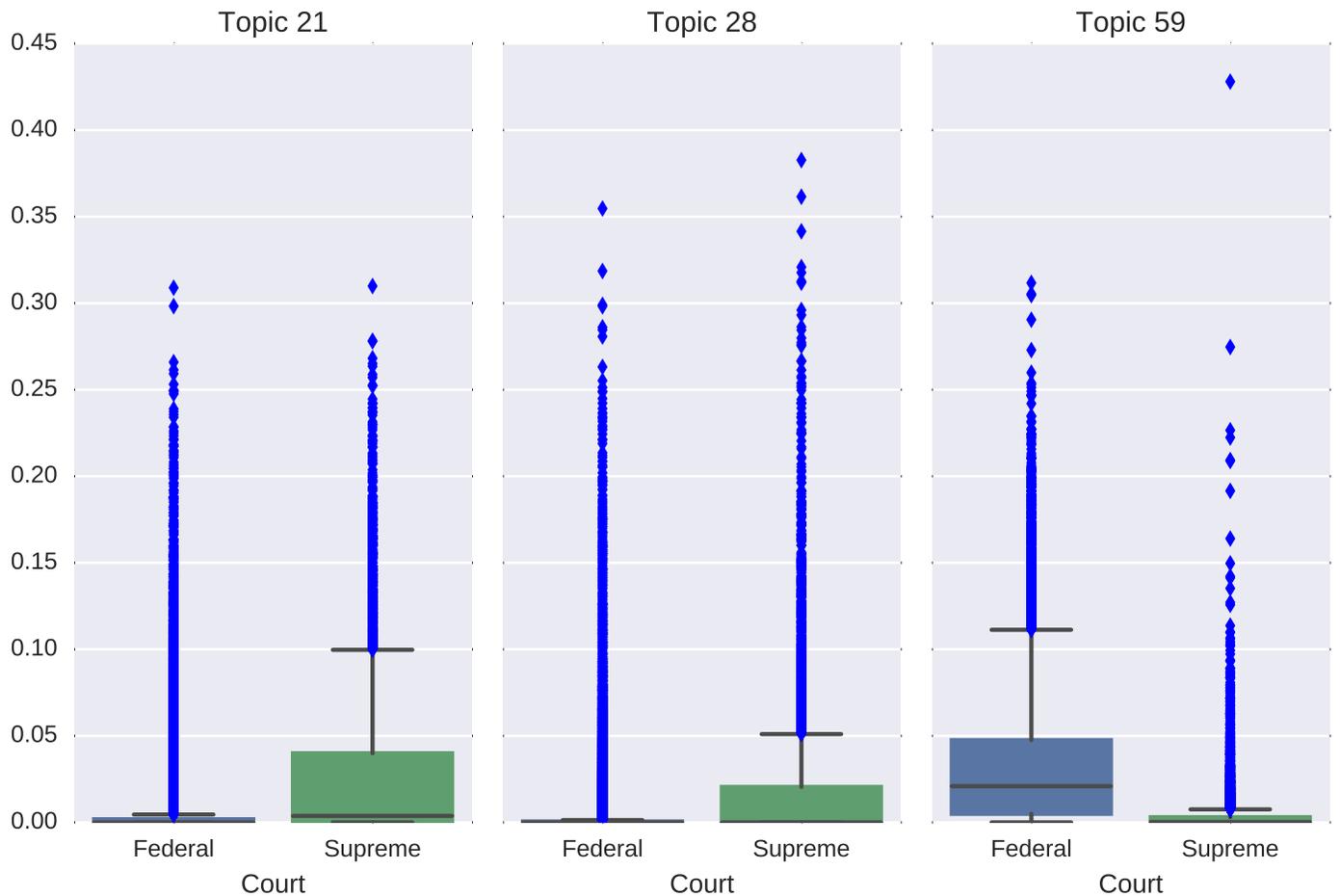
**Figure 3:** Out-of-sample classification accuracy using (1) the topic model (100 topics) and (2) word frequencies (frequencies for all 21,695 words). The SCDB issueArea classification is being predicted. Accuracy assessed using 20-fold cross-validation.



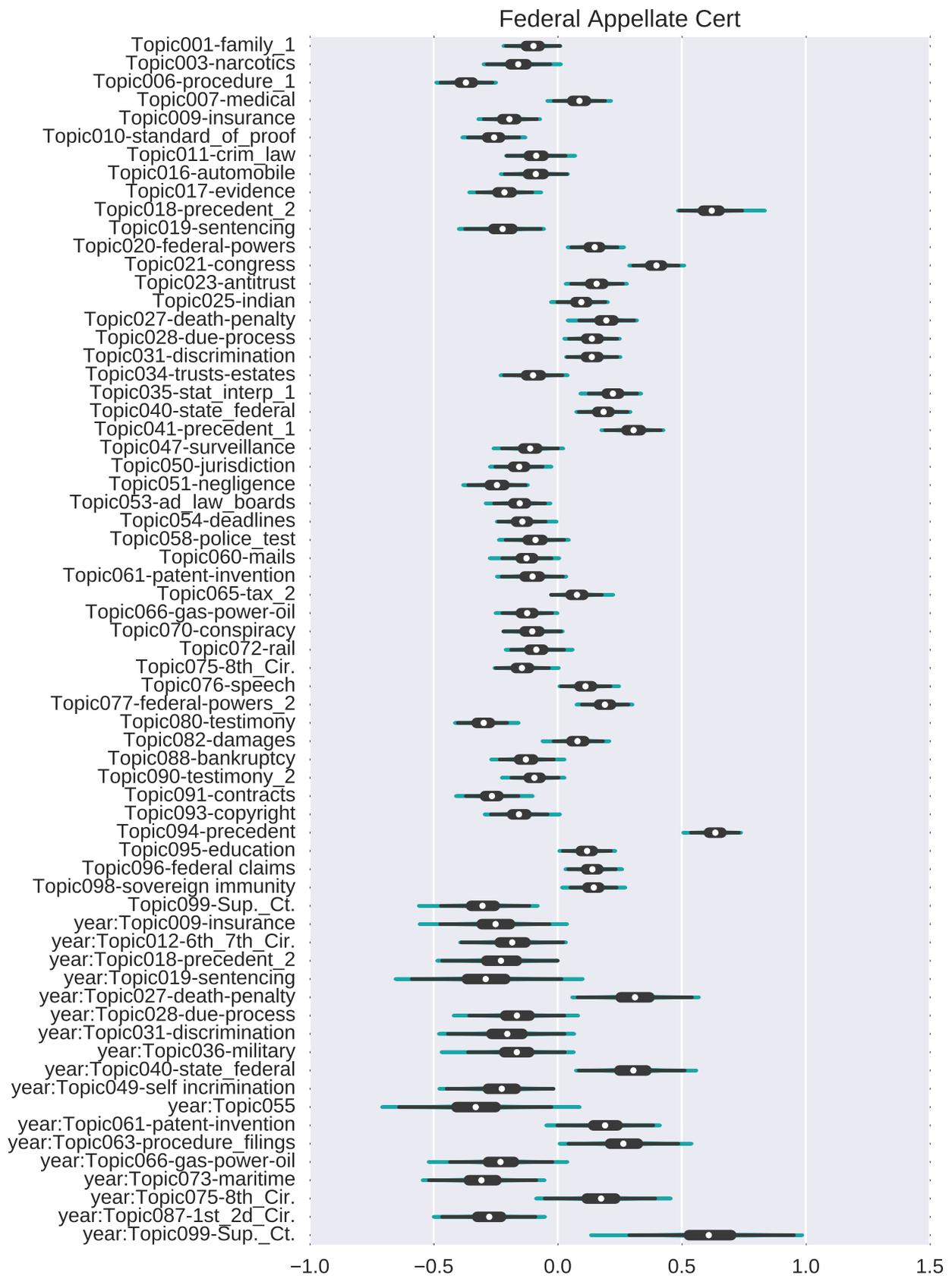
**Figure 4:**  $p$ -values associated with the test of Hypothesis 1 for each Supreme Court decision in the corpus. Unsurprisingly, the topics in Supreme Court cases tend to resemble the topics in the Federal Appellate case under review.



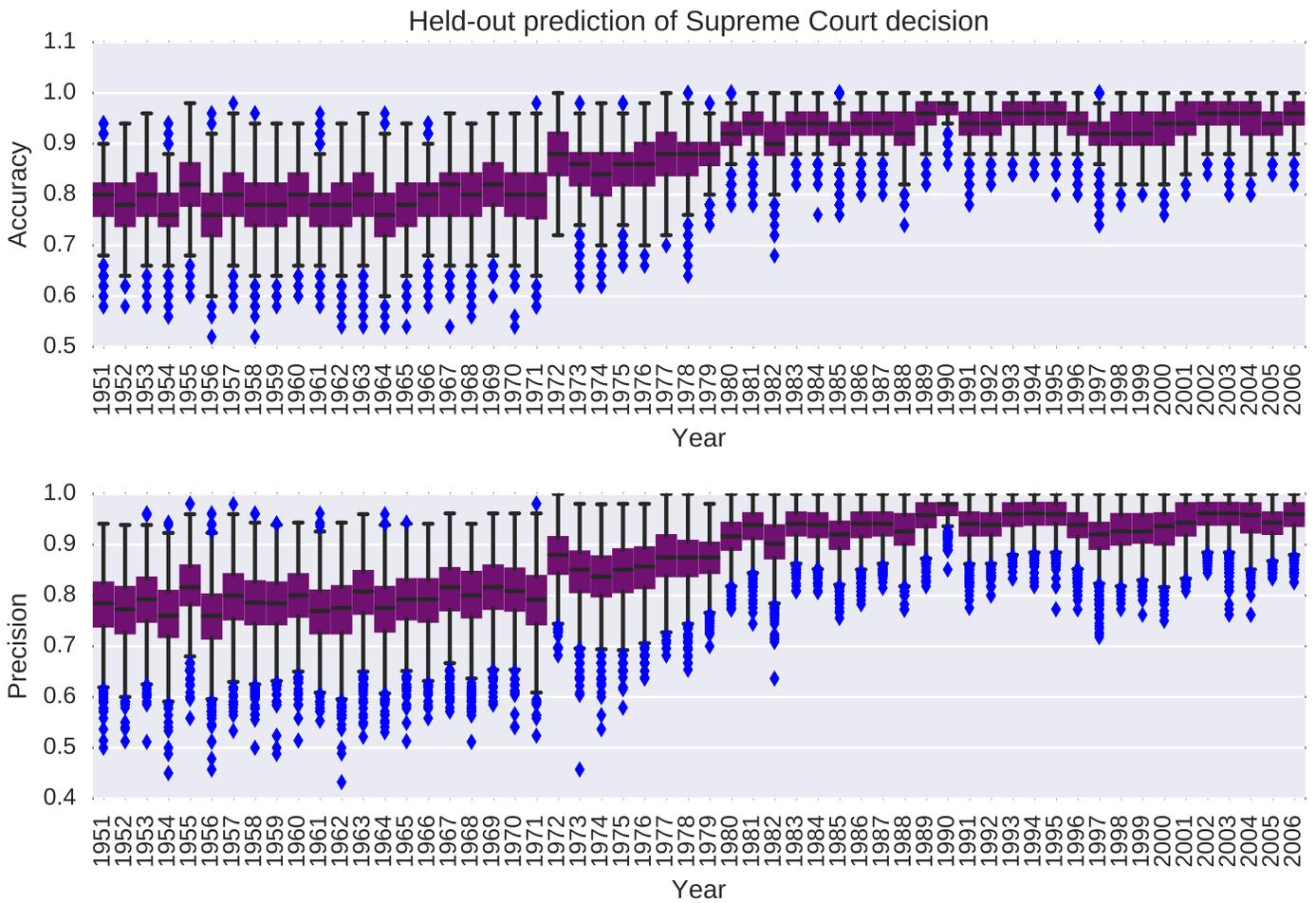
**Figure 5:** Each decision has a topic share, the percentage of words associated the topic. The figure shows a sample of results for those topics where the difference in average topic shares between the Supreme Court and the Federal Appellate Courts is large.



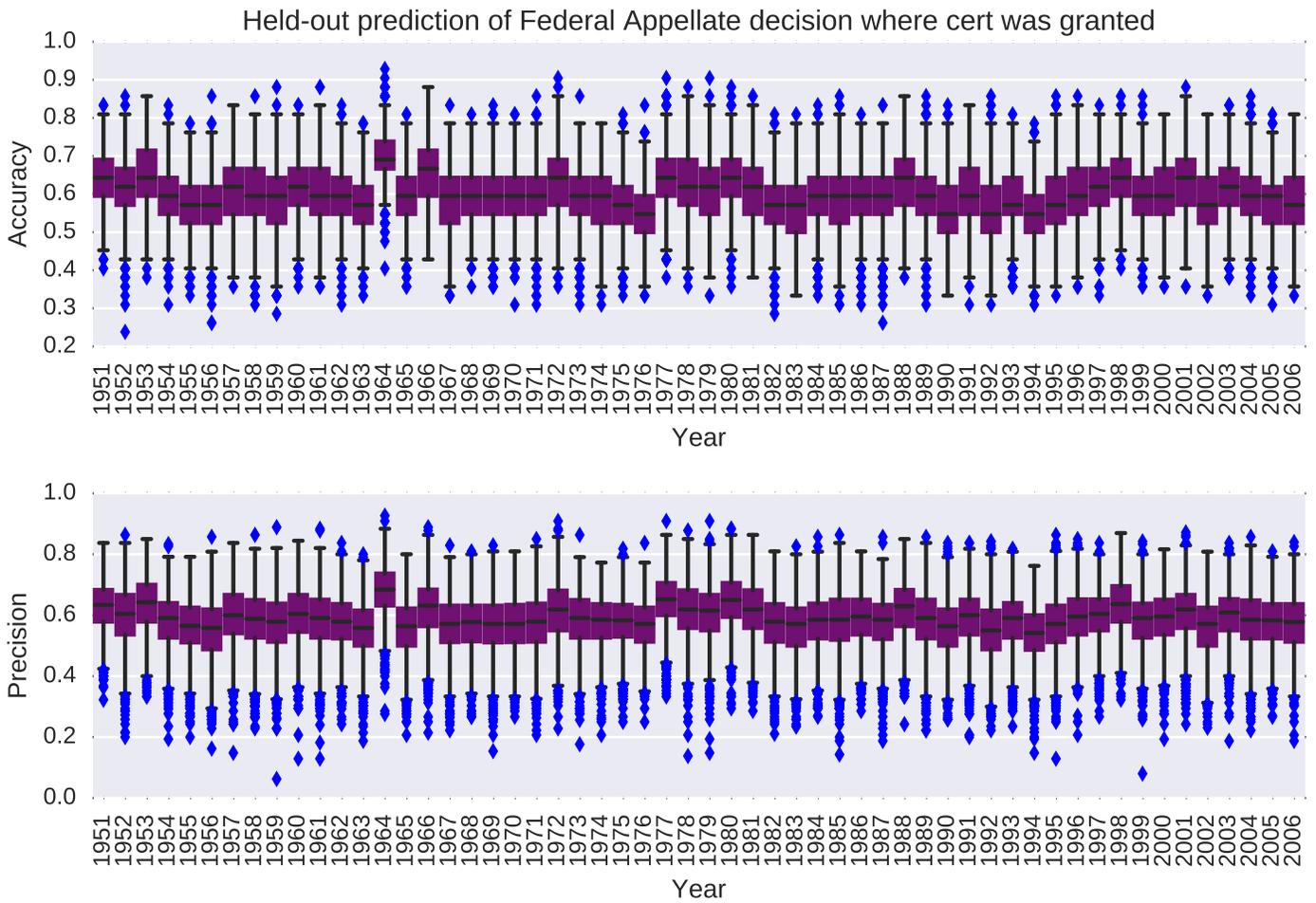
**Figure 6:** Comparing the collection of Federal Appellate decisions that are referenced by Supreme Court decisions with the collection of Supreme Court decisions reveals differences that are less pronounced. The differences in average topic shares are among the largest for the topics above.



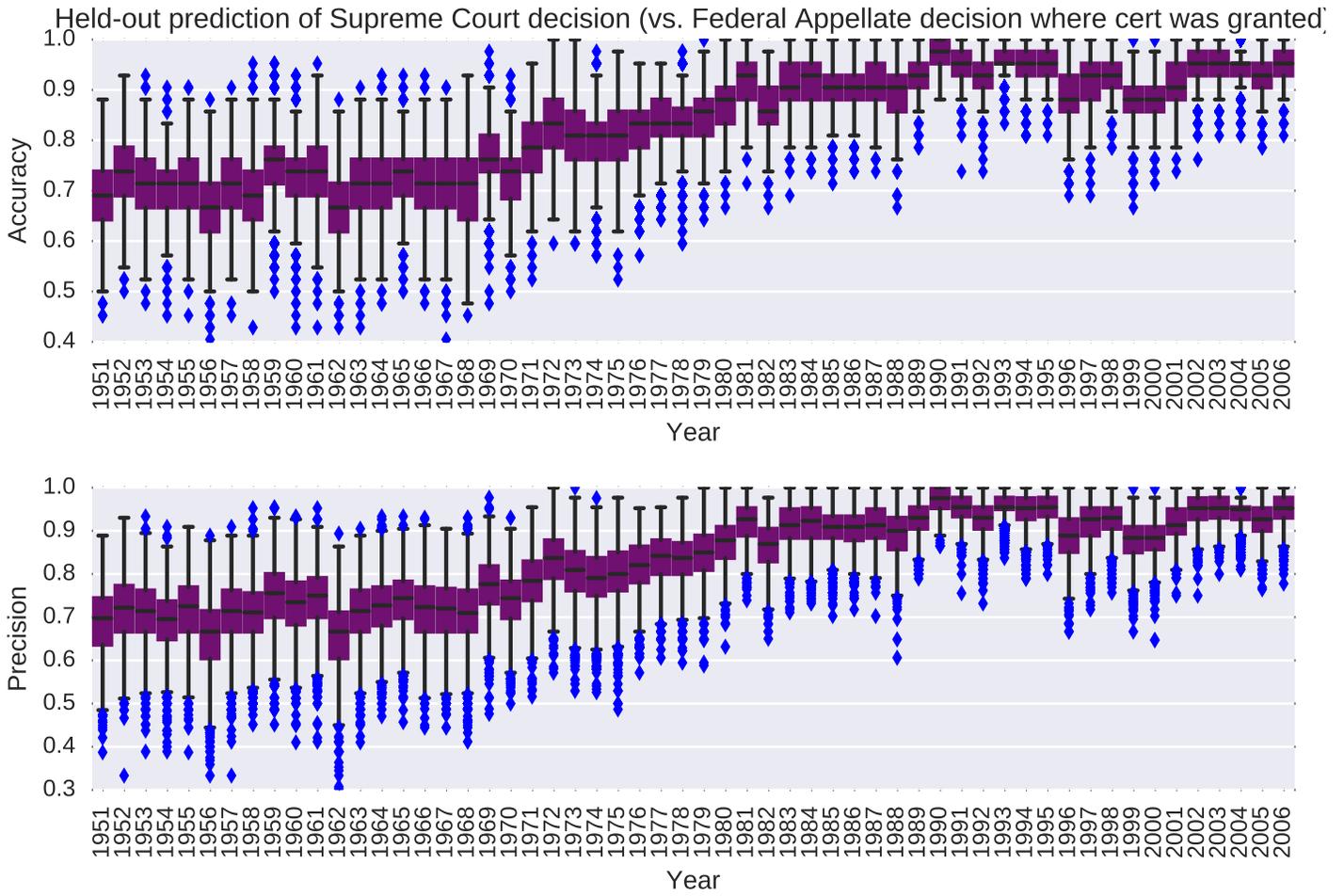
**Figure 7:** Coefficient confidence intervals in regression predicting whether or not a Federal Appellate decision is selected for review. Only coefficients with p-values less than 0.05 are shown. The input data—topic proportions and year—have been centered and rescaled to have mean 0 and standard deviation 0.5.



**Figure 8:** Held-out prediction of Supreme Court vis-à-vis Federal Appellate Court decisions. For each year a random sample of equal size is taken from the decisions from each jurisdiction. Half of the sample is held out and logistic regression model is fit using the remaining half. This process is repeated many times to calculate the range of prediction accuracy and precision shown.



**Figure 9:** Held-out prediction of Federal Appellate decisions granted cert vis-à-vis Federal Appellate decisions not granted cert. For each year a random sample of equal size is taken from the decisions from each jurisdiction. Half of the sample is held out and logistic regression model is fit using the remaining half. This process is repeated many times to calculate the range of prediction accuracy and precision shown.



**Figure 10:** Held-out prediction of Supreme Court decisions vis-à-vis Federal Appellate decisions granted cert. For each year a random sample of equal size is taken from the decisions from each jurisdiction. Half of the sample is held out and logistic regression model is fit using the remaining half. This process is repeated many times to calculate the range of prediction accuracy and precision shown.