# Mapping the Geometry of Law Using Document Embeddings

Elliott Ash and Daniel L. Chen*

September 30, 2018

**Abstract**

Recent work in natural language processing represents language objects (words and documents) as dense vectors that encode the relations between those objects. This chapter explores the application of these methods to legal language, with the goal of understanding judicial reasoning and the relations between judges. In an application to federal appellate courts, we show that these vectors encode information that distinguishes courts, time, and legal topics. The vectors do not reveal spatial distinctions in terms of political party or law school attended, but they do highlight generational differences across judges. We conclude the chapter by outlining a range of promising future applications of these methods.

## 1 Introduction

Law is embedded in language. In this chapter, we ask what can be gained by applying to the law new techniques from natural language processing that translate words and documents into vectors within a space. Vector representations of words and documents are information-dense—in the sense of retaining information about semantic content and meaning—while also being computationally tractable. This combination of information density and computational tractability opens up a wide potential realm of mathematical tools that can be used to generate quantitative and empirically testable insights into the law.

*Elliott Ash, Assistant Professor of Law, Economics, and Data Science ETH Zurich, ashe@ethz.ch. Daniel L. Chen, Professor of Economics, University of Toulouse, daniel.chen@iast.fr. We thank Brenton Arnaboldi, David Cai, Matthew Willian, and Lihan Yao for helpful research assistance. We thank Michael Livermore, Daniel Rockmore, and participants at the Santa Fe Institute Law as Data Workshop for helpful feedback on this research.

This new approach to legal studies addresses the shortcomings of existing methods for studying legal language. For starters, the formal theory work in law and economics requires strong simplifying assumptions that treat the law metaphorically. The case-space literature, for example, assume the language of law to be a function over a geometric space, where the law separates the fact space into "liable" and "not liable" or "guilty" and "not guilty."[1] The case space models give us some intuition into the legal reasoning process. But they have been somewhat limited empirically because it has been infeasible to measure the legal case space.

Likewise, because law consists of text, the standard empirical research methods are somewhat limited in the questions that can be asked. The traditional empirical legal studies literature has relied on small-scale data sets, where legal variables are manually coded (e.g. Songer and Haire, 1992). Hand-coding legal documents is labor-intensive and requires subjective and simplifying decisions.

Meanwhile, recent work in computational linguistics has made breakthroughs in vector representations of language (Jurafsky and Martin, 2014). Topic models such as latent dirichlet allocation serve to automate the coding of texts by generating topics as sets of words that tend to co-occur (Blei et al., 2003; Blei, 2012). These algorithms have provided a window on the relations between documents at scale.

An active literature in computational legal studies has begun to apply these methods to legal documents. Livermore et al. (2016) use a topic model to understand agenda formation on the U.S. Supreme Court (see also Carlson et al., 2015). Leibon et al. (2018) use a network model to represent the geometric relations between U.S. Supreme Court cases. Ganglmair and Wardlaw (2017) apply a topic model to debt contracts, while Ash et al. (2018b) apply one to labor union contracts.

But legal scholars are still mostly unfamiliar with a new counterpart to topic models: embeddings models. The success of word embedding models, such as Google's Word2Vec algorithm, is that they "learn" the conceptual relations between words; a trained model can produce synonyms, antonyms, and analogies for any given word (Mikolov et al., 2013; Levy et al., 2015). These word vectors serve well as features in down-stream prediction tasks by encoding a good deal of information in relatively rare word features. More recently, "document embeddings" have built upon the success of word embeddings to represent words and documents in a joint geometric space (Le and Mikolov, 2014). Like word embeddings, these document embeddings have advantages in terms of interpretability and serve well in

[1] Cameron and Kornhauser (2017) provide a recent review of this literature.

prediction and classification tasks.

This chapter serves to introduce the method of document embeddings into the law and to illustrate the method using a corpus of U.S. appellate court cases. Out data include the universe of U.S. Supreme Court and U.S. Circuit Court cases for the years 1887 through 2013. We construct document embeddings for each opinion in the corpus. We then construct judge vectors by taking the average of the document embeddings for the cases authored by the judge. These case vectors are used to analyze the geometry of federal appellate caselaw.

We ask whether the information recovered by our model provides a meaningful signal about the legal content in cases. We find that spatial clustering in these embeddings encode differences between cases on different courts, between cases in different years, and between cases in different legal topics. The vectors can also discriminate judges based on birth cohorts, but does not do well in encoding the partisan affiliation of judges or law school attended. We also demonstrate that the vectors can show which judges are similar to each other in their legal writing.

We provide a concluding discussion in Section 4. In addition, we outline a range of potential future applications for the use of embeddings models in computational analysis of law. These include looking at analogies and associations, experimenting with structured and categorical embeddings, and constructing embeddings for citation networks.

# 2 Embeddings Models and the Law

A first-order problem in empirical analysis of text data is the high dimensionality of text. There are an arbitrary number of approaches for representing plain text as data. One must trade off informativeness, interpretability, and computational tractability (Ash, 2017). For example, one could represent a document as a frequency distribution over words. But with a large vocabulary, say 20,000 words, a document is still a high-dimensional vector.

Word embeddings came about as a dimension reduction approach in deep learning models for prediction tasks in computational linguistics (Mikolov et al., 2013). Such a prediction task would include, for example, predicting the next word in a sequence given a set of words in a sentence. To that end, the model represents a word as a small and dense vector (say 100 dimensions). Initially, words are randomly distributed across the vector space. But the word locations then become features in a learning model: they move around during training to improve performance on a prediction task. In natural language settings, this process typically leads to words clustering near similar words.

Document embeddings, such as Le and Mikolov's (2014) paragraph vectors, use a separate embedding layer for both the word and the document to solve the prediction task. These models locate documents in a vector space, where documents that contain similar language tend to be located near to each other in the space. Embedding models are different from topic models (e.g. Blei, 2012) because the dimensions have a spatial interpretation, rather than a topic-share interpretation.

Embeddings models have become popular because the spatial relations between the vectors encode useful and meaningful information (Levy et al., 2015). To illustrate, a word embedding can identify similar words in the vocabulary. For example, "judge" might be close to "jury" but far away from "flowerpot." Similarly, a document embedding can identify similar cases in a corpus of decisions based on use of similar language. For example, *Engel v. Vitale* (1962) might be spatially close to *Everson v. Board of Education* (1947), since they are both early U.S. Supreme Court decisions that deal with religious freedoms in the states (this turns out to be true in our data). Finally, a judge embedding constructed from these documents could be used to identify similar judges in the legal system. For example, Antonin Scalia might be close to Clarence Thomas, since they are both conservative judges who tend to use originalist arguments (also true in our data).

# 3 Application to Federal Appellate Courts

This section illustrates the use of document embeddings in the federal appellate courts. We begin by discussing the data and how the document vectors are constructed. We then explore the visual relations between the cases. Finally, we explore similarity relations between judges.

## 3.1 Data and Documents

The analysis utilizes a corpus of all U.S. Supreme Court cases, and all U.S. Circuit Court cases, for the years 1887 through 2013. We have detailed metadata for each opinion; we mainly use the court, date, case topic, and authoring judge. The circuit court data does not include unpublished opinions. Per curiam opinions and discretionary opinions (concurrences/dissents) are excluded from the analysis.

For case topic, we use the 7-category "General Issue" designation coded for Donald Songer's Court of Appeals Database. To make these categories, Songer's research team

classified cases to a single major topic from crime, civil rights, first amendment, due process, privacy, labor relations, economics/regulation, and other.

The cases are linked to biographical information on the judges obtained from the Federal Judicial Center. This includes a plethora of demographic and career details by judge. In the illustrative analysis, we use birth date and political affiliation of appointing president.

Finally, the data set includes the full text of the authored judicial opinions. We remove HTML markup and citations. We then have each case as a list of tokens. These tokens provide the inputs for the embeddings model.

## 3.2 Construction of Document Vectors

The next step is to construct document vectors for each case $i$. The model we use is Doc2Vec (Le and Mikolov, 2014), implemented in the Python package gensim. The objective function solved by this model is to iterate over the corpus and try to predict a given word using its context (a window of neighboring words), as well as a bag-of-words representation of the whole document. The model uses an embedding layer for the context features and the document features. Therefore the geometric location of documents encodes predictive information for the context-specific frequencies of words in the document.

We feed the case documents in random order into Doc2Vec, using standard parameter choices (Dai et al., 2015). We used the distributed bag-of-words model over the distributed memory model, with 200 dimensions per document vector. Other parameter choices include a context window of size 10, capping the vocabulary at 100,000 words (based on document frequency), and excluding documents shorter than 40 words in length. The model iterates through the corpus in random order for five epochs. As this chapter is an exploration and illustration, we did not substantially explore the parameter space on these margins.

## 3.3 Vector Centering and Aggregation

We now have a set of vectors $\vec{i}$ for each case $i$. Following the advice of the embeddings literature,[2] we normalized each vector to length one. Each case has an authoring judge $j$, working in court $c$ at year $t$. Besides author and time, the other metadata feature is the case topic $k$. We would like to use these categories for descriptive statistics, as well as to "control" for these features for more targeted analysis.

[2]See Omer Levy, "Should I do normalization to word embeddings?". *Quora*, 7 November 2015.

For visualization and other analysis we would like to center and aggregate the document vectors in several ways. Let $I_j$ be the set of cases authored by $j$. Let $I_{jt}$ be the set of cases authored by $j$ at year $t$. One could construct a vector representation for a judge using

$$\vec{j} = \frac{1}{|I_j|} \sum_{i \in I_j} \vec{i}$$

where $|\cdot|$ gives the count of the set. Similarly, the vector for judge $j$ at year $t$ would be given by

$$\vec{jt} = \frac{1}{|I_{jt}|} \sum_{i \in I_{jt}} \vec{i}$$

and the vector for all cases on topic $k$ in court $c$ during year $t$ would be given by

$$\vec{ckt} = \frac{1}{|I_{ckt}|} \sum_{i \in I_{ckt}} \vec{i}.$$

Meanwhile, the same notation and corresponding aggregation formula could be used to construct a vector for a year, $\vec{t}$, for a court $\vec{c}$, for a topic $\vec{k}$, or for the cases in court $c$ during a particular year $t$, $\vec{ct}$.

We are interested in recovering the ideological component of the judge vectors. Therefore we explore the following steps to center the document vectors before aggregating. Represent the year-centered vector for case $i$ as $\vec{i}_t = \vec{i} - \vec{t}_i$, where $\vec{t}_i$ corresponds to the average vector for all cases in the same year as $i$. Similarly, let a subscripted judge vector $\vec{j}_t$ be defined as

$$\vec{j}_t = \frac{1}{|I_j|} \sum_{i \in I_j} \vec{i}_t$$

the average for judge $j$ of the year-centered vectors $\vec{i}_t$.

The preferred centering specification depends on the context of the analysis. We center by interacted groups, in particular. In the results below, we variously center by topic-year $\vec{kt}$, by court-year $\vec{kt}$, and by court-topic-year $\vec{ckt}$. Only after this centering step do we aggregate by judge and perform analysis of the spatial relations between vectors. The hope is that the remaining spatial variation is purged of court-specific, topic-specific, and year-specific differences in language. The remaining variation will provide a cleaner summary of the ideological differences between judges.

Here we have used the unweighted average of the case vectors, where each case is weighted equally. Future work might explore the use of other weighting schemes. A sensible

alternative would be to weight the cases by their length (in words or sentences), for example. In addition, it would be reasonable to weight the cases by the number of citations they later received – as a proxy for importance. Finally, one might normalize the vectors after centering and/or aggregating.

## 3.4 Visual Structure of Case Vectors and Judge Vectors

In this section we present a variety of visualizations to understand better the spatial relationships encoded by our case vectors and judge vectors. Our visualization methods is a t-SNE (t-Distributed Stochastic Neighbor Embedding) plot (Maaten and Hinton, 2008), which projects the vectors down to two dimensions for visualization purposes. We use t-SNE plots because the dimension reduction algorithm is designed to preserve local distances between points and therefore recover informative clusters (Lee and Verleysen, 2007). In tests, we got better visualizations from t-SNE than other manifold learning algorithms, such as principal components, multi-dimensional scaling, or isomap.

We begin by exploring the institutional, temporal, and judge-level features encoded in the vectors. For Figure 1, we centered the case vectors by topic interacted with year, as described in Section 3.3. We then averaged by judge and plotted the judge vectors. The vectors are labeled by court. One can see that, conditional on topic and year, the document vectors separate the courts quite well. This is consistent with systematic differences in legal language across courts, conditional on topic and year, being captured by the embedding.

For Figure 2, we centered on court interacted with topic. We then average by court-year and plotted the court-year-level averaged vectors. The dots are labeled and colored by the decade the case was published. One can see a steady linear development of case law across the geometric space. This shows that, controlling for court and topic factors, the embedding captures systematic differences in language across time.

For Figure 3, the cases were centered on judge interacted with year; this residualizes out any judge-level time-varying components of language. We then averaged and plotted by topic-year. The labels and colors distinguish the seven-digit general issue topic. We can see that the document embeddings discriminate topics, effectively capturing differences in language across recognized issue areas.

Next we look at whether the vectorized language in the case vectors encodes information about judge characteristics. For Figure 4, we centered on an interacted groupings for court, topic, and year. This centering controls for any time-varying topic and court level language variation. We then averaged by judge and plotted the judge vectors. The labels and colors

Figure 1: Centered by Topic-Year, Averaged by Judge, Labeled by Court

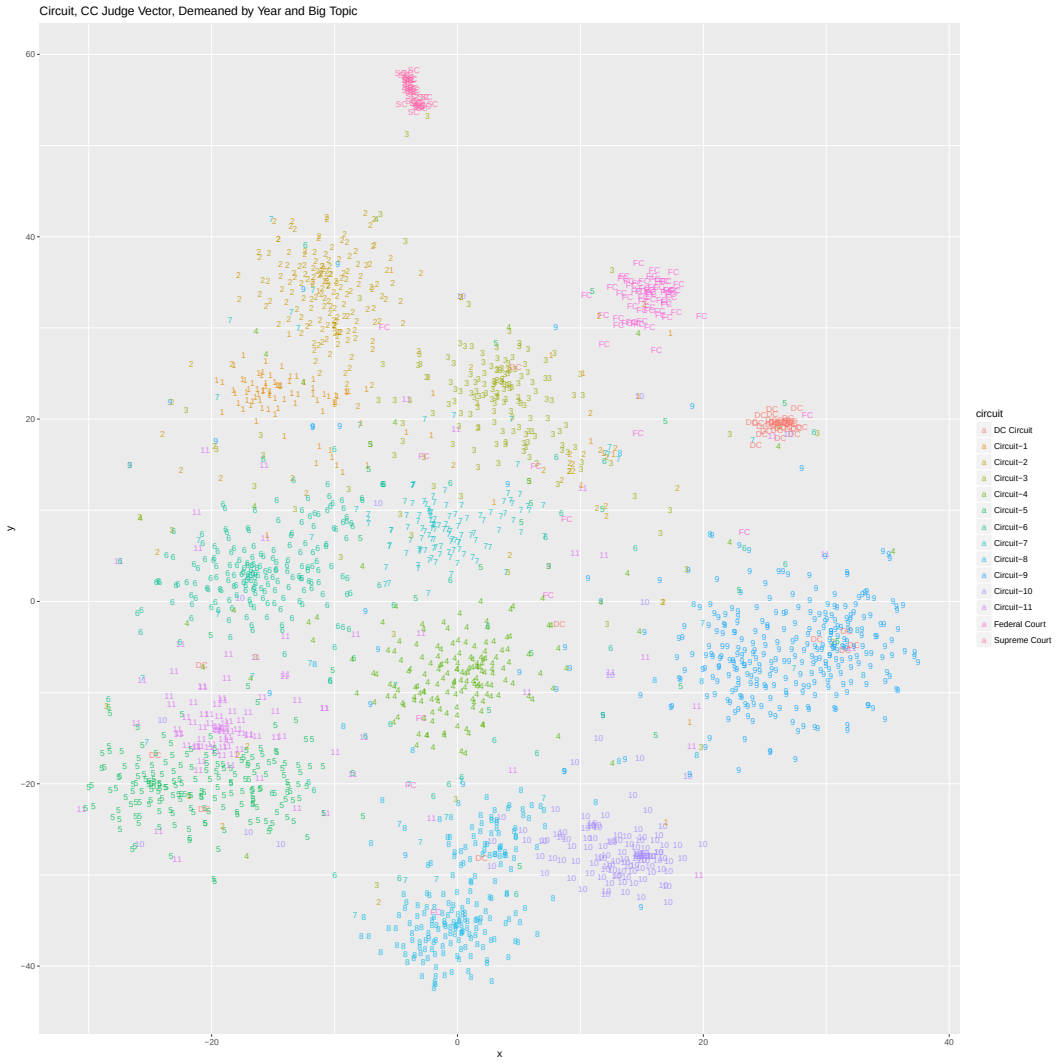


Figure 2: Centered by Court-Topic, Averaged by Court-Year, Labeled by Decade

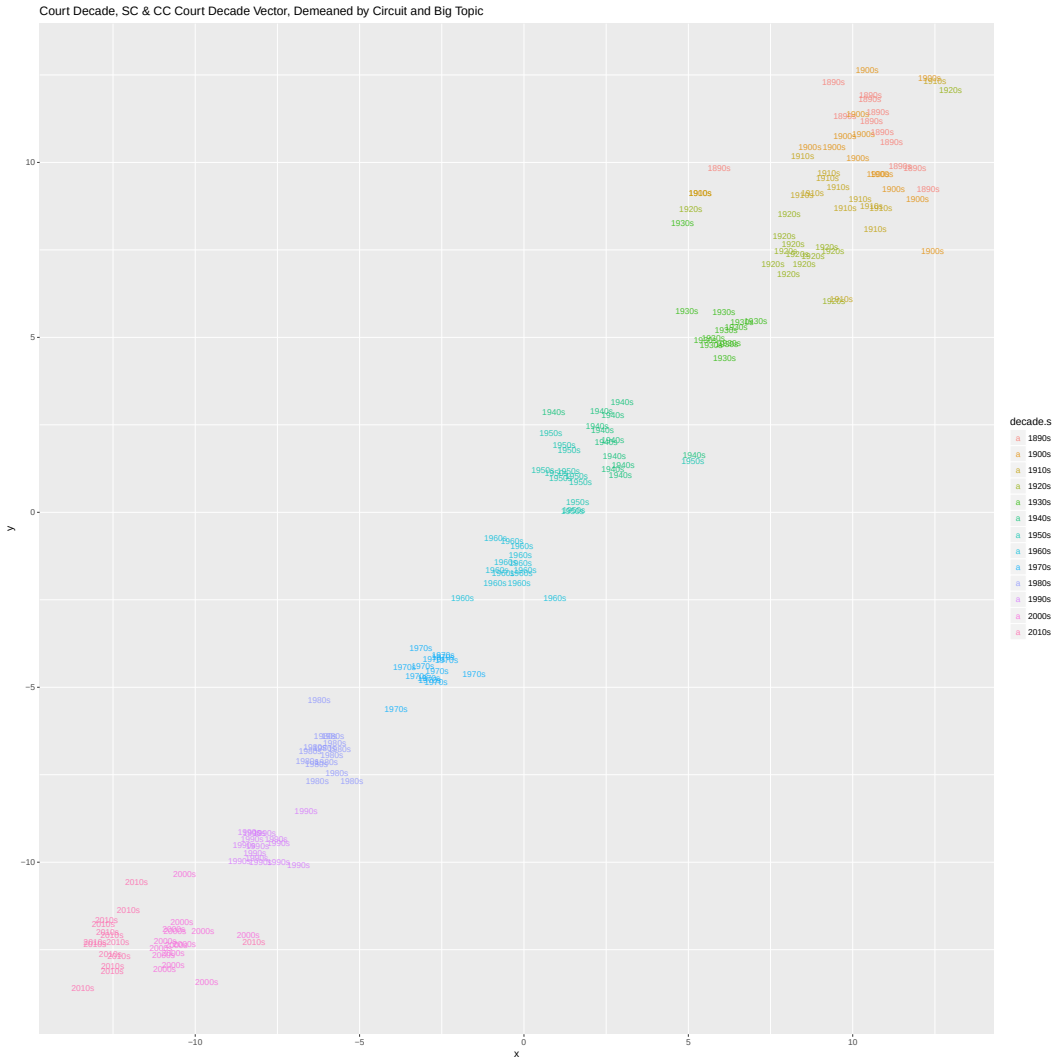
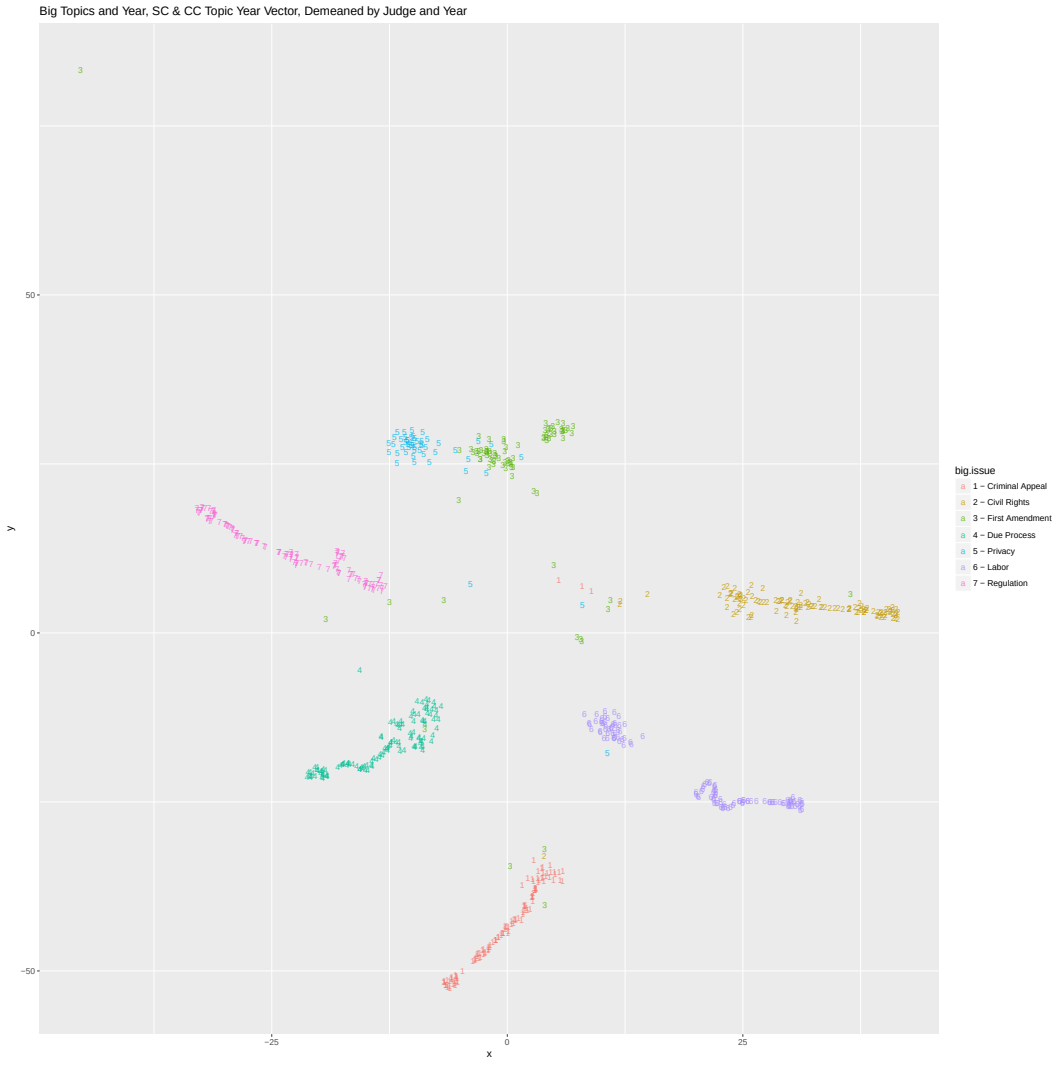


Figure 3: Centered by Judge-Year, Averaged by Topic-Year, Labeled by Topic



are by political party – Democrat or Republican. These are randomly distributed across the graph. It appears that the language features encoded by the document embeddings are not informative about political party.

Figure 5 considers another judicial biographical feature: birth cohort. As before, we centered on court-topic-year and averaged/plotted by judge. In this case, the labels and colors are by birth cohort decade (1910s through 1950s). In stark contrast to political party, there is clear segmentation across the geometric space across cohorts. Remember that this is conditioned on court-topic-year, so is not driven by time trends over the sample. The vectorized language recovers differences in the legal language used by judges from different generations.

Finally, for Figure 6, we consider law school attended as a final source of linguistic differences across judges. Conditional on court, topic, and year, we see apparent random distributions across the space in terms of law school. As with political party, it seems like language or ideological differences by school do not show up in the vectors.

## 3.5 Analysis of Relations Between Judges

This section uses our vector representation of judges to produce a similarity metric between courts and judges. We adopt a measure of vector similarity that is used often for document classification. The *cosine similarity* between two vectors,

$$s(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\| \, \|\vec{w}\|},$$

which is equal to one minus the cosine of the angle between the vectors. It takes a value between -1 and 1. In the case of word embeddings, high similarity means that the words are often used in similar language contexts.

In the case of judges, we can say that similarities approaching one mean that the judges tend to use similar language in their opinions. Similarities approaching -1 meaning the judges rarely use the same language. Similarities near zero mean that the judges are as similar to each other as would be expected from two randomly selected judges in the population.

First we look at similarity between court vectors to complement the spatial representation in Figure 1. We centered the vectors by topic and year, and then aggregated by court. We then computed the pair-wise similarities between the court vectors. These are reported in Table 1.

The colors provide a gradient for similarity, with green meaning the courts are relatively

Figure 4: Centered by Court-Topic-Year, Averaged by Judge, Labeled by Political Party

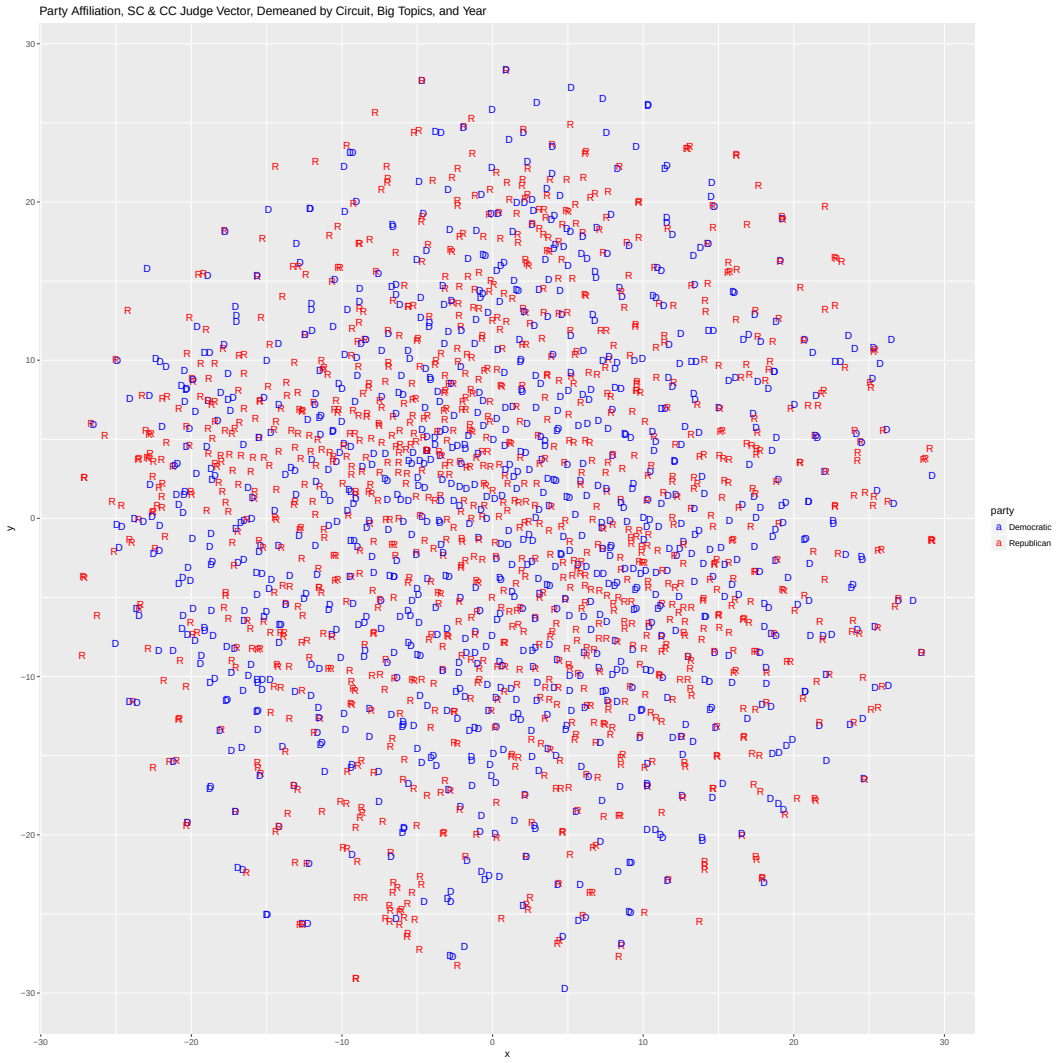


Figure 5: Centered by Court-Topic-Year, Averaged by Judge, Labeled by Judge Birth Cohort

Figure 6: Centered by Court-Topic-Year, Averaged by Judge, Labeled by Law School Attended

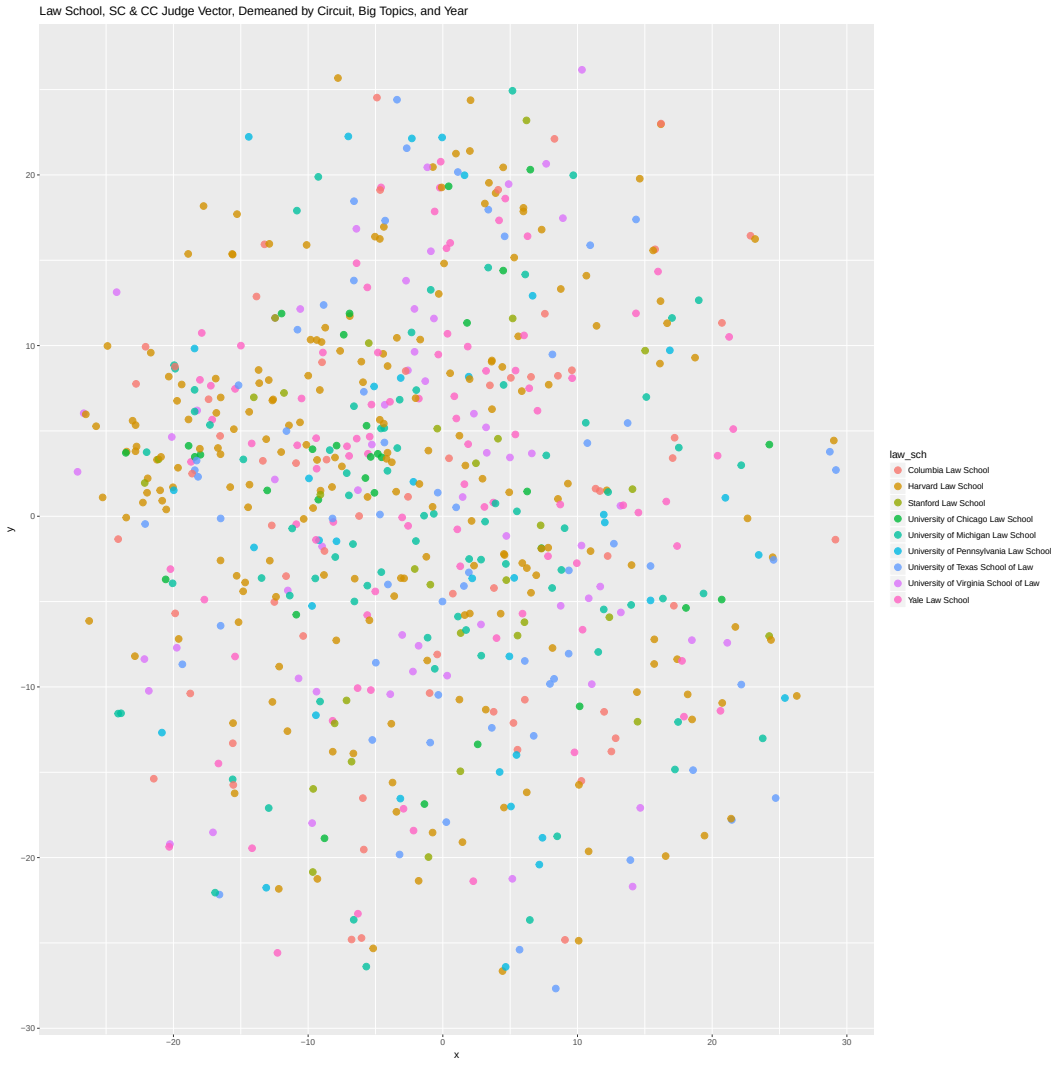


Table 1: Pair-Wise Similarities Between Federal Appellate Courts

| | SCOTUS | 1st Circ. | 2nd. Circ. | 3rd. Circ. | 4th Circ. | 5th Circ. | 6th Circ. | 7th Circ. | 8th Circ. | 9th Circ. | 10th Circ. | 11.th Circ. | D.C. Circ. | Fed. Circ. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SCOTUS | 1.000 | | | | | | | | | | | | | |
| 1st Circ. | 0.022 | 1.000 | | | | | | | | | | | | |
| 2nd. Circ. | -0.008 | 0.302 | 1.000 | | | | | | | | | | | |
| 3rd. Circ. | -0.001 | 0.135 | 0.207 | 1.000 | | | | | | | | | | |
| 4th Circ. | -0.045 | -0.045 | -0.081 | 0.126 | 1.000 | | | | | | | | | |
| 5th Circ. | -0.105 | -0.196 | -0.298 | -0.269 | 0.038 | 1.000 | | | | | | | | |
| 6th Circ. | -0.074 | -0.185 | -0.148 | 0.009 | 0.069 | -0.107 | 1.000 | | | | | | | |
| 7th Circ. | -0.097 | -0.052 | -0.014 | -0.055 | -0.162 | -0.257 | 0.029 | 1.000 | | | | | | |
| 8th Circ. | -0.137 | -0.215 | -0.296 | -0.214 | -0.150 | -0.184 | 0.050 | -0.022 | 1.000 | | | | | |
| 9th Circ. | 0.039 | -0.137 | -0.140 | -0.182 | -0.147 | -0.121 | -0.220 | -0.265 | -0.150 | 1.000 | | | | |
| 10th Circ. | -0.111 | -0.249 | -0.361 | -0.179 | -0.189 | 0.017 | 0.006 | -0.158 | 0.218 | 0.042 | 1.000 | | | |
| 11.th Circ. | -0.086 | -0.191 | -0.240 | -0.215 | 0.067 | 0.713 | -0.039 | -0.224 | -0.192 | -0.084 | 0.026 | 1.000 | | |
| D.C. Circ. | 0.846 | -0.085 | -0.058 | 0.011 | -0.010 | -0.062 | -0.097 | -0.177 | -0.111 | 0.067 | -0.025 | 0.011 | 1.000 | |
| Fed. Circ. | 0.178 | 0.200 | 0.132 | 0.116 | 0.124 | -0.150 | -0.154 | -0.082 | -0.255 | -0.116 | -0.260 | -0.181 | 0.094 | 1.000 |

similar and red meaning they are relatively dissimilar. The table has some interesting features. First, the D.C. Circuit is most similar to the Supreme Court of the United States, which is intuitive since they are both located in Washington, D.C. and focus on issues of federal government functioning such as separation of powers. Second, the 11th circuit is similar to the 5th circuit, which is intuitive since the 11th Circuit used to be a part of the 5th Circuit and they share many legal precedents.

Next we look at similarity between judge vectors. This analysis is related to recent work comparing judges based on writing style features (e.g. Carlson et al., 2015). Starting with the Supreme Court, we center the document vectors on topic, and year. Then we take the average of these centered vectors by judge as our representation of judge writing, reasoning, and beliefs.

Table 2 (continued in Table 3) reports the pair-wise similarities between a selection of recently sitting Supreme Court judges. Overall, there are limited immediate insights and the results are mixed. For example, it is intuitive that Scalia is close to Thomas. But counter-intuitively, Scalia is even closer to Souter, Stevens, and O'Connor. Another example: Intuitively, Brennan is close to Thurgood Marshall; but counter-intuitively, he is closer to White and Stewart. Overall, the judge vectors do not seem to encode similarities between Supreme court judges very well. This may be due to the relatively few decisions that they author. In particular, the relative dissimilarity between Kagan and most other justices is likely due to her having only a handful of decisions in the corpus.

One interesting feature of our model is that it represents both circuit court judges and

Table 2: Pair-Wise Similarities between Supreme Court Judges

| | AFortas | AMKennedy | AScalia | BRWhite | CThomas | DHSouter | EKagan | EWarren | FFrankfurter | FMVinson | HABlackmun | HLBlack | JGRoberts | JPStevens |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AFortas | 1.000 | | | | | | | | | | | | | |
| AMKennedy | 0.733 | 1.000 | | | | | | | | | | | | |
| AScalia | 0.735 | 0.974 | 1.000 | | | | | | | | | | | |
| BRWhite | 0.834 | 0.908 | 0.913 | 1.000 | | | | | | | | | | |
| CThomas | 0.686 | 0.962 | 0.958 | 0.854 | 1.000 | | | | | | | | | |
| DHSouter | 0.718 | 0.967 | 0.967 | 0.878 | 0.962 | 1.000 | | | | | | | | |
| EKagan | 0.454 | 0.674 | 0.659 | 0.514 | 0.697 | 0.654 | 1.000 | | | | | | | |
| EWarren | 0.855 | 0.732 | 0.730 | 0.863 | 0.684 | 0.709 | 0.407 | 1.000 | | | | | | |
| FFrankfurter | 0.807 | 0.604 | 0.604 | 0.752 | 0.554 | 0.580 | 0.324 | 0.913 | 1.000 | | | | | |
| FMVinson | 0.717 | 0.542 | 0.542 | 0.675 | 0.494 | 0.521 | 0.310 | 0.838 | 0.906 | 1.000 | | | | |
| HABlackmun | 0.823 | 0.919 | 0.923 | 0.970 | 0.880 | 0.901 | 0.557 | 0.814 | 0.695 | 0.620 | 1.000 | | | |
| HLBlack | 0.873 | 0.706 | 0.706 | 0.847 | 0.655 | 0.689 | 0.381 | 0.943 | 0.930 | 0.854 | 0.803 | 1.000 | | |
| JGRoberts | 0.569 | 0.869 | 0.861 | 0.728 | 0.862 | 0.850 | 0.679 | 0.575 | 0.445 | 0.389 | 0.734 | 0.521 | 1.000 | |
| JPStevens | 0.775 | 0.965 | 0.966 | 0.956 | 0.932 | 0.949 | 0.611 | 0.781 | 0.656 | 0.588 | 0.963 | 0.767 | 0.826 | 1.000 |
| LFPowell | 0.818 | 0.908 | 0.910 | 0.980 | 0.852 | 0.882 | 0.508 | 0.841 | 0.720 | 0.643 | 0.975 | 0.819 | 0.725 | 0.958 |
| PStewart | 0.874 | 0.847 | 0.856 | 0.969 | 0.797 | 0.820 | 0.468 | 0.924 | 0.838 | 0.750 | 0.939 | 0.905 | 0.656 | 0.906 |
| RBGinsburg | 0.699 | 0.950 | 0.952 | 0.853 | 0.961 | 0.953 | 0.702 | 0.679 | 0.546 | 0.477 | 0.882 | 0.660 | 0.849 | 0.933 |
| RHJackson | 0.758 | 0.546 | 0.541 | 0.694 | 0.494 | 0.524 | 0.305 | 0.864 | 0.925 | 0.878 | 0.640 | 0.903 | 0.367 | 0.602 |
| SAAlito | 0.560 | 0.846 | 0.848 | 0.697 | 0.866 | 0.836 | 0.702 | 0.554 | 0.438 | 0.399 | 0.710 | 0.503 | 0.872 | 0.790 |
| SDOConnor | 0.761 | 0.964 | 0.962 | 0.950 | 0.930 | 0.944 | 0.588 | 0.777 | 0.660 | 0.591 | 0.955 | 0.757 | 0.802 | 0.976 |
| SGBreyer | 0.683 | 0.953 | 0.950 | 0.846 | 0.963 | 0.950 | 0.708 | 0.674 | 0.557 | 0.506 | 0.865 | 0.649 | 0.863 | 0.928 |
| SSotomayor | 0.556 | 0.743 | 0.747 | 0.621 | 0.774 | 0.742 | 0.587 | 0.522 | 0.439 | 0.372 | 0.641 | 0.471 | 0.723 | 0.697 |
| TMarshall | 0.827 | 0.900 | 0.898 | 0.962 | 0.857 | 0.876 | 0.543 | 0.832 | 0.717 | 0.626 | 0.968 | 0.830 | 0.725 | 0.948 |
| WEBurger | 0.811 | 0.871 | 0.873 | 0.967 | 0.813 | 0.836 | 0.464 | 0.843 | 0.734 | 0.658 | 0.953 | 0.813 | 0.702 | 0.924 |
| WHRehnquist | 0.788 | 0.932 | 0.940 | 0.974 | 0.885 | 0.905 | 0.537 | 0.816 | 0.705 | 0.636 | 0.966 | 0.790 | 0.762 | 0.963 |
| WJBrennan | 0.871 | 0.896 | 0.894 | 0.976 | 0.844 | 0.869 | 0.533 | 0.909 | 0.806 | 0.723 | 0.957 | 0.892 | 0.725 | 0.943 |
| WODouglas | 0.872 | 0.720 | 0.722 | 0.859 | 0.674 | 0.707 | 0.412 | 0.938 | 0.924 | 0.847 | 0.819 | 0.972 | 0.536 | 0.785 |

Table 3: Pair-Wise Similarities between Supreme Court Judges (cont.)

| | LFPowell | PStewart | RBGinsburg | RHJackson | SAAlito | SDOConnor | SGBreyer | SSotomayor | TMarshall | WEBurger | WHRehnquist | WJBrennan | WODouglas |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LFPowell | 1.000 | | | | | | | | | | | | |
| PStewart | 0.954 | 1.000 | | | | | | | | | | | |
| RBGinsburg | 0.854 | 0.794 | 1.000 | | | | | | | | | | |
| RHJackson | 0.669 | 0.772 | 0.493 | 1.000 | | | | | | | | | |
| SAAlito | 0.688 | 0.642 | 0.860 | 0.369 | 1.000 | | | | | | | | |
| SDOConnor | 0.946 | 0.898 | 0.921 | 0.597 | 0.778 | 1.000 | | | | | | | |
| SGBreyer | 0.841 | 0.786 | 0.959 | 0.492 | 0.885 | 0.926 | 1.000 | | | | | | |
| SSotomayor | 0.621 | 0.598 | 0.745 | 0.365 | 0.741 | 0.708 | 0.744 | 1.000 | | | | | |
| TMarshall | 0.961 | 0.936 | 0.868 | 0.666 | 0.690 | 0.926 | 0.845 | 0.605 | 1.000 | | | | |
| WEBurger | 0.971 | 0.950 | 0.809 | 0.672 | 0.665 | 0.921 | 0.801 | 0.613 | 0.931 | 1.000 | | | |
| WHRehnquist | 0.970 | 0.939 | 0.876 | 0.637 | 0.729 | 0.972 | 0.873 | 0.672 | 0.937 | 0.964 | 1.000 | | |
| WJBrennan | 0.968 | 0.971 | 0.849 | 0.749 | 0.691 | 0.929 | 0.841 | 0.620 | 0.964 | 0.950 | 0.946 | 1.000 | |
| WODouglas | 0.833 | 0.913 | 0.681 | 0.905 | 0.525 | 0.766 | 0.673 | 0.488 | 0.851 | 0.820 | 0.796 | 0.904 | 1.000 |

supreme court judges in the same geometric space. As done previously, we center all the document vectors on court, topic, and year. We then aggregate by judge. For Table 4, we computed the vector similarity between each circuit court judge and each supreme court judge. We then ranked the circuit court judges by this similarity. The table shows, for each supreme court judge, the top 5 circuit court judges on this ranking. As with the pair-wise similarities between supreme court judges, these rankings are not particularly intuitive or informative.

As mentioned, a possible reason for the lackluster results in the Supreme Court is the small number of opinions they publish. Therefore we round out this analysis by looking at a notable circuit court judge, Richard A. Posner, who published over 3,300 opinions during his tenure. The document vectors are de-meaned by court, year, and topic. Then they are aggregated by judge. Then we rank all circuit court judges by the similarity of their vector to Posner's vector.

These are reported in Table 5. Interestingly, the most similar judge is Frank Easterbrook, who, like Posner is known for the use of economic analysis in opinions. Stephen Breyer has a published article in *The Economic Journal* on "economic reasoning and judicial review" (Breyer, 2009). Posner has a conservative reputation, and we see other conservative judges such as Neil Gorsuch and Antonin Scalia. Henry Friendly makes an appearance – he is a well-known pragmatist, as is Posner. Finally, Michael McConnell co-write law articles with Posner. The document vectors, as trained in this example, seem more informative about the connections between circuit court judges than between Supreme Court judges.

Table 4: Most Similar Circuit Court Judges to each Supreme Court Judge

| **W E Burger** | **A M Kennedy** | **A Scalia** |
|---|---|---|
| MARBLEY, ALGENON L. | SARGUS, EDMUND A., JR. | ROBERTS, VICTORIA A. |
| MURRAY, HERBERT F. | NICKERSON, EUGENE H. | VANCE, SARAH SAVOIA |
| HULL, THOMAS GRAY | NOTTINGHAM, EDWARD WILLIS, JR. | LAKE, SIMEON TIMOTHY, III |
| O'SULLIVAN, CLIFFORD | PECK, JOHN WELD | SHAW, CHARLES A. |
| DOTY, DAVID S. | JOHNSEN, HARVEY | O'NEILL, THOMAS N., JR. |
| **C Thomas** | **D H Souter** | **E Warren** |
| KEELEY, IRENE PATRICIA M. | MOTZ, DIANA GRIBBON | ZAVATT, JOSEPH C. |
| FISHER, JOE J. | MARRERO, VICTOR | DYER, DAVID PATTERSON |
| MCCORD, LEON | DIAMOND, GUSTAVE | SWAN, THOMAS W. |
| SMITH, WILLIAM F. | WANGELIN, H. KENNETH | WHITAKER, SAMUEL |
| KEENAN, BARBARA MILANO | BOOCHEVER, ROBERT | MCCORD, LEON |
| **H A Blackmun** | **H L Black** | **J G Roberts** |
| CORDOVA, VALDEMAR A. | THOMPSON, JOSEPH W. | STEIN, SIDNEY H. |
| SINGLETON, JOHN V., JR. | MINER, ROGER J. | GLEESON, JOHN |
| AGEE, G. STEVEN | MACKINNON, GEORGE E. | WILKINS, WILLIAM W. |
| WHITE, JEFFREY S. | FUSTE, JOSE ANTONIO | MURRAY, HERBERT F. |
| DAVIS, EDWARD BERTRAND | JOHNSON, ALBERT WILLIAMS | VAN DUSEN, FRANCIS |
| **J P Stevens** | **R B Ginsburg** | **S A Alito** |
| PERRY, CATHERINE DELORES | GANEY, J. CULLEN | CAHILL, CLYDE S., JR. |
| GIBSON, KIM R. | FORRESTER, J. OWEN | HARPER, ROY WINFIELD |
| SNEED, JOSEPH T. | CHASE, HARRIE B. | ELLIOTT, JAMES ROBERT |
| JENSEN, D. LOWELL | LEAVY, EDWARD | HIGGINS, THOMAS A. |
| MCKEOWN, M. MARGARET | BEA, CARLOS T. | WEST, SAMUEL H. |
| **S D OConnor** | **S G Breyer** | **S Sotomayor** |
| BARRY, MARYANNE TRUMP | SUTTLE, DORWIN W. | ROBRENO, EDUARDO C. |
| DECKER, BERNARD MARTIN | WOODS, GEORGE E., JR. | PICKERING, CHARLES WILLIS SR. |
| WILKINS, PHILIP C. | FAIRCHILD, THOMAS | NUGENT, DONALD C. |
| BRIGGLE, CHARLES GUY | TEVRIZIAN, DICKRAN M., JR. | FARNAN, JOSEPH J., JR. |
| DOOLING, MAURICE TIMOTHY | WEINFELD, EDWARD | LACEY, FREDERICK B. |
| **T Marshall** | **W H Rehnquist** | **W J Brennan** |
| VAN SICKLE, FREDERICK L. | MCAULIFFE, STEVEN JAMES | RESTANI, JANE A. |
| COFFRIN, ALBERT W. | DUNCAN, ROBERT M. | YOUNG, GORDON E. |
| BOOTLE, WILLIAM A. | KARLTON, LAWRENCE KATZ | NICHOLS, PHILIP, JR. |
| MORTON, L. CLURE | GREEN, CLIFFORD SCOTT | MATSCH, RICHARD P. |
| AGUILAR, ROBERT P. | MCNICHOLS, ROBERT J. | PUTNAM, WILLIAM LE BARON |

Table 5: Most Similar Circuit Court Judges to Richard A. Posner

| Circuit Judge Name | Similarity | Rank | Circuit Judge Name | Similarity | Rank |
|---|---|---|---|---|---|
| POSNER, RICHARD A. | 1.000 | 1 | TONE, PHILIP W. | 0.459 | 16 |
| EASTERBROOK, FRANK H. | 0.663 | 2 | SIBLEY, SAMUEL | 0.459 | 17 |
| SUTTON, JEFFREY S. | 0.620 | 3 | SCALIA, ANTONIN | 0.456 | 18 |
| NOONAN, JOHN T. | 0.596 | 4 | COLLOTON, STEVEN M. | 0.445 | 19 |
| NELSON, DAVID A. | 0.592 | 5 | DUNIWAY, BENJAMIN | 0.438 | 20 |
| CARNES, EDWARD E. | 0.567 | 6 | GIBBONS, JOHN J. | 0.422 | 21 |
| FRIENDLY, HENRY | 0.566 | 7 | BOGGS, DANNY J. | 0.420 | 22 |
| KOZINSKI, ALEX | 0.563 | 8 | BREYER, STEPHEN G. | 0.414 | 23 |
| GORSUCH, NEIL M. | 0.559 | 9 | GOODRICH, HERBERT | 0.412 | 24 |
| CHAMBERS, RICHARD H. | 0.546 | 10 | LOKEN, JAMES B. | 0.410 | 25 |
| FERNANDEZ, FERDINAND F. | 0.503 | 11 | WEIS, JOSEPH F. | 0.408 | 26 |
| EDMONDSON, JAMES L. | 0.501 | 12 | SCALIA, ANTONIN (SCOTUS) | 0.406 | 27 |
| KLEINFELD, ANDREW J. | 0.491 | 13 | BOUDIN, MICHAEL | 0.403 | 28 |
| WILLIAMS, STEPHEN F. | 0.481 | 14 | RANDOLPH, A. RAYMOND | 0.397 | 29 |
| KETHLEDGE, RAYMOND M. | 0.459 | 15 | MCCONNELL, MICHAEL W. | 0.390 | 30 |

# 4 Discussion and Future Work

To recap, we applied the document vectorization algorithm Doc2Vec to twelve decades of opinion texts from the U.S. appellate courts. While previous work has applied LDA to large corpora of legal opinions, this is the first to introduce document embeddings to the area of empirical legal studies. Our analysis of the resulting vectors serve to validate their informativeness, in terms of legal jurisdiction (distinguishing courts), time (distinguishing decades), and topics (distinguishing broad legal areas). These vectors would therefore be useful for downstream prediction or inference tasks on these categories. The method has the advantage of requiring few subjective decisions by the researcher, and the resulting features are easier to work with than high-dimensional sparse representations such as N-grams. Moreover, we have show the

In terms of distinguishing judges, the results are more mixed. The vectors capture judge birth cohort, and provide some intuitive rankings for the similarity of circuit judges to Richard Posner. But the vectors do not show a clear signal for political party or judge law school. Similarities between Supreme Court judges, and the similarities of circuit judges to Supreme Court judges, are not intuitive or informative.

One interpretation of these results is that judicial language is not very politicized along

partisan lines. This would be consistent with the finding in Ash et al. (2017) that judicial language is less polarized than congressional language. In contrast, differences in policy approach, such as use of economic analysis, might be more salient for distinguishing judges. This would be consistent with the finding in Ash, Chen, and Naidu (2018) on the importance of economics (language and training) in judicial decision-making.

But there is the alternative possibility that Doc2vec representations of language are not rich enough to encode some dimensions of judicial ideology. Richer representations, such as those constructed from grammatical relations between words (Levy and Goldberg, 2014), may be needed. Another possibility is that document embeddings may require a large number of documents to form coherent ideological dimensions. Understanding the limitations of Doc2Vec and related models is important for future research. https://www.overleaf.com/project/5bb0ff55389
In the rest of this section, we outline some parallel and future work in using embeddings models for empirical analysis of law. This research may serve to address the limitations identified during the research for this chapter.

## 4.1 Analogies and Associations

The exploratory analysis taken in this chapter focused on distances and similarities. But the geometric nature of embeddings models allow for richer analysis using matrix algebra. In particular, an intriguing use of word embeddings is to solve analogies. A well-known example is that word embeddings "know" that "man" is to "woman" as "king" is to "queen," through the vector algebra king - man + woman = queen (Mikolov et al., 2013). In a legal milieu, Ash (2016) shows that "personal income tax" - "person" + "corporation" = "corporate income tax."

Dai et al. (2015) show that document embeddings also encode analogical relations between documents, with an application to Wikipedia articles. The document vector for the "Christina Aguilera" article, minus the vector for the "America" article, plus the vector for the "Japan" article, results in the vector for the Japanese pop star "Ayumi Hamasaki." In the case of the law, a document embedding could say something like "*Everson vs. Board of Education* is to *Engel v. Vitale* as *Griswold v. Connecticut* is to *Roe v. Wade.*" These cases share an analogical relation, in that the latter case is a related application of the constitutional principle articulated in the former case. In the vector math, that would be represented as *Everson* - *Engel* + *Griswold* = *Roe*. Finally, a judge embedding could say something like "Scalia is to Thomas as Ginsburg is to Breyer," in the sense that Scalia - Thomas + Breyer = Ginsburg.

This discussion of analogies is exemplary of the feature that directions in the embedding space encode semantic meaning, for example those related to singular-vs-plural, verb tense, etc. Bolukbasi et al. (2016) show how to isolate a vector direction for a semantic concept, such as gender, in the embedding space. Construct a list of word pairs who share the gendered analogical relation (man-woman, men-women, boy-girl, boys-girls, father-mother, etc.), and then take the average of the vectors defined by the pairwise differences. This "gender" vector defines a semantic concept, rather than any particular word or pair of words. It can then be used to identify and analyze the use of gendered language.

In the law, we would be interested in isolating other types of language dimensions, notably legal and political concepts and distinctions. For example, there might be a direction for liberal vs conservative, or procedural vs substantive. Are there directions or clusters for originalists, or pragmatists, or economic analysis?

Some of the recent work on embeddings has used these associational features to analyze cultural biases, including sexism and racism. Caliskan et al. (2017) show that the results of implicit association tests are reproduced in aggregate language associations. Garg et al. (2018) and Kozlowski et al. (2018) use long-run historical corpora to analyze trends in biased language over the last century.

The issue of unconscious bias is of particular significance in the legal system (e.g. Fagan and Ash, 2017). Rachlinski et al. (2009) show that trial judges demonstrate the same implicit biases as the broader population on a standard psychological test. But that test was confidential, and it was not matched with the judges' actual decisions. Constructing a language-based measure of bias, available for all judges from their written opinions, would be quite useful for understanding the importance of prejudice in the judicial system.

Toward this end, Ash et al. (2018a) analyze implicit associations in judicial language. The broad descriptive results from Caliskan et al. (2017) are replicated in the judiciary, and relations between "innocent" and "guilty" are also analyzed. Male names tend to have a stronger connotation with "guilty" (relative to "innocent") than female names. In addition, black- and hispanic-associated names are more closely related to guilty than Causian-associated names.

Future work should analyze whether and how these biases in language are associated to biases in decisions. One could ask, for example, whether judges with a lexical bias against blacks also tend to reject discrimination complaints, or to give longer criminal sentences to blacks. Similarly, having more traditional gender views, as detected in one's implicit gender bias, might be reflected in more conservative judicial decisions related to gender discrimination cases. We could also look for peer effects, and see whether sitting with a

biased judge has an impact on a peer judge's subsequent decisions.

## 4.2 Structured Embeddings and Categorical Embeddings

The document embeddings developed in the previous section were trained on the whole corpus. The emebdding model did not explicitly model a time component, a court component, or other metadata categories. Differences across courts, time, and judges were encoded only through aggregating by different categories. Future work might explicitly account for differences between these categories in how the embeddings are constructed.

Along these lines, recent work in embeddings models seeks to include these relations more flexibly and elegantly as a part of the data generating process. Rudolph and Blei (2017) provide a model for learning dynamic embeddings, and look at how language has changed over time in the U.S. Congress over the last century. Rudolph et al. (2017) provide a model for structured group embeddings, and allow word and document vectors to have a group component and an individual component.

In parallel work, we found difficulties in initial applications of structured embeddings to judge groups (Ash et al., 2018a). Word similarities across groups seem to be sensitive to model parameters. Systematic differences in word similarities between Republican and Democrat judges can flip based on the embedding dimension and vocabulary size, for example. While structured embeddings do not work off the shelf, we expect that there is still potential in this research area.

As discussed, the Doc2Vec embeddings did not do a good job of discriminating judges on ideology. This may be because language style of written decisions may not encode ideology. This information may be mostly contained in the direction of the decision (e.g., for or against plaintiffs), or some interaction between the decision and the language. Embedding layers in the the deep learning literature provide an alternative approach for identifying spatial relations between judges in prediction of decisions.

As described, Word2Vec and Doc2Vec work by locating words together that are most similarly predictive for a deep learning task. In that case, a word is the embedded categorical variable, but embedding layers can be used for any sort of categorical variable. In future work, the judge identity could be represented with an embedding lookup layer to a relatively low-dimensional dense vector space. The location of the judge vectors, initialized randomly, would be endogenous to the model. As the model goes through further training, the locations of these vectors would be pushed around to improve predictiveness. As a by-product of the model, the judges that locate together in the vector space would be predicted to behave

similarly on the court, holding other factors constant. This type of model may work to analyze ideological dimensions of judging.

## 4.3 Embedding of Citation Networks

In this chapter, the focus has been on the language of opinions as representing legal ideas. But we all know that in a common law system, the previous cases cited are a major expression of the ideological content of a decision. Ash et al. (2017) show that citations are more predictive of the political party of a judge than the writing style. Therefore, in the context of the geometry of law, citations could be included as features in the document embedding. This might reveal more differences, such as those between political parties.

Another approach to embedding citations is based on Rudolph et al. (2017). In that paper, the model predicts occurrence of a product in a grocery shopping cart based on the co-occurrence of other products. In the legal analogue, cases could be treated as a bundle of citations to precedents, in the same way that Rudolph et al. (2017) treat grocery baskets as a bundle of products. The citation embedding model would predict the presence of a particular citation using the list of co-occurring citations. As with word embeddings, cases that tend to be cited together would locate near each other in the embedding space. The model would thereby construct a "precedent space" as opposed to a language space.

An intriguing feature of the grocery cart model is that the learned parameters encode complementarity or substitutability of items. In the context of Rudolph et al. (2017), that means coffee being substitutable with tea but complementary with milk, for example. In the context of the law, we would learn which precedents are complementary (tending to be cited together), and which are substitutable (tend to appear in similar contexts, but not together). By pairing substitutability metrics with ideological valence (liberal versus conservative), we can analyze the parallel histories of liberal and conservative jurisprudence in the United States.

# References

Ash, E. (2016). The political economy of tax laws in the u.s. states. Technical report. 4.1

Ash, E. (2017). Emerging tools for a 'driverless' legal system: Comment. *Journal of Institutional and Theoretical Economics*. 2

Ash, E., Chen, D., and Liu, W. (2017). The (non-)polarization of u.s. circuit court judges, 1930-2013. Technical report. 4, 4.3

Ash, E., Chen, D. L., and Ornaghi, A. (2018a). Implicit bias in the judiciary: Evidence from judicial language associations. Technical report. 4.1, 4.2

Ash, E., MacLeod, W. B., and Naidu, S. (2018b). The language of contract: Promises and power in union collective bargaining agreements. Technical report. 1

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84. 1, 2

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022. 1

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357. 4.1

Breyer, S. (2009). Economic reasoning and judicial review. *Economic Journal*, 119(535):F215–F135. 3.5

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. 4.1

Cameron, C. and Kornhauser, L. (2017). What courts do . . . and how to model it. Technical report, NYU Law and Economics Research Paper. 1

Carlson, K., Livermore, M. A., and Rockmore, D. (2015). A quantitative analysis of writing style on the us supreme court. *Wash. UL Rev.*, 93:1461. 1, 3.5

Dai, A. M., Olah, C., and Le, Q. V. (2015). Document embedding with paragraph vectors. 3.2, 4.1

Fagan, J. and Ash, E. (2017). New policing, new segregation? from ferguson to new york. *Georgetown Law Journal*. 4.1

Ganglmair, B. and Wardlaw, M. (2017). Complexity, standardization, and the design of loan agreements. Technical report. 1

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644. 4.1

Jurafsky, D. and Martin, J. H. (2014). *Speech and language processing*, volume 3. Pearson London. 1

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2018). The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*. 4.1

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*. 1, 2, 3.2

Leibon, G., Livermore, M., Harder, R., Riddell, A., and Rockmore, D. (2018). Bending the law: geometric tools for quantifying influence in the multinetwork of legal opinions. *Artificial Intelligence and Law*, 26(2):145–167. 1

Levy, O. and Goldberg, Y. (2014). Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308. 4

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225. 1, 2

Livermore, M. A., Riddell, A., and Rockmore, D. (2016). Agenda formation and the us supreme court: A topic model approach. *Arizona Law Review*. 1

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605. 3.4

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119. 1, 2, 4.1

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., and Guthrie, C. (2009). Does unconscious bias affect trial judges? *Notre Dame Law Review*, 84(3):1195–1246. 4.1

Rudolph, M. and Blei, D. (2017). Dynamic bernoulli embeddings for language evolution. *arXiv preprint arXiv:1703.08052*. 4.2

Rudolph, M., Ruiz, F., Athey, S., and Blei, D. (2017). Structured embedding models for grouped data. In *Advances in Neural Information Processing Systems*, pages 250–260. 4.2, 4.3

Songer, D. R. and Haire, S. (1992). Integrating alternative approaches to the study of judicial voting: Obscenity cases in the u.s. courts of appeals. *American Journal of Political Science*, 36(4):963–982. 1