Text as Strategic Choice

Yanchuan Sim

CMU-LTI-16-014

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis committee

Noah Smith (chair), University of Washington
Eduard Hovy, Carnegie Mellon University
Daniel Neill, Carnegie Mellon University
Jing Jiang, Singapore Management University
Philip Resnik, University of Maryland, College Park

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies.

Abstract

People write everyday — articles, blogs, emails — with a purpose and an audience in mind. Politicians adapt their speeches to convince their audience, news media slant stories for their market, and teenagers on social media seek social status among their peers through their posts. Hence, language is *purposeful* and *strategic*. In this thesis, we introduce a framework for text analysis that make explicit the purposefulness of the author and develop methods that consider the interaction between the author, her text, and her audience's responses. We frame the authoring process as a decision theoretic problem — the observed *text* is the result of an author maximizing her utility.

We will explore this perspective by developing a set of novel statistical models that characterize authors' strategic behaviors through their utility functions. We consider three particular domains — political campaigns, the scientific community, and the judiciary — using our models and develop the necessary tools to evaluate our assumptions and hypotheses. In each of these domains, our models yield better response prediction accuracy and provide an interpretable means of investigating the underlying processes. Together, they exemplify our approach to text modeling and data exploration.

Throughout this thesis, we will illustrate how our models can be used as tools for in-depth exploration of text data and hypothesis generation.

Acknowledgements

This thesis marks an important milestone in my journey. A journey that began sixteen years ago with me typing `PRINT "Hello world!"` into the QBasic interpreter.[1] Along the way, I am very very fortunate to have the support of many people and it is my pleasure to thank those who made this possible.

First and foremost, it is difficult to overstate my gratitude to my advisor, Noah Smith, for countless hours of helpful and inspiring advice both in science and in life. Throughout graduate school, he provided invaluable guidance and patience without which I would have gotten lost. I could not have imagined a better advisor and mentor.

Besides my advisor, I would like to thank the rest of my thesis committee: Eduard Hovy, Jing Jiang, Daniel Neill, and Philip Resnik for their insightful comments from which this thesis has greatly benefited. I thank my undergraduate advisor at UIUC, Julia Hockenmaier, for introducing me to natural language processing and patiently teaching me the expectation-maximization algorithm.[2]

In my daily work, I have been blessed with the smartest, friendliest, and most amazing group of ARK-mates: Waleed Ammar, David Bamman, Dallas Card, Elizabeth Clark, Brendan O'Connor, Dipanjan Das, Jesse Dodge, Chris Dyer, Jeff Flanigan, Kevin Gimpel, Swapna Gottipati, Lingpeng Kong[3], Lucy Lin, Fei Liu, Nelson Liu, Kelvin Luu, Bill Mc-

[1] This statement merely sets a lower bound for my true age.

[2] Julia also encouraged us to attend Noah's talk at UIUC on text driven forecasting. I did and the rest is history.

[3] Special thanks for keeping my Mandarin ability in working condition and "making sure" that I have Chinese food more than once a week.

Dowell, George Mulcaire, Minghui Qiu, Rohan Ramanath, Bryan Routledge, Maarten Sap, Nathan Schneider, Swabha Swayamdipta, Sam Thomson, Tae Yano, Dani Yogatama, and Xiaote Zhu. There is no better group of people to be stuck in an ark with than *yinz*. Of course, I am also indebted to my many brilliant student colleagues at CMU and UW for providing a stimulating and fun environment in which to learn and grow.

I thank the Agency for Science, Technology and Research (A*STAR) for the scholarship that supported me over the last 9 years and gave me the incredible opportunity to meet with and learn from the greatest minds in the world. To my fellow friends from home who ended up with me in Pittsburgh over the years: Kawa Cheung, Weiling Neo, Junyang Ng, Jiaqi Tan, Chinyen Tee, Lionel Wong, and Yingying Wu, each of you made Pittsburgh my *kampung Singapura*.

Last but not least, I want to thank my family. Joy Sim, while in my heart you will always be my baby sis, I thank you for being my built-in best friend, for your advice that is wise beyond your years, for listening to my complaints over the last few years, and for being my sister. Mom and dad, thank you for unconditionally supporting me from afar, for raising me, for loving me, and for letting me know that there is always a home for me. 妈妈，爸爸，多谢你们多年的无私付出和养育之恩。这篇论文和我今日的成就一切因你们而有意义。

To my family I dedicate this thesis.

Contents

Chapter 1

Introduction

All models are wrong but some are useful.

George Box (1979)

People write everyday — books, articles, blogs, emails, tweets — with a purpose and an audience in mind. Politicians adapt their speeches to convince their audience, news media slant stories for their market (Gentzkow and Shapiro 2010), and teenagers on social media seek social status among their peers through their posts. Hence, language is *purposeful* and *strategic*. In this thesis, we introduce a framework for text analysis that make explicit the purposefulness of the author and develop methods that consider the interaction between the author, her text, and her audience's responses.

Today, text modeling research centers on finding solutions to these three classes of problems:

1. *What is in the text?*

   A significant part of NLP research — syntactic parsing, semantics, information extraction, etc — is centered around content. NLP researchers are interested in building models and tools that help us understand and synthesize the volumes of text available today. Techniques such as topic modeling, which we will use and build upon in this

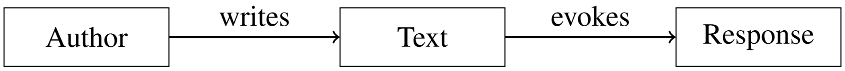


Figure 1.1: The relationship between author, text, and response.

thesis, allow us to efficiently explore text collections and discover hidden thematic structure. This problem is a cornerstone of NLP research, and finding computational representations of text and its meaning will form a foundation for answering the next two questions.

2. *Who wrote the text?*

   This problem is focused on the author. Modern text analysis methods have made significant advancements in learning author (or group) attributes, such as identity (Stamatatos 2009), gender (Bamman et al. 2014a), demographic attributes (Eisenstein et al. 2010), or political orientations (Conover et al. 2011) from text. The text reflects the *social attributes* of the author; a conservative politician is more likely to write about lowering taxes, whereas a teenager will be more likely to tweet about Taylor Swift than Exxon's latest financial statement.

3. *How does the audience respond (to the text)?*

   Text is written for an audience, even if that audience is just the author herself (e.g., a diary); it is important to consider the response of the reader upon reading the text. Therefore, this problem is centered around the response *evoked* by the text. For instance, consider a movie-goer checking up on the latest movie reviews to decide which movies to catch on her next cinema outing; her actions will be influenced by the text (reviews) she consumed. Unsurprisingly, response prediction is an active area of NLP research, whether it is predicting movie revenues from movie reviews (Joshi et al. 2010), blog post's popularity (Yano et al. 2009), retweeting activity (Petrović et al. 2011), roll call votes from bills (Gerrish and Blei 2011), or citations of a scientific paper (Yogatama et al. 2011).

Figure 1.1 represents the current text modeling paradigm we describe. The solutions

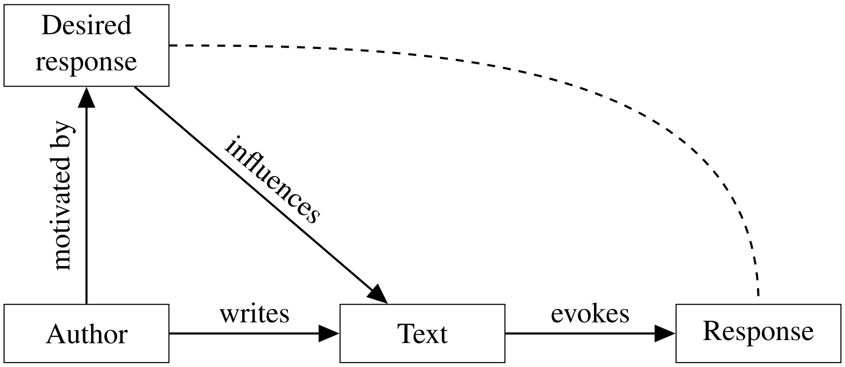


Figure 1.2: Our proposed relationship between author, text and response.

to the first problem ("what") give us representations of the content of text. Using these representations, researchers answer the second problem ("who") by characterizing the authors through their language. Likewise, much research effort has gone into solving the third problem ("how") of modeling readers' responses from text (Kogan et al. 2009). However, current research treats the problems of "who" and "how" separately despite the fact that we know authors are motivated to evoke responses from their audience — politicians give speeches to rally their constituents while editors prioritize reporting on events that interest their readers.

This thesis proposes an alternative framework where we explicitly model the author's motivation (the desired response) and how this influences the text (illustrated in Figure 1.2). The text is still influenced by the author's attributes but also motivated by the "desired response" of the author. In this thesis, models of the desired response often follow (but not limited to) the same modeling assumptions as that for the "true" response function (represented by the dotted line in Figure 1.2) but they do not necessarily share same outcome.[1] We frame this as a decision theoretic problem. We treat the observed *text* as the result of an author maximizing her utility. This utility function encapsulates the author's strategy; within the constraints of her attributes (e.g., her language, political orientation, expertise), she produces text that maximizes the utility of her desired response (e.g., votes, re-tweets,

[1] A politician may tailor her speech to influence her constituents to vote for her but the "true" response is not known until after she gives her speech.

citations).

Consider an example: a lawyer presenting his argument in court. His audience is the judge, who will listen to him (as well as his opposing counsel), and his desired response is for the judge to decide the case in his favor. On the other hand, he is constrained by what he can say; for instance, he must follow procedural rules, invoke legal arguments that are relevant to the case, and stay within the allotted time limit. Within these constraints, he has a choice of arguments and he knows from experience the different "costs" and "benefits" of each. The arguments that we observe are the result of his strategic choice.

1.1 Settings and Assumptions

In this work, we investigate texts where the authors' motivations are evident. We find instances of such texts in the realm of politics and the law, such as presidential campaign speeches and Supreme Court proceedings. The models are often unsupervised, and it may not be straightforward to evaluate their quality. Where possible, we validate our models and methods based on their predictive ability on downstream tasks (e.g., predicting the responses). Otherwise, we perform qualitative assessments of our model by validating them with extrinsic hypotheses; part of the contribution of my thesis is in developing techniques for careful assessments of these models.

We exploit the machinery of probabilistic graphical models[2] in developing these models of text. Graphical models provide a powerful framework for articulating the statistical assumptions and relationships between variables of interest (both observed and hidden). Furthermore, there are well-established approximate inference algorithms[3] for graphical models such as variational methods (Jordan et al. 1999; Wainwright and Jordan 2008) and Markov chain Monte Carlo (MCMC) methods. In this thesis, we regularly use MCMC methods such as Gibbs sampling (Geman and Geman 1984; Steyvers and Griffiths 2006)

[2] See Koller and Friedman (2009) for an in-depth study of probabilistic graphical models.

[3] Exact inference in high dimensional Bayesian networks is NP-hard (Cooper 1990), and therefore we resort to approximate methods.

and Metropolis-Hastings (Hastings 1970; Metropolis et al. 1953) for inference.

As with model-based approaches, our models are simplified views of the world and are not guaranteed to recover the "truth." We depend on our modeling assumptions that could very well be wrong; for example, there may be confounding variables that we fail to take into account. The inference algorithms are approximate; they only give us estimates of the model's parameters, which are often learned from limited and noisy data. Nevertheless, we must exercise caution when drawing causal conclusions from our models and remind readers that the results may simply reflect our modeling assumptions and/or artifacts of the data. While we seek external validation of our model where possible, we also point out pitfalls and limitations associated with our methods when discussing our results and conclusions.

1.2 Organization of this Thesis

This thesis is organized around different settings involving authors and their strategic behaviors. In each setting, we first introduce the data and describe the authoring decisions. Then, we describe a probabilistic model of the data that encodes our assumptions of the relationships between the observed evidence, hidden structures, and our utility function. Finally, we evaluate our models by comparing against simpler treatments of the data.

We start by considering the setting of U.S. presidential campaigns in Chapter 2. We developed a procedure for measuring ideological signaling in campaign speeches which allowed us to find empirical evidence suggestive of strategic behavior of candidates during their campaign.

In Chapter 3, we investigate author utility in the setting of scientific authorship within the computational linguistics community. Our model enables us to characterize scientist authors through their utility functions, which were learned from the data. We found that our author utility model achieves better predictive performance than simpler baselines.

Chapter 4 examines strategic behavior of amici curiae[4] in the Supreme Court of the United States. We present several versions of utility models for the Supreme Court where we use our utility framework to analyze the Supreme Court. We evaluate our approach on vote prediction and achieved improved accuracy over existing methods. In addition, we demonstrate some of the counterfactual analyses that we perform using our approach.

Lastly, Chapter 5 concludes with a discussion of broader issues, future directions, and opportunities for using our utility framework in other research fields.

1.3 Thesis Statement

In this thesis, we argue that text is evidence of an author's strategic behavior: it encodes her social attributes and beliefs about audience responses. We examine this claim by developing a set of novel statistical models that characterize authors' strategic behaviors through their utility functions. These models enable us to make inferences about authors' strategic choices by using text and responses as evidence. We explore three particular domains — political campaigns, the judiciary, and the scientific community — using our models and develop the necessary tools to evaluate our assumptions and hypotheses. In each of these domains, we construct models of authors' strategic behaviors that yield better response prediction accuracy and provide an interpretable means of investigating the underlying processes. Together, they exemplify our approach to text modeling and data exploration.

We view our models as tools for in-depth exploration of text data and hypothesis generation. They enable us to make inferences about how authors behave and provide us with deeper insights into understanding how language is being used for influence. We envision that these models will be useful to authors of actuating texts (e.g., politicians, journalists, lobbyists) by recommending content that will help them better frame their writings to pursue their agendas. From a social science standpoint, this will be an analysis framework that uses text to facilitate a more complex analysis of individual actor's behaviors and offers

[4] Latin for "friends of the court." See Chapter 4 for a brief description of the procedures in the Supreme Court.

new quantitative standpoints for researchers.

Chapter 2

Observing Strategic Behavior in U.S. Presidential Campaign Speeches

The ballot is stronger than the bullet.

Abraham Lincoln

The artful use of language is central to politics, and the language of politicians has attracted considerable interest among scholars of political communication and rhetoric (Charteris-Black 2005; Deirmeier et al. 2012; Hart 2009; Hart et al. 2013) and computational linguistics (Fader et al. 2007; Gerrish and Blei 2011; Thomas et al. 2006, *inter alia*). In American politics, the presidential election season is prime time where candidates running for office give speeches and write books and manifestos expounding their ideas. Election campaign speeches are important — they are heard by millions of people nationwide at rallies, on television, and more recently, through social media. A great campaign speech has the ability to connect with the audience and inspire voters to a candidate's cause. Thus, speeches are an integral part of a candidate's campaign and there are numerous instances where *strategic* *behaviors* can be observed.

In this chapter, we will describe our work previously published as (Sim et al. 2013), and carried out in collaboration with Brice Acree, Justin Gross, and Noah Smith. In Sim et al. (2013), we analyzed political speeches during the 2008 and 2012 U.S. presidential campaigns to identify one such instance of strategic behavior — that successful primary candidates ideologically "move to the center" before a general election. In fact, presidential candidate Mitt Romney's own aide infamously proclaimed in 2012:

> *I think you hit a reset button for the fall campaign (i.e., the general election). Everything changes. It's almost like an Etch-a-Sketch. You can kind of shake it up and we start all over again.*
>
> — Eric Fehrnstrom, Spokesman for 2012 presidential candidate Mitt Romney

The statement became news and a source of derision, but hardly comes as a surprise; the "Etch-a-Sketch" is a perfect analogy for what political scientists expect to happen during a presidential campaign. Classic Downsian median voter theorem predicts that candidates will appeal to median partisan voters during the primary, before converging to the point of the median voter in the national electorate for the general election (Black 1948; Downs 1957; Hotelling 1929). Therefore, it is to be expected that when a set of voters that are more ideologically concentrated are replaced by a set who are more widely dispersed across the ideological spectrum, as occurs in the transition between the United States primary and general elections, that candidates will present themselves as more moderate in an effort to capture enough votes to win. This observation is often stated, but not yet, to our knowledge, tested empirically and we will attempt to do so here.

2.1 Introduction

Do political candidates in fact stray ideologically at opportune moments? More specifically, can we measure candidates' ideological positions from their prose at different times? Following much work on *classifying* the political ideology expressed by a piece of text (Hillard et al. 2008; Laver et al. 2003; Monroe and Maeda 2004), we start from the assumption that a candidate's choice of words and phrases reflects a deliberate attempt to signal common cause with a target audience, and as a broader strategy, to respond to political competitors. Our central hypothesis is that, despite candidates' intentional vagueness, differences in position — among candidates or over time — can be automatically detected and described as *proportions* of ideologies expressed in a speech.

In this work, we operationalize ideologies in a novel empirical way, exploiting political writings published in explicitly ideological books and magazines (§2.2).[1] The corpus then serves as evidence for a probabilistic model that allows us to automatically infer compact, human-interpretable lexicons of cues strongly associated with each ideology.

These lexicons are used, in turn, to create a low-dimensional representation of political speeches: a speech is a sequence of cues interspersed with lags. Lags correspond to the lengths of sequences of non-cue words, which are treated as irrelevant to the inference problem at hand. In other words, a speech is represented as a series alternating between cues signaling ideological positions and uninteresting filler.

Our main contribution is a probabilistic technique for inferring proportions of ideologies expressed by a candidate (§2.3). The inputs to the model are the cue-lag representation of a speech and a domain-specific topology relating ideologies to each other. The topology tree (shown in Figure 2.1) encodes the closeness of different ideologies and, by extension, the odds of transitioning between them within a speech. Bayesian inference is used to manage uncertainty about the associations between cues and ideologies, probabilities of traversing each of the tree's edges, and other parameters.

We demonstrate the usefulness of the measurement model by showing that it accurately recovers pre-registered beliefs regarding narratives widely accepted — but not yet tested empirically — about the 2008 and 2012 U.S. Presidential elections (§2.4).

[1] We consider general positions in terms of broad ideological groups that are widely discussed in current political discourse (e.g., "Far Right," "Religious Right," "Libertarian,"' etc.).

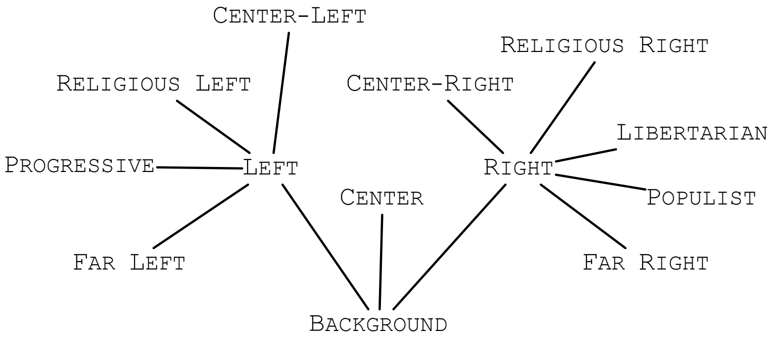


Figure 2.1: Ideology tree showing the labels for the ideological corpus in §2.2.1 (excluding BACKGROUND) and corresponding to states in the HMM (§2.3.3).

2.2 First Stage: Cue Extraction

We first present a data-driven technique for automatically constructing "cue lexicons" from texts labeled with ideologies by domain experts.

2.2.1 Ideological Corpus

We start with a collection of contemporary political writings whose authors are perceived as representative of one particular ideology. Our corpus consists of two types of documents: books and magazines. Books are usually written by a single author, while each magazine consists of regularly published issues with collections of articles written by several authors. Justin Gross, who is a political science domain expert and a co-author of this work, manually labeled each element in a collection of 112 books and 10 magazine titles[2] with one of three coarse ideologies: LEFT, RIGHT, or CENTER. Documents that were labeled LEFT and RIGHT were further broken down into more fine-grained ideologies, shown in Figure 2.1.[3] Table 2.1 summarizes key details about the ideological corpus.

[2] There are 765 magazine issues, which are published biweekly to quarterly, depending on the magazine. All of a magazine's issues are labeled with the same ideology.

[3] We cannot claim that these texts are "pure" examples of the ideologies they are labeled with (i.e., they may contain parts that do not match the label). By finding relatively few terms strongly associated with texts sharing a label, our model should be somewhat robust to impurities, focusing on those terms that are indicative of whatever drew the expert to identify them as (mostly) sharing an ideology.

Total tokens	32,835,190	
Total types	138,235	
Avg. tokens per book	77,628	
Avg. tokens per mag. issue	31,713	
Breakdown by ideology:	*Documents*	*Tokens*
Left	0	0
Far Left	112	3,334,601
Center-Left	196	7,396,264
Progressive Left	138	7,257,723
Religious Left	7	487,844
Center	5	429,480
Right	97	3,282,744
Far Right	211	7,392,163
Libertarian Right	88	1,703,343
Center-Right	9	702,444
Populist Right	5	407,054
Religious Right	6	441,530

Table 2.1: Ideology corpus statistics. Note that not all documents are labeled with finer-grained ideologies.

In addition to ideology labels, individual chapters within the books were manually tagged with topics that the chapter was about. For instance, in Barack Obama's book *The Audacity of Hope*, his chapter titled "Faith" is labeled as Religious. Not all chapters have clearly defined topics, and as such, these chapters are simply labeled Misc. Magazines are not labeled with topics because each issue of a magazine generally touches on multiple topics. There are a total of 61 topics.[4]

[4] The following 61 topics were used to label chapters of books in our ideological corpus: Abortion, Arts, Capitalism, Civil-Liberties, Class, Competition, Congress, Constitution, Courts, Crime, Debt, Defense, Democrats' Lies, Democrats–Failings, Democrats–General, Economy, Education, Elections, Energy, Environment, Family, Federalism, Financial Institutions, Foreign Affairs, Freedom, Future of Nation, GOP Lies, GOP–Failings, GOP–General, Gender, Government Scope, Guns, Health, History (General), History (GW Bush Era), History (Clinton Era), History (Democrats '08), History (GOP '08), History (Obama Era), History (Reagan Era), History (Reagan/Bush Sr.Era), Immigration, Infrastructure, International, Labor, Lobbying, Media, Miscellaneous, Morality, Parties, Patriotism, Personal (Author), Race, Religion, Same-Sex Marriage & Gay Rights, Science, Spending, Tax, Terrorism, Trade, Welfare.

Each magazine was labeled with a document-specific topic. The magazines are: *American Conservative* (Right), *American Prospect* (Progressive Left, *International Socialist Review* (Far Left), *Monthly Review* (Far Left), *Mother Jones* (Progressive Left), *Reason* (Libertarian Right), *The Freeman* (Libertarian Right), *The New American* (Far Right), *The New Republic* (Center-Left), *Z Magazine* (Far Left).

2.2.2 Cue Discovery Model

We use the ideological corpus to infer ideological cues: terms that are strongly associated with an ideology. For example, a candidate who utters the term “death tax” rather than “estate tax”, or “pro-abortion” instead of “pro-choice” is clearly signaling her ideological tendencies. Some instances of distinguishing word choices are a result of explicit partisan marketing strategies, but others may represent more subtle clues (Luntz 2007). Thus, we propose an method for automatically inferring these cue terms from our labeled ideological corpus. Because our ideologies are organized hierarchically, we required a technique that can account for multiple effects within a single text. We further require that the sets of cue terms be small, so that they can be inspected by domain experts. We therefore turn to the sparse additive generative (SAGE) models introduced by Eisenstein et al. (2011).

Like other probabilistic language models, SAGE assigns probability to a text as if it were a bag of terms. It differs from most language models in parameterizing the distribution using a generalized linear model, so that different effects on the log-odds of terms are additive. In our case, we define the probability of a term w conditioned on attributes of the text in which it occurs. These attributes include both the ideology and its coarsened version (e.g., a FAR RIGHT book also has the attribute RIGHT). For simplicity, let $\mathcal{A}(d)$ denote the set of attributes of document d and $\mathcal{A} = \bigcup_d \mathcal{A}(d)$. The parametric form of the distribution is given, for term w in document d, by:

$$p(w \mid \mathcal{A}(d); \boldsymbol{\eta}) = \frac{\exp\left(\eta_w^0 + \sum_{a \in \mathcal{A}(d)} \eta_w^a\right)}{Z(\mathcal{A}(d), \boldsymbol{\eta})}$$

Each of the η weights can be a positive or negative value influencing the probability of the word, conditioned on various properties of the document. When we stack an attribute a’s weights into a vector across all words, we get an $\boldsymbol{\eta}^a$ vector, understood as an effect on the term distribution. (We use $\boldsymbol{\eta}$ to refer to the collection of all of these vectors.) The effects in our model, described in terms of attributes, are:

- $\boldsymbol{\eta}^0$, the background (log) frequencies of words, fixed to the empirical frequencies in the corpus. Hence the other effects can be understood as *deviations* from this background distribution.

- $\boldsymbol{\eta}^{i_c}$, the coarse ideology effect, which takes different values for LEFT, RIGHT, and CENTER.
- $\boldsymbol{\eta}^{i_f}$, the fine ideology effect, which takes different values for the fine-grained ideologies corresponding to the leaves in Figure 2.1.
- $\boldsymbol{\eta}^{t}$, the topic effect, taking different values for each of the 61 manually assigned topics. We further include one effect for each magazine series (of which there are 10) to account for each magazine's idiosyncrasies (topical or otherwise).
- $\boldsymbol{\eta}^{d}$, a document-specific effect, which captures idiosyncratic usage within a single document.

Note that the effects above are not mutually exclusive, although some effects never appear together due to constraints imposed by their semantics (e.g., no book is labeled both LEFT and RIGHT).

When estimating the parameters of the model (the $\boldsymbol{\eta}$ vectors), we impose a sparsity-inducing ℓ_1 prior that forces many weights to zero. The objective is:

$$\max_{\boldsymbol{\eta}} \sum_{d} \sum_{w \in d} \log p(w \mid \mathcal{A}(d); \boldsymbol{\eta}) - \sum_{a \in \mathcal{A}} \lambda_a \|\boldsymbol{\eta}^a\|_1$$

This objective function is convex but requires special treatment due to non-differentiability when any elements are zero; we use the OWL-QN algorithm to solve it (Andrew and Gao 2007).[5]

[5] Our approach to learning $\boldsymbol{\eta}$ differs from Eisenstein et al. (2011). They postulated that the components of $\boldsymbol{\eta}$ are drawn from a compound model $\int \mathcal{N}(\eta; \mu, \sigma)\mathcal{E}(\sigma; \tau) d\sigma$, where $\mathcal{E}(\sigma; \tau)$ indicates the Exponential distribution and learned the parameters using variational methods (Beal 2003).

In fact, the compound model described above is equivalent to the Laplace distribution $\mathcal{L}(\eta; \mu, \tau)$ (Figueiredo 2003; Lange and Sinsheimer 1993). Moreover, a zero mean Laplace prior has the same effect as placing an L_1 regularizer on $\boldsymbol{\eta}$ which can be optimized easily using standard convex optimization approaches. In cases where the labels are known *a priori*, we have found this approach to infer SAGE vectors to be simpler and just as effective. We released software at `https://bitbucket.org/skylander/ark-sage/` to learn SAGE vectors using our approach.

To reduce the complexity of the hyperparameter space (the possible values of all λ_a) and to encourage similar levels of sparsity across the different effect vectors, we let, for each ideology attribute a,

$$\lambda_a = \frac{|\mathcal{V}(a)|}{\max_{a' \in \mathcal{A}} |\mathcal{V}(a')|}\lambda$$

where $\mathcal{V}(a)$ is the set of term types appearing in the data with attribute a (i.e., its vocabulary) , and λ is a hyperparameter we can adjust to control the amount of sparsity in the SAGE vectors. For the non-ideology effects, we fix $\lambda_a = 10$ (not tuned).

2.2.3 Bigram and Trigram Lexicons

After estimating parameters, we are left with sparse $\boldsymbol{\eta}^a$ for each attribute. We are only interested, however, in the ideological attributes $\mathcal{I} \subset \mathcal{A}$. For an ideological attribute $i \in \mathcal{I}$, we take the terms with positive elements of this vector to be the cues for ideology i; call this set $\mathcal{L}(i)$ and let $\mathcal{L} = \bigcup_{i \in \mathcal{I}} \mathcal{L}(i)$.

Because political texts use a fair amount of multi-word jargon, we initially represented each document as a bag of unigrams, bigrams, and trigrams, ignoring the fact that these “overlap” with each other.[6] While this would be inappropriate in language modeling and is inconsistent with our model’s independence assumptions among words, it is sensible since our goal is to identify cues that are statistically associated with attributes like ideologies.

Preliminary trials revealed that unigrams tend to dominate in such a model, since their frequency counts are so much higher. Further, domain experts found them harder to interpret out of context compared to bigrams and trigrams. We therefore included only bigrams and trigrams as terms in our final cue discovery model.

[6] Generative models that produce the same evidence more than once are sometimes called “deficient,” but model deficiency does not necessarily imply that the model is ineffective. Some of the IBM models for statistical machine translation provide a classic example (Brown et al. 1993).

2.2.4 Validation

The term selection method we have described can be understood as a form of feature selection that reasons globally about the data and tries to control for some effects that are not of interest (topic or document idiosyncrasies). We compared the approach to two classic, simple methods for feature selection: ranking based on pointwise mutual information (PMI) and weighted average PMI (WAPMI) (Cover and Thomas 2006; Schneider 2005). Selected features were used to classify the ideologies of held-out documents from our corpus.[7] We evaluated these feature selection methods within naïve Bayes classification in a 5-fold cross-validation setup. We vary λ for the SAGE model and compare the results to equal-sized sets of terms selected by PMI and WAPMI. We consider SAGE with and without topic effects.

Figure 2.2 visualizes accuracy against the number of features for each method. Bigrams and trigrams consistently outperform unigrams (McNemar's, $p < 0.05$). Otherwise, there are no significant differences in performance except WAPMI with bigrams/trigrams at its highest point. SAGE with topics is slightly (but not significantly) better than without. We conclude that SAGE is a competitive choice for cue discovery, noting that a principled way of controlling for topical and document effects — offered by SAGE but not the other methods — may be even more relevant to our task than classification accuracy.

We ran SAGE on the the full ideological book corpus, including topic effects, and setting $\lambda = 30$, obtained a set of $|\mathcal{L}| = 8,483$ cue terms. Table 2.2 presents top cue terms associated with various ideologies and Figure 2.3 shows a heatmap of similarities among SAGE vectors.

Informal study As a check of face validity for the vocabularies uncovered by SAGE, we invited several people — four scholars of American politics and three U.S. citizens with a moderate to high interest in contemporary politics — to examine a set of top terms

[7] The text was tokenized and stopwords removed. Punctuation, numbers, and web addresses were normalized. Tokens appearing less than 20 times in training data, or in fewer than 5 documents were removed.

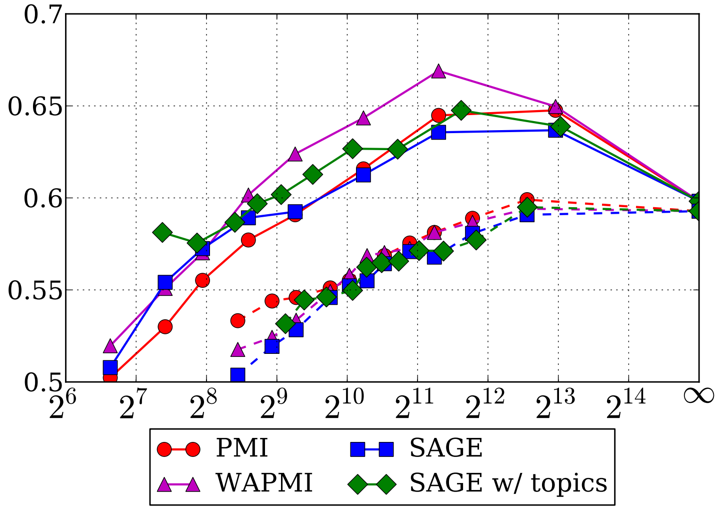


Figure 2.2: Plot of average classification accuracy for 5-fold cross validation against the number of features. Dashed lines refer to using only unigram features, while solid lines refer to using bigram and trigram features.

for each fine-grained ideological category. They were given brief descriptions of each class, including prominent prototypical individuals exemplifying each, and asked to match term sets to ideologies. On average, respondents correctly identified about 70% of ideologies using only a handful of terms from each. Experts correctly matched coarse ideologies (LEFT and RIGHT) to appropriate lists of top terms 76% of the time (85% LEFT and 70% RIGHT). Several fine (and fairly distinct) ideologies were correctly labeled by all, or nearly all, respondents (LIBERTARIAN, FAR LEFT, RELIGIOUS LEFT and RELIGIOUS RIGHT). Centrist ideologies were sometimes confused but nearly always identified as one of the three centrist categories (CENTER-RIGHT, CENTER-LEFT and CENTER). Of sub-ideologies on the left and right, all but one — POPULIST RIGHT — were given the correct label more than any other label. Relatively few mistakes were made in which LEFT and RIGHT were mixed up.

Far Left (2,802)	monopoli_capit, class_struggl, capitalist_economi, social_movement, occupi_movement, polit_economi, capitalist_system, trade_union, labor_movement, rule_class, develop_countri, world_bank, work_peopl, labor_power, econom_crisi
Progressive Left (2,319)	#_peopl, recent_year, abu_ghraib, #_state, execut_director, public_info, state_depart, public_polici, vice_presid, #_centuri, mental_ill, john_kerri, make_sens, polit_parti, presidenti_elect
Religious Left (941)	biolog_famili, progress_religion, nuclear_famili, mother_teresa, bad_theologi, religi_issu, earli_church, tax_collector, god_love, religi_commun, american_creed, earli_christian, luke_#, church_leader, matthew_#
Left (2,580)	north_carolina, econom_polici, execut_director, public_opinion, cell_phone, mental_ill, #_state, abu_ghraib, earli_#, decad_ago, west_bank, presidenti_elect, good_job, air_forc, homeland_secur
Center-Left (3,050)	stanlei_kauffmann, modern_art, young_woman, eighteenth_centuri, al_jazeera, nineteenth_centuri, good_deal, young_man, twentieth_centuri, long_ago, mitt_romnei, great_deal, presidenti_campaign, twenti_year, al_qaeda
Center (1,230)	long_beach, debt_limit, stock_option, countri_music, averag_american, corpor_america, origin_intent, georg_washington, cousin_john, tax_increas, loan_offic, alexand_hamilton, debt_ceil, park_lot, proof_text
Center-Right (1,450)	governor_bush, class_voter, health_care, republican_presid, georg_bush, state_polic, move_forward, miss_america, middl_eastern, water_buffalo, fellow_citizen, sam_club, polit_career, american_life, work_class
Right (2,415)	mcmanu_sourc, foreign_aid, north_korea, nation_review, north_american, georg_washington, communist_parti, arm_forc, emphasi_ad, european_union, limit_govern, constitut_convent, presid_georg, presidenti_candid, #_minut
Libertarian (2,268)	medic_marijuana, realiti_taught, intuit_tempt, raw_milk, rand_paul, econom_freedom, health_care, govern_intervent, market_economi, commerc_claus, militari_spend, govern_agenc, due_process, drug_war, govern_polici
Populist Right (1,155)	corpor_america, work_men, border_secur, nation_interest, big_busi, nation_media, birth_rate, hundr_year, special_interest, million_peopl, american_citizen, immigr_law, open_border, mass_immigr, border_patrol
Religious Right (960)	d&c_#, dai_saint, holi_spirit, matthew_#, john_#, jim_walli, modern_liber, individu_liberti, posit_law, god_word, jesu_christ, elementari_school, natur_law, limit_govern, emerg_church
Far Right (2,410)	mcmanu_sourc, ron_paul, north_american, emphasi_ad, american_citizen, foreign_aid, european_union, world_govern, communist_parti, north_korea, constitut_convent, govern_spend, #_centuri, #_amend, polic_offic

Table 2.2: Top 20 cue terms associated with each coarse/fine-grained ideology, with the total number of terms in brackets. The terms are ordered by log-deviation weights in the $\boldsymbol{\eta}^i$ vectors. # denotes any numeral.

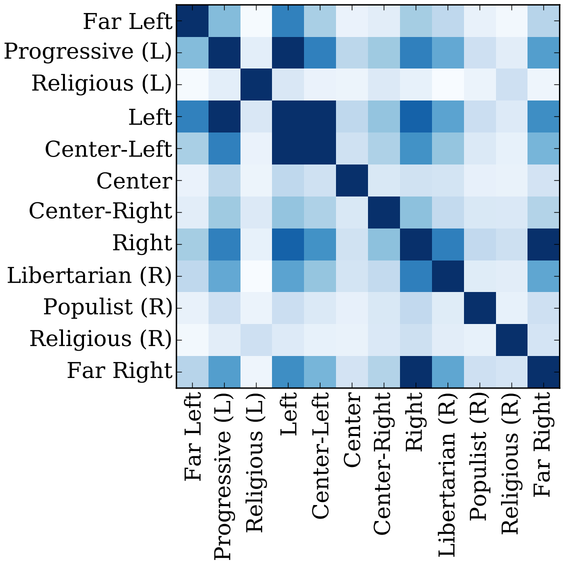


Figure 2.3: Heatmap showing Jaccard similarity between SAGE ideology vectors. Darker shade implies more overlap between the corresponding SAGE effect vectors. We notice that religious and populist cues are outliers.

2.3 Second Stage: Cue-Lag Ideological Proportions

In this section, we propose a technique for measuring ideology proportions in the prose of political candidates. We adopt a Bayesian approach that manages our uncertainty about the cue lexicon $\mathcal{L}$, the tendencies of political speakers to "flip-flop" among ideological types, and the relative "distances" among different ideologies. The representation of a candidate's ideology as a mixture among discrete, hierarchically related categories can be distinguished from continuous representations ("scaling" or "spatial" models) often used in political science, especially to infer positions from Congressional roll-call voting patterns (Clinton et al. 2004; Poole and Rosenthal 1985, 2000). Moreover, the ability to draw inferences about individual policy-makers' ideologies from their votes on proposed legislation is severely limited by institutional constraints on the types of legislation that is actually subject to recorded votes.

2.3.1 Political Speeches Corpus

We gathered transcribed speeches given by candidates of the two main parties (Democrats and Republicans) during the 2008 and 2012 Presidential election seasons. Each election season is comprised of two stages: (i) the primary elections, where candidates seek the support of their respective parties to be nominated as the party's Presidential candidate, and (ii) the general elections where the parties' chosen candidates travel across the states to garner support from all citizens. Each candidate's speeches are partitioned into *epochs* for each election; e.g., those that occur before the candidate has secured enough pledged delegates to win the party nomination are from the "primary", while those after are from the "general".

Table 2.3 presents a breakdown of the candidates and speeches in our corpus.

Party	Primary 2008	General 2008	Primary 2012	General 2012
Democrats*	167	-	-	-
Republicans†	50	-	49	-
Obama (D)	78	81	-	99
McCain (R)	9	159	-	-
Romney (R)	8	‡(13)	19	19

*Democrats in our corpus are: Joe Biden, Hillary Clinton, John Edwards, and Bill Richardson in 2008 and Barack Obama in both 2008 and 2012.

†Republicans in our corpus are: Rudy Giuliani, Mike Huckabee, John McCain, and Fred Thompson in 2008, Michelle Bachmann, Herman Cain, Newt Gingrich, Jon Huntsman, Rick Perry, and Rick Santorum in 2012, and Ron Paul and Mitt Romney in both 2008 and 2012.

‡For Romney, we have 13 speeches which he gave in the period 2008–2011 (between his withdrawal from the 2008 elections and before the commencement of the 2012 elections). While these speeches are not technically part of the regular Presidential election campaign, they can be seen as his preparation towards the 2012 elections, which is particularly interesting as Romney has been accused of having inconsistent viewpoints.

Table 2.3: Breakdown of number of speeches in our political speech corpus by epoch. On average, 2,998 tokens, and 95 cue terms are found in each speech document.

2.3.2 Cue-Lag Representation

Our measurement model only considers ideological cues; other terms are treated as filler. We therefore transform each speech into a **cue-lag** representation.

The representation is a sequence of alternating cues (elements from the ideological lexicon $\mathcal{L}$) and integer "lags" (counts of non-cue terms falling between two cues). This will allow us to capture the intuition that a candidate may use longer lags between evocations of different ideologies, while nearby cues are likely to be from similar ideologies.

To map a speech into the cue-lag representation, we simply match all elements of $\mathcal{L}$ in the speech and replace sequences of other words by their lengths. When a trigram cue strictly includes a bigram cue, we take only the trigram. When two cues partially overlap, we treat them as consecutive cue terms and set the lag to 0. Figure 2.4 shows an example of our cue-lag representation.

Original sentence	Just compare this President's record with **Ronald Reagan's** first term. **President Reagan** also faced an **economic crisis**. In fact, in 1982, the **unemployment rate** peaked at nearly 11 percent. But in the two years that followed, he delivered a true recovery – **economic growth** and **job creation** were three times higher than in the Obama Economy.
Cue-lag representation	$\ldots \xrightarrow{6}$ ronald_reagan $\xrightarrow{2}$ presid_reagan $\xrightarrow{3}$ econom_crisi $\xrightarrow{5}$ unemploy_rate $\xrightarrow{17}$ econom_growth $\xrightarrow{1}$ job_creation $\xrightarrow{9} \ldots$

Figure 2.4: Example of the cue-lag representation.

2.3.3 CLIP: An Ideology HMM

The model we use to infer ideologies, **cue-lag ideological proportions** (CLIP), is a hidden Markov model. Each state corresponds to an ideology (Figure 2.1) or BACKGROUND. The emission from a state consists of (i) a cue from $\mathcal{L}$ and (ii) a lag value. The high-level generative story for a single speech with T cue-lag pairs is as follows:

1. Parameters are drawn from conjugate priors (details in §2.3.3).

2. Let the initial state be the BACKGROUND state.

3. For $i \in \{1, 2, \ldots, T\}$:[8]

   (a) Transition to state s_i based on the transition distribution, discussed in §2.3.3. This transition is conditioned on the previous state s_{i-1} and the lag at timestep $i-1$, denoted by ℓ_{i-1}.

   (b) Emit cue term w_i from the lexicon $\mathcal{L}$ and lag ℓ_i based on the emission distribution, discussed in §2.3.3.

We turn next to the transitions and emissions.

Ideology Topology and Transition Parameterization

CLIP assumes that each cue term uttered by a politician is generated from a hidden state corresponding to an ideology. The ideologies are organized into a tree based on their hierarchical relationships; see Figure 2.1. In this study, the tree is fixed according to our domain knowledge of current American politics; in future work it might be enriched with greater detail or its structure learned automatically.

The ideology tree is used in defining the transition distribution in the HMM, but not to directly define the topology of the HMM. Importantly, each state may transition to any other state, but the transition *distribution* is defined using the graph, so that ideologies that are closer to each other will tend to be more likely to transition to each other. To transition between two states s_i and s_j, a walk must be taken in the tree from vertex s_i to vertex s_j. We emphasize that the walk corresponds to a *single* transition — the speaker does not emit anything from the states passed through along the path.

A simplified version of our transition distribution, for exposition, is given as follows:

$$p_{tree}(s_j \mid s_i; \boldsymbol{\zeta}, \boldsymbol{\theta}) = \left(\prod_{\langle u,v\rangle \in Path(s_i,s_j)} (1-\zeta_u)\theta_{u,v}\right) \zeta_{s_j}$$

[8] The length of the sequence is assumed to be exogenous, so that no stop state needs to be defined.

$Path(s_i, s_j)$ refers to the sequence of edges in the tree along the unique path from s_i to s_j. Each of these edges $\langle u, v \rangle$ must be traversed, and the probability of doing so, conditioned on having already reached u, is $(1 - \zeta_u)$ — i.e., not stopping in u — times $\theta_{u,v}$ — i.e., selecting vertex v from among those that share an edge with u. Eventually, s_j is reached, and the walk ends, incurring probability ζ_{s_j}.

In order to capture the intuition that a longer lag after a cue term should increase the entropy over the next ideology state, we introduce a **restart** probability, which is conditioned on the length of the most recent lag, ℓ. The probability of restarting the walk from the BACKGROUND state is a noisy-OR model with parameter ρ. This gives the transition distribution:

$$\begin{aligned} p(s_j \mid s_i, \ell; \boldsymbol{\zeta}, \boldsymbol{\theta}, \rho) = {} & (1-\rho)^{\ell+1} p_{tree}(s_j \mid s_i; \boldsymbol{\zeta}, \boldsymbol{\theta}) \\ & + (1 - (1-\rho)^{\ell+1}) p_{tree}(s_j \mid s_{\text{BACKGROUND}}; \boldsymbol{\zeta}, \boldsymbol{\theta}) \end{aligned}$$

Note that, if $\rho = 1$, there is no Markovian dependency between states (i.e., there is always a restart), so CLIP reverts to a mixture model.

This approach allows us to parameterize the full set of $|\mathcal{I}|^2$ transitions with $O(|\mathcal{I}|)$ parameters.[9] Since the graph is a tree and the walks are not allowed to backtrack, the only ambiguity in the transition is due to the restart probability; this distinguishes CLIP from other algorithms based on random walks (Brin and Page 1998; Collins-Thompson and Callan 2005; Mihalcea 2005; Toutanova et al. 2004).

Emission Parameterization

Recall that, at time step t, CLIP emits a cue from the lexicon $\mathcal{L}$ and an integer-valued lag. For each state s, we let the probability of emitting cue w be denoted by $\psi_{s,w}$; $\boldsymbol{\psi}_s$ is a multinomial distribution over the entire lexicon $\mathcal{L}$. This allows our approach to handle

[9] More precisely, there are $|\mathcal{I}|$ edges (since there are $|\mathcal{I}| + 1$ vertices including BACKGROUND), each with a $\boldsymbol{\theta}$-parameter in each direction. For a vertex with degree d, however, there are only $d - 1$ degrees of freedom, so that there are $2|\mathcal{I}| - (|\mathcal{I}| + 1) = |\mathcal{I}| - 1$ degrees of freedom for $\boldsymbol{\theta}$. There are $|\mathcal{I}|$ $\boldsymbol{\zeta}$-parameters and a single ρ, for a total of $2|\mathcal{I}|$ degrees of freedom.

ambiguous cues that can associate with more than one ideology, and also to associate a cue with a different ideology than our cue discovery method proposed, if the signal from the data is sufficiently strong. We assume each lag to be generated by a Poisson distribution with global parameter ν.

Inference and Learning

Above we described CLIP's transitions and emissions. Because our interest is in measuring proportions — and, as we will see, in *comparing* those proportions across speakers and campaign periods — we require a way to allow variation in parameters across different conditions. Specifically, we seek to measure differences in time spent in each ideology state. This can be captured by allowing each speaker to have a different $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ in each stage of the campaign. On the other hand, we expect that a speaker draws from his ideological lexicon similarly across different epochs — there is a single $\boldsymbol{\psi}$ shared between different epochs.

In order to manage uncertainty about the parameters of CLIP, to incorporate prior beliefs based on our ideology-specific cue lexicons $\{\mathcal{L}(i)\}$, and to allow sharing of statistical strength across conditions, we adopt a Bayesian approach to inference. This will allow principled exploration of the posterior distribution over the proportions of interest.

We place a symmetric Dirichlet prior, α, on the tree walk probabilities $\boldsymbol{\theta}$. For the cue emission distribution associated with ideology i, $\boldsymbol{\psi}_{s_i}$, we use an *informed* Dirichlet prior with two different values, β_{cue} for cues in $\mathcal{L}(i)$, and a smaller β_{def} for those in $\mathcal{L} \setminus \mathcal{L}(i)$.[10]

Exact inference in this model is intractable, so we resort to an approximate inference technique based on Markov Chain Monte Carlo simulation. As the Dirichlet distributions are conjugate priors to the multinomial, we can integrate out the latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. For a speech d from epoch e at the ith term, we jointly sample its ideology $s_{e,d,i}$ and the

[10] This implies that a term can, in the posterior distribution, be associated with an ideology i of whose $\mathcal{L}(i)$ it was not a member. In fact, this occurred frequently in our runs of the model.

restart indicator $r_{e,d,i}$ conditioned on that of all other terms. For simpler notation, we drop the document and epoch level subscripts and denote $\boldsymbol{s}_{-i}$ and $\boldsymbol{r}_{-i}$ as the current state and restart assignments for all other terms except the ith term in document d. The sampling equation for s_i and r_i is given by:

$$\begin{aligned} p(s_i = k, r_i = r \mid \boldsymbol{s}_{-i}, \boldsymbol{r}_{-i}, \boldsymbol{w}, \boldsymbol{t}, \alpha, \boldsymbol{\beta}, \rho, \boldsymbol{\zeta}) & \\ & \propto [(1-\rho)^r (1-(1-\rho))^{(1-r)}]^{\ell_i + 1} \\ & \times \frac{n_{k,w_i}^{-i} + \beta_{k,w_i}}{\sum_{w \in \Sigma} n_{k,w}^{-i} + \beta_{k,w}} \times \zeta_k \\ & \times \left[\prod_{(u,v) \in Path(v_0,k)} \frac{(1-\zeta_u)(f_{u,v}^{-i} + \alpha)}{\sum_{v' \in \mathcal{I}} f_{u,v'}^{-i} + \alpha} \right]^r \\ & \times \left[\prod_{(u,v) \in Path(s_{i-1},k)} \frac{(1-\zeta_u)(f_{u,v}^{-i} + \alpha)}{\sum_{v' \in \mathcal{I}} f_{u,v'}^{-i} + \alpha} \right]^{1-r} \\ & \times \left[\prod_{(u,v) \in Path(k,s_{i+1})} \frac{(1-\zeta_u)(f_{u,v}^{-i} + \alpha)}{\sum_{v' \in \mathcal{I}} f_{u,v'}^{-i} + \alpha} \right]^{1-r_{i+1}} \end{aligned} \tag{2.1}$$

where $n_{k,w}^{-i}$ is the number of times word w is generated by ideology k and $f_{u,v}^{-i}$ is the number of times we traversed the edge (u, v) of the ideology tree during epoch e.

At each iteration, we performed collapsed Gibbs sampling to resample the ideology state ($\boldsymbol{s}$) and restart indicator variable ($\boldsymbol{r}$) for every cue term in every speech according to Equation (2.1). Likewise, we used slice sampling (with vague priors) for the hyperparameters, α, $\boldsymbol{\beta}$, ρ, and $\boldsymbol{\zeta}$.

We ran our Gibbs sampler for 75,000 iterations, discarding the first 25,000 iterations for burn-in, and collected samples at every 10 iterations. Further, we perform the slice sampling step at every 5,000 iterations. For each candidate, we collected 5,000 posterior samples which we use to infer her ideological proportions.

To determine the amount of time a candidate spends in each ideology, we denote the unit of time in terms of half the lag before and after each cue term, i.e., when a candidate draws a cue term from an ideology during timestep i, we say that he spends $\frac{1}{2}(\ell_{i-1} + \ell_i)$ amount of time in that ideology. Averaging over all the samples returned by our sampler and normalizing it by the length of the documents in each epoch, we obtain a candidate's expected ideological proportions within an epoch.

2.4 Pre-registered Hypotheses

The traditional way to evaluate a text analysis model in NLP is, of course, to evaluate its output against gold-standard judgements by humans. In the case of recent political speeches, however, we are doubtful that such judgments can be made objectively at a fine-grained level. While we are confident about gross categorization of books and magazines in our ideological corpus (§2.2.1), many of which are *overtly* marked by their ideological assocations, we believe that human estimates of ideological proportions, or even association of particular tokens with ideologies they may evoke, may be overly clouded by the variation in annotator ideology and domain expertise.

We therefore adopt a different method for evaluation — **pre-registration**. Pre-registration is a practice which involves researchers committing to their research predictions and methods before starting their experiments. It is commonly performed in other fields such as psychology and social sciences.[11] Monogan (2013) presents a case for study registration in the field of political science, arguing that it encourages transparency in experimental design and reporting results[12]. Here, we have adopted it in a data-driven NLP setting to deal with subjective texts. Our approach has been embraced by NLP researchers working with subjective text domains such as literature (Bamman et al. 2014b), and we anticipate that pre-registration will be used by researchers when dealing with computational models of subjective texts.

[11] In political science, the Political Science Registered Studies Dataverse (`http://dvn.iq.harvard.edu/dvn/dv/registration`) is a repository where scholars can document their research designs and expectations before completing their experiments.

[12] Gelman (2013) published a commentary on supporting preregistration of studies from a statistician's perspective.

Before running our model, we identified a set of hypotheses, which we pre-registered as expectations. These expectations are categorized into groups based on their strength and relevance to judging the validity of the model. *Strong* hypotheses are those that constitute the lowest bar for face validity; if violated, they suggest a flaw in the model. *Moderate* hypotheses are those that match the intuition of domain experts conducting the research, or extant theory. Violations suggest more examination is required, and may raise the possibility that further testing might be pursued to demonstrate the hypothesis is false. Our 13 principal hypotheses are enumerated in Table 2.4.

2.5 Evaluation

We compare the posterior proportions inferred by CLIP with several baselines:

- HMM: rather than our CLIP model, a fully connected, traditional transition matrix is used.
- MIX: a mixture model; at each timestep, we *always* restart ($\rho = 1$). This eliminates Markovian dependencies between ideologies at nearby timesteps, but still uses the ideology tree in defining the probabilities of each state through $\boldsymbol{\theta}$.
- NORES: where we *never* restart ($\rho = 0$). This strengthens the Markovian dependencies.

In MIX, there are no temporal effects between cue terms, although the structure of our ideology tree encourages the speaker to draw from coarse-grained ideologies over fine-grained ideologies. On the other hand, the strong Markovian dependency between states in NORES would encourage the model to stay local within the ideology tree. In our experiments, we will see how that the ideology tree and the random treatment of restarting both contribute to our model's inferences.

Table 2.4 presents a summary of which hypotheses the models' inferences are in accordance with. CLIP is not consistently outperformed by any of the competing base-

Hypotheses	CLIP	HMM	Mix	NoRes
Sanity checks (strong):				
S1. Republican primary candidates should tend to draw more from Right than from Left.	*12/12	10/13	13/13	12/13
S2. Democratic primary candidates should tend to draw more from Left than from Right.	4/5	5/5	5/5	5/5
S3. In general elections, Democrats should draw more from the Left than the Republicans and vice versa for the Right.	4/4	4/4	3/4	0/4
S total	20/21	19/22	21/22	17/22
Primary hypotheses (strong):				
P1. Romney, McCain and other Republicans should almost never draw from Far Left, and extremely rarely from Progressive.	29/32	*21/31	27/32	29/32
P2. Romney should draw more heavily from the Right than Obama in both stages of the 2012 campaign.	2/2	2/2	1/2	1/2
Primary hypotheses (moderate):				
P3. Romney should draw more heavily on words from the Libertarian, Populist, Religious Right, and Far Right in the primary compared to the general election. In the general election, Romney should draw more heavily on Center, Center-Right and Left vocabularies.	2/2	2/2	0/2	2/2
P4. Obama should draw more heavily on words from the Progressive in the 2008 primary than in the 2008 general election.	0/1	0/1	0/1	1/1
P5. In the 2008 general election, Obama should draw more heavily on the Center, Center-Left, and Right vocabularies than in the 2008 primary.	1/1	1/1	1/1	1/1
P6. In the 2012 general election, Obama should sample more from the Left than from the Right, and should sample more from the Left vocabularies than Romney.	2/2	2/2	0/2	0/2
P7. McCain should draw more heavily from the Far Right, Populist, and Libertarian in the 2008 primary than in the 2008 general election.	0/1	1/1	1/1	1/1
P8. In the general 2008, McCain should draw more heavily from the Center, Center-Right, and Left vocabularies than in the 2008 primary.	1/1	1/1	1/1	1/1
P9. McCain should draw more heavily from the Right than Obama in both stages of the campaign.	2/2	2/2	2/2	1/2
P10. Obama and other Democrats should very rarely draw from Far Right.	6/7	5/7	7/7	4/7
P total	45/51	37/50	40/51	41/51

Table 2.4: Pre-registered hypotheses used to validate the measurement model and the number of statements evaluated correctly by different models. *Some differences were not significant at $p = 0.05$ and are not included in the results.

lines.

2.5.1 Sanity Checks — S1, S2, and S3

CLIP correctly identifies sixteen LEFT/RIGHT alignments of primary candidates (S1, S2), but is unable to determine one candidate's orientation; it finds Jon Huntsman to spend roughly equal proportions of speech-time drawing on LEFT and RIGHT cue terms. Interestingly, Huntsman, who had served as U.S. Ambassador to China under Obama, was considered the one moderate in the 2012 Republican field. MIX correctly identifies all thirteen Republicans, while NORES places McCain from the 2008 primaries as mostly LEFT-leaning and HMM misses three of thirteen, including Perry and Gingrich, who might be deeply disturbed to find that they are misclassified as LEFT-leaning. As for the Democratic primary candidates (S2), CLIP's one questionable finding is that John Edwards spoke slightly more from the RIGHT than the LEFT. For the general elections (S3), CLIP and HMM correctly identify the relative amount of time spent in LEFT/RIGHT between Obama and his Republican competitors. NORES had the most trouble, missing all four. CLIP finds Obama spending slightly more time on the RIGHT than on the LEFT in the 2008 general elections but nevertheless, Obama is still found to spend more time engaging in LEFT-speak than McCain.

2.5.2 Name Interference

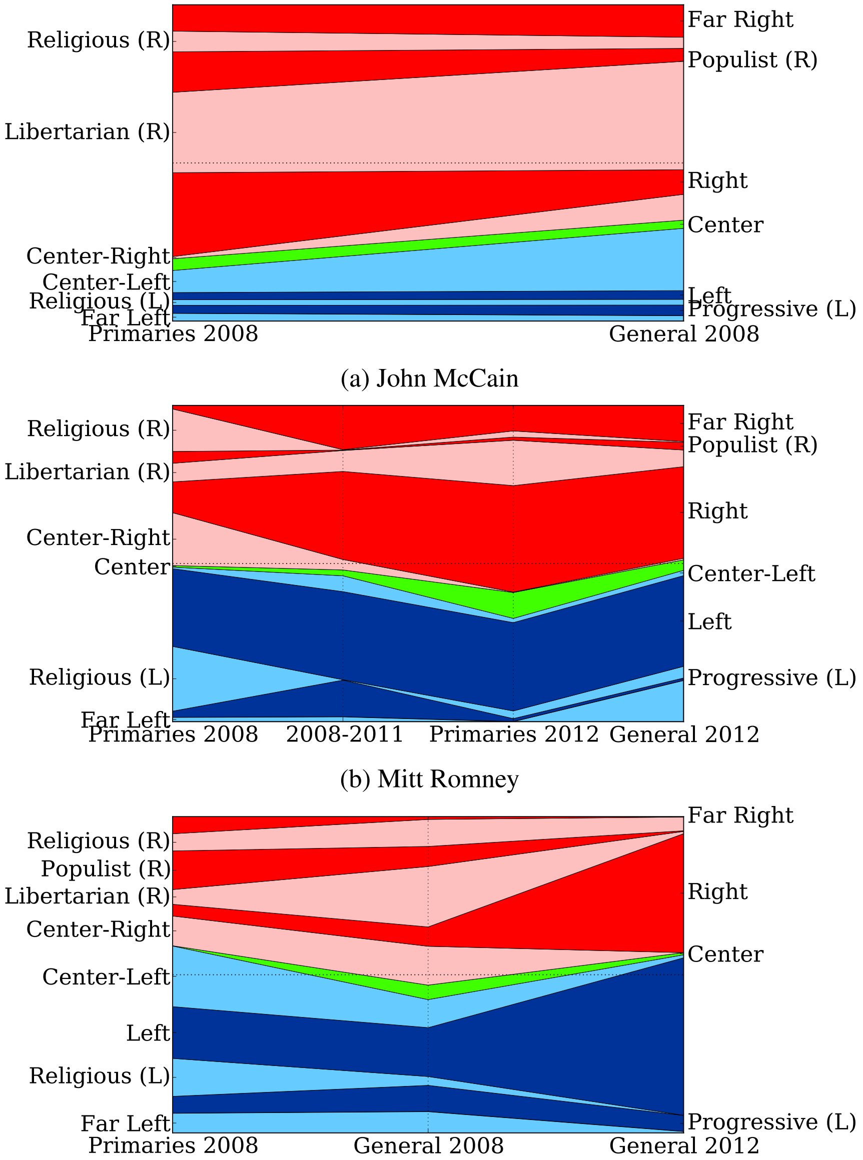
When we looked at the cue terms actually used in the speeches, we found one systematic issue: the inclusion of candidates' names as cue terms. Terms mentioning John McCain are associated with the RIGHT, so that Barack Obama's mentions of his opponent are taken as evidence for rightward positioning; in total, mentions of McCain contributed 4% absolute to Obama's RIGHT ideological proportion. In future work, we believe filtering candidate names in the first stage will be beneficial.

2.5.3 Strong Hypotheses — P1 and P2

CLIP and the variants making use of the ideology tree were in agreement on most of the strong primary hypotheses. Most of these involved our expectation that the Republican candidates would rarely draw on FAR LEFT and PROGRESSIVE LEFT. Our qualitative hypotheses were not specific about how to quantify "rare" or "almost never." We chose to find a result inconsistent with a P1 hypothesis any time a Republican had proportions greater than 5% for either ideology. The notable deviations for CLIP were Fred Thompson (13% from the PROGRESSIVE LEFT during the 2008 primary) and Mitt Romney (12% from the PROGRESSIVE LEFT between the 2008 and 2012 elections, 13% from the FAR LEFT during the 2012 general election). This model did no worse than other variants here and much better than one: HMM had 10 inconsistencies out of 32 opportunities, suggesting the importance of the ideology tree as a source of prior knowledge.

2.5.4 "Etch-a-Sketch" Hypotheses

Hypotheses P3, P4, P5, P7, and P8 are all concerned with differences between the primary and general elections: successful primary candidates are expected to "move to the center." A visualization of CLIP's proportions for McCain, Romney, and Obama is shown in Figure 2.5, with their speeches grouped together by different epochs. The model is in agreement with most of these hypotheses. It did not confirm P4 — Obama appears to CLIP to be more PROGRESSIVE in the 2008 general election than in the primary, though the difference is small (3%) and may be within the margin of error. Likewise, in P7, the difference between McCain drawing from FAR RIGHT, POPULIST and LIBERTARIAN between the 2008 primary and general elections is only 2% and highly uncertain, with a 95% credible interval of 44–50% during the primary (vs. 47–50% in the general election).



(a) John McCain

(b) Mitt Romney

(c) Barack Obama

Figure 2.5: Proportion of time spent in each ideology by John McCain, Mitt Romney, and Barack Obama during the 2008 and 2012 Presidential election seasons.

2.5.5 Fine-grained Ideologies

Fine-grained ideologies are expected to account for smaller proportions, so that making predictions about them is quite difficult. This is especially true for primary elections, where a broader palette of ideologies is expected to be drawn from, but we have fewer speeches from each candidate. CLIP's inconsistency with P10, for example, comes from assigning 5.4% of Obama's 2008 primary cues to FAR RIGHT.

2.5.6 Aggregated Time Spent on Left vs Right

Figure 2.6 shows the proportions of time spent in LEFT and RIGHT, aggregating over candidates, by season and party, according to CLIP. We note that general election candidates (shaded) tend to be closer to the center.

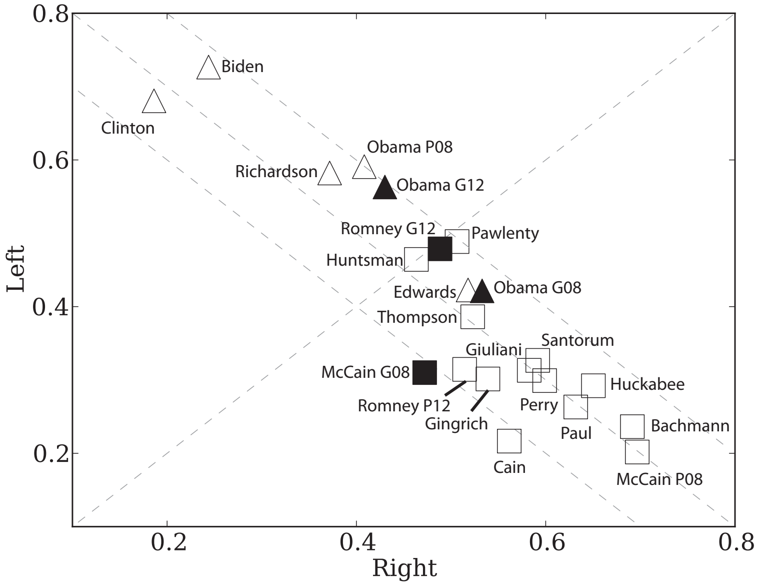


Figure 2.6: Proportions of time spent in LEFT and RIGHT vocabularies for presidential candidates during the 2008 and 2012 presidential election seasons for CLIP. Republican candidates are denoted in squares, while Democratic party candidates are denoted in triangles. Solid and hollow points denote the primary and general election seasons respectively.

CLIP's inferences on the corpus of political speeches can be browsed at http://www.ark.cs.cmu.edu/CLIP.

2.5.7 2016 Presidential Primaries

We downloaded the speeches of the current primaries using transcripts of videos uploaded to `http://www.c-span.org/` for the candidates — Hilary Clinton and Bernie Sanders for the Democrats, Ted Cruz and Donald Trump for the Republicans. The dataset includes rally speeches from each candidate's announcement up until April 21, 2016. When running the CLIP model, we used the same cue lexicon, which is extracted from books published before 2012; this means that the cue words may not be kept up to date with new terms that are used in this election season.[13] Hence, we should take caution when interpreting the results of CLIP on the 2016 primaries. Table 2.5 illustrates some of the top cue terms used in the different election years.

2008	2012	2016 primary
__NUM___million	god_bless	middl_class
health_insur	insur_compani	america_great
nation_secur	wall_street	__NUM___million
presid_bush	auto_industri	make_america
year_ago	small_busi	suprem_court
__NUM___year	move_forward	year_ago
white_hous	work_hard	presid_obama
tax_cut	young_peopl	__MONEY___million
georg_bush	tax_break	south_carolina
small_busi	class_famili	small_busi
__NUM___centuri	creat_job	__NUM___peopl
middl_class	__MONEY___trillion	__MONEY___billion
john_mccain	__NUM___million	american_peopl
wall_street	american_peopl	health_care
__MONEY___billion	presid_obama	wall_street
senat_mccain	health_care	young_peopl
senat_obama	__NUM___year	__NUM___year
american_peopl	tax_cut	hillari_clinton
health_care	middl_class	donald_trump
unit_state	unit_state	unit_state

Table 2.5: Top 20 cue terms used in each of the last 3 election seasons.

[13] Previously, about 6.3% of the tokens are found in the cue lexicon, while it is 5.4% for the 2016 speeches. Since the cue extraction stage is done independently of the CLIP stage, it means that we can easily update the cue lexicons by having domain experts label recently published "ideological" books to obtain new cue words.

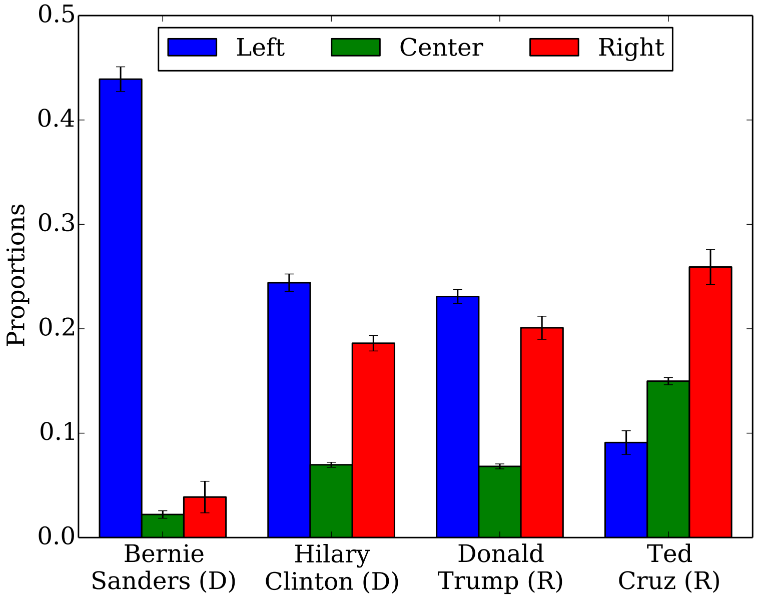


Figure 2.7: Proportions of time spent in LEFT, CENTER, and RIGHT vocabularies for key candidates during the 2016 presidential election seasons estimated by CLIP.

Figure 2.7 illustrates the ideological proportions of the CLIP model on the 2016 primaries. Bernie Sanders and Ted Cruz are known as the more ideologically extreme candidates of their respective parties and this is reflected in the ideological proportions in Figure 2.7. However, the model found little difference in proportions between Hilary Clinton and Donald Trump, which suggests that it could be an artifact of using outdated cue lexicon.

2.6 Related Work

Many political science theories rely upon the concept of ideology, yet ideology is never directly observed (Converse 2006). Instead, researchers attempt to infer ideology based on other characteristics — policy preferences, partisanship, demographics, roll call votes, survey responses, *etc*. As early as the 1960s, automated systems has been used to model ideological beliefs at a sophisticated level, involving the actors, actions and goals, but they require manually constructed knowledge bases (Abelson and Carroll 1965; Carbonell 1978;

Sack 1994). Poole and Rosenthal (1985) introduced the ideal points model, which is based on item response theory, to infer latent positions of congressional lawmakers from roll call data and many researchers have built on their groundbreaking work in many ways (Clinton et al. 2004; Jackman 2001; Londregan 1999; Martin and Quinn 2002, *inter alia*).

However, unlike roll-call votes, texts such as speeches and manifestos allow for considerable nuance and richness, which will aid scholars in accurately placing political actors in an ideological space. In the political science community, Laver et al. (2003) and Slapin and Proksch (2008) present popular techniques for extract political positions on a left-right spectrum from text documents by modeling the frequency of word occurrences; while Gerrish and Blei (2011) and Gerrish and Blei (2012a) augmented roll call votes with text from congressional bills using probabilistic topic models to uncover legislators' positions on specific political issues, putting them in a multidimensional ideological space. Likewise, taking advantage of the proliferation of text today, numerous techniques have been developed to identify topics and perspectives in the media (Fortuna et al. 2009; Gentzkow and Shapiro 2005, 2010; Lin et al. 2008); determine the political leanings of a document or author (Efron 2004; Fader et al. 2007; Laver et al. 2003; Mullen and Malouf 2006); recognize stances in debates (Anand et al. 2011; Somasundaran and Wiebe 2009); finding association between social media behavior and campaign contributions (Yano et al. 2013); or predict legislators' votes using floor debate speeches (Thomas et al. 2006). Going beyong lexical indicators, Greene and Resnik (2009) investigated syntactic features to identify perspectives or implicit sentiment.

2.7 Conclusions

We emphasize that CLIP and its variants are intended to quantify the ideological content candidates express in *speeches*, not necessarily their *beliefs* (which may not be perfectly reflected in their words), or even how they are described by pundits and analysts (who draw on far more information than is expressed in speeches). CLIP's deviations from the hypotheses are suggestive of potential improvements to cue extraction (§2.2), but also of incorrect hypotheses. We expect future research to explore a richer set of linguistic cues and attributes beyond ideology (e.g., topics and framing on various issues). We see CLIP as a potentially powerful text analysis method to support substantive inquiry in political science, such as following trends in expressed ideology over time.

In this chapter, we introduced CLIP, a domain-informed, Bayesian model of ideological proportions in political language. We also showed how ideological cues could be discovered from a lightly labeled corpus of ideological writings, and then incorporated into CLIP. Together, we find empirical evidence of strategic behavior and ideological signaling in candidates' speeches. However, in the CLIP model, we did not explicitly take into account the attributes of a candidate nor the responses he seeks; we will consider models that do so in the subsequent chapters.

Chapter 3

Modeling Strategic Behavior in the Scientific Community

> The difference between the almost right word and the right word is really a large matter. ’tis the difference between the lightning bug and the lightning.
>
> *Mark Twain, The Wit and Wisdom of Mark Twain (1987)*

The work described in this chapter is previously published as Sim et al. (2015).

As we have seen in Chapter 2, politicians exhibit strategic behavior in response to change in their audiences’ demographics. Despite a politician’s “shift towards the ideological center,” he tends to remain on the same side, i.e., a Democrat stays on the left of center and vice versa for a Republican. After all, the speech is a result of many decisions; one of which is making a trade off between his individual preferences and appealing to his constituents. This is a form of strategy and the balancing act between two (or more needs) can be inferred from their textual artifacts.

In this chapter, we will examine a different instantiation of strategic behavior — authoring scientific papers — and consider explicitly the trade offs facing a scientist-author.

Just like campaign speeches, authoring a scientific paper is a complex process involving a many decisions that may be influenced by factors such as institutional incentives, attention-seeking, and pleasure derived from research on topics that excite us. Here, we propose that text collections and associated metadata can be analyzed to reveal optimizing behavior by authors. Using the ACL Anthology Network Corpus (Radev et al. 2013), we introduce a probabilistic model of some of the important aspects of that process: that authors have individual preferences and that writing a paper requires trading off among the preferences of authors as well as extrinsic rewards in the form of community response to their papers. Furthermore, the preferences (of individuals and the community) and tradeoffs may vary over time.

Thus, the author utility model incorporates assumptions about how authors decide what to write, how joint decisions work when papers are coauthored, and how individual and community preferences shift over time. Central to our model is a low-dimensional topic representation shared by authors (in defining individual author's preferences), papers (i.e., what they are "about"), and the community as a whole (in responding with citations). Empirically, we find that

1. topics discovered by generative models outperform a strong text regression baseline (Yogatama et al. 2011) for citation count prediction;
2. such models do better at that task *without* modeling author utility as we propose; and
3. the author utility model leads to better predictive accuracy when answering the question, "given a set of authors, what are they likely to write?"

It can also be used for exploration and to generate hypotheses. In §3.3, we provide an intriguing example relating author tradeoffs to age within the research community.

Unlike Chapter 2, our methods in this chapter infer two kinds of quantities about an author: her associations with interpretable research topics, which might correspond to relative expertise or merely to preferences among topics to write about; and a tradeoff coefficient that estimates the extent to which she writes papers that will be cited versus papers close to her preferences. These two quantities allows us to characterize an author's utility function and better model her strategic behavior.

3.1 Author Utility Model

3.1.1 Notation and Representations

In the following, a document d will be represented by a vector $\boldsymbol{\theta}_d \in \mathbb{R}^K$. The dimensions of this vector might correspond to elements of a vocabulary, giving a "bag of words" encoding; in this chapter they correspond to latent topics.

Document d is assumed to elicit from the scientific community an observable response y_d, which might correspond to the number of citations of the paper.

Each author a is associated with a vector $\boldsymbol{\eta}_a \in \mathbb{R}^K$, with dimensions indexed the same as documents. Below, we will refer to this vector as a's "preferences," though it is important to remember that they could also capture an author's *expertise*, and the model makes no attempt to distinguish between them. We use "preferences" because it is a weaker theoretical commitment.

We describe the components of our model — author utility (§3.1.2), coauthorship (§3.1.3), topics (§3.1.4), and temporal dynamics (§3.1.5) — then give the full form in §3.1.6.

3.1.2 Modeling Utility

Our main assumption about author a is that she is an optimizer: when writing document d she seeks to increase the response y_d while keeping the contents of d, $\boldsymbol{\theta}_d$, "close" to her preferences $\boldsymbol{\eta}_a$. We encode her objectives as a utility function to be maximized with respect to $\boldsymbol{\theta}_d$:

$$U(\boldsymbol{\theta}_d) = \kappa_a y_d - \frac{1}{2}\|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2 \tag{3.1}$$

where $\boldsymbol{\epsilon}_{d,a}$ is an author-paper-specific idiosyncratic randomness that is unobserved to us but assumed known to the author. This is a common assumption in discrete choice models. It is often called a "random utility model" (McFadden 1974).

Notice the tradeoff between maximizing the response y_d and staying close to one's preferences. We capture these competing objectives by formulating the latter as a squared Euclidean distance between $\boldsymbol{\eta}_a$ and $\boldsymbol{\theta}_d$, and encoding the tradeoff between extrinsic (citation-seeking) and intrinsic (preference-satisfying) objectives as the (positive) coefficient κ_a. If κ_a is large, a might be understood as a citation-maximizing agent; if κ_a is small, a might appear to care much more about writing certain kinds of papers ($\boldsymbol{\eta}_a$) than about citation.

This utility function considers only two particular facets of author writing behavior; it does not take into account other factors that may contribute to an author's objective. For this reason, some care is required in interpreting quantities like κ_a. For example, divergence between a particular $\boldsymbol{\eta}_a$ and $\boldsymbol{\theta}_d$ might suggest that a is open to new topics, not merely hungry for citations. Other motivations, such as reputation (notoriously difficult to measure), funding maintenance, and the preferences of peer referees are not captured in this model. Similarly for preferences $\boldsymbol{\eta}_a$, a large value in this vector might reflect a's skill or the preferences of a's sponsors rather than a's personal interest in the topic.

Next, we model the response y_d. We assume that responses are driven largely by topics, with some noise, so that

$$y_d = \boldsymbol{\beta}^\top \boldsymbol{\theta}_d + \xi_d \tag{3.2}$$

where $\xi_d \sim \mathcal{N}(0, 1)$. Because the community's interest in different topics varies over time, $\boldsymbol{\beta}$ is given temporal dynamics, discussed in §3.1.5.

Under this assumption, the author's *expected* utility assuming she is aware of $\boldsymbol{\beta}$ (often called "rational expectations" in discrete choice models), is:

$$\mathbb{E}[U(\boldsymbol{\theta}_d)] = \kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta}_d - \frac{1}{2}\|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2$$

(This is obtained by plugging the expected value of y_d, from Eq. 3.2, into Eq. 3.1.)

An author's decision will therefore be

$$\hat{\boldsymbol{\theta}}_d = \arg\max_{\boldsymbol{\theta}} \kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta} - \frac{1}{2}\|\boldsymbol{\theta} - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2$$

Optimality implies that $\hat{\boldsymbol{\theta}}_d$ solves the first-order equations

$$\kappa_a \beta_j - (\hat{\theta}_{d,j} - (\eta_{a,j} + \epsilon_{d,a,j})) = 0, \; \forall 1 \leq j \leq K \tag{3.3}$$

Eq. 3.3 highlights the tradeoff the author faces: when $\beta_j > 0$, the author will write more on $\theta_{d,j}$, while straying too far from $\eta_{a,j}$ incurs a penalty.

3.1.3 Modeling Coauthorship

Matters become more complicated when multiple authors write a paper together. Suppose the document d is authored by set of authors $\boldsymbol{a}_d$. We model the joint expected utility of $\boldsymbol{a}_d$ in writing $\boldsymbol{\theta}_d$ as the average of the group's utility.[1]

$$\mathbb{E}[U(\boldsymbol{\theta}_d)] = \frac{1}{|\boldsymbol{a}_d|} \sum_{a \in \boldsymbol{a}_d} \left(\kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta}_d - \frac{1}{2} c_{d,a} \|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2 \right)$$

where the "cost" term is scaled by $c_{d,a}$, denoting the fractional "contribution" of author a to document d. Thus, $\sum_{a \in \boldsymbol{a}_d} c_{d,a} = 1$, and we treat $\boldsymbol{c}_d$ as a latent categorical distribution to be inferred. The first-order equation becomes

$$\sum_{a \in \boldsymbol{a}_d} \kappa_a \boldsymbol{\beta} - c_{d,a}(\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})) = \mathbf{0} \tag{3.4}$$

[1] This assumption is a convenient starting place, but we can imagine revisiting it in future work. For example, an economist and a linguist with different expertise might derive "utility" from the collaboration that is nonlinear in each one's individual preferences (Anderson 2012). Further, contributions by complementary authors are not expected to be independent of each other.

3.1.4 Modeling Document Content

As noted before, there are many possible ways to represent and model document content $\boldsymbol{\theta}_d$. We treat $\boldsymbol{\theta}_d$ as (an encoding of) a mixture of topics. Following considerable past work, a "topic" is defined as a categorical distribution over observable tokens (Blei et al. 2003a; Hofmann 1999). Let $\boldsymbol{w}_d$ be the observed bag of tokens constituting document d. We assume each token is drawn from a mixture over topics:

$$p(\boldsymbol{w}_d \mid \boldsymbol{\theta}_d) = \sum_{\boldsymbol{z}_d} \prod_{i=1}^{N_d} p(z_{d,i} \mid \boldsymbol{\theta}_d) p(w_{d,i} \mid \boldsymbol{\phi}_{z_{d,i}})$$

where N_d is the number of tokens in document d, $z_{d,i}$ is the topic assignment for d's ith token $w_{d,i}$, and $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K$ are topic-term distributions. Note that $\boldsymbol{\theta}_d \in \mathbb{R}^K$; we define $p(z \mid \boldsymbol{\theta}_d)$ as a categorical draw from the softmax-transformed $\boldsymbol{\theta}_d$ (Blei and Lafferty 2007).

Using topic mixtures instead of a bag of words provides us with a low-dimensional interpretable representation that is useful for analyzing authors' behaviors and preferences. Each dimension j of an author's preference is grounded in topic j. If we ignore document responses, this component of model closely resembles the author-topic model (Rosen-Zvi et al. 2004), except that we assumed a logistic-normal prior instead of Dirichlet prior for the topic mixtures.

3.1.5 Modeling Temporal Dynamics

Individual preferences shift over time, as do those of the research community. We extend our model to allow variation at different timesteps. Let $t \in \langle 1, \ldots, T \rangle$ index timesteps (in our experiments, each t is a calendar year). We let $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\eta}_a^{(t)}$, and $\kappa_a^{(t)}$ denote the community's response coefficients, author a's preferences, and author a's tradeoff coefficient at timestep t.

Again, we must take care in interpreting these quantities. Do changes in community interest drive authors to adjust their preferences or expertise? Or do changing author preferences aggregate into community-wide shifts? Or do changes in the economy or funding availability change authors' tradeoffs? Our model cannot differentiate among these different causal patterns. Our method is useful for tracking these changes, but it does not provide an explanation for *why* they take place.

Modeling the temporal dynamics of a vector-valued random variable can be accomplished using a multivariate Gaussian distribution. Following Yogatama et al. (2011), we assume the prior for $\boldsymbol{\beta}_j^{(\cdot)} = \langle \beta_j^{(1)}, \ldots, \beta_j^{(T)} \rangle$ has a tridiagonal precision matrix $\Lambda(\lambda, \alpha) \in \mathbb{R}^{T \times T}$:

$$\Lambda(\lambda, \alpha) = \lambda \begin{pmatrix} 1+\alpha & -\alpha & 0 & 0 & \ldots \\ -\alpha & 1+2\alpha & -\alpha & 0 & \ldots \\ 0 & -\alpha & 1+2\alpha & -\alpha & \ldots \\ 0 & 0 & -\alpha & 1+2\alpha & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

The two hyperparameters α and λ capture, respectively, autocorrelation (the tendency of $\beta_j^{(t+1)}$ to be similar to $\beta_j^{(t)}$) and overall variance. This approach to modeling time series allows us to capture temporal dynamics while sharing statistical strength of evidence across all time steps.

We use the notation $\mathcal{T}(\lambda, \alpha) \equiv \mathcal{N}(\mathbf{0}, \Lambda(\lambda, \alpha))$ for this multivariate Gaussian distribution, instances of which are used as priors over response coefficients $\boldsymbol{\beta}$, author preferences $\boldsymbol{\eta}_a$, and (transformed) author tradeoffs $\log \kappa_a$.

3.1.6 Full Model

Table 3.1 summarizes all of the notation. The log-likelihood our model assigns to a collec-

Observed evidence

$w_{d,i}$	ith token in document d
V	vocabulary size
N_d	number of tokens in document d
y_d	response to document d
$\mathcal{A}$	the set of authors
$\boldsymbol{a}_d$	set of authors of document d ($\subseteq \mathcal{A}$)
T	number of timesteps
$\mathcal{D}_t$	the set of documents from timestep t
$\mathcal{D}$	the set of all documents ($= \bigcup_{t=1}^{T} \mathcal{D}_t$)

Latent variables

$\boldsymbol{\beta}^{(t)}$	response coefficients at time t ($\in \mathbb{R}^K$)
$\boldsymbol{\eta}_a^{(t)}$	author a's topic preferences at time t ($\in \mathbb{R}^K$)
$\kappa_a^{(t)}$	author a's tradeoff coefficient at time t ($\in \mathbb{R}_{\geq 0}$)
$\boldsymbol{\theta}_d$	document d topic associations ($\in \mathbb{R}^K$)
$c_{d,a}$	author a contrtibution to document d ($\sum_{a \in \boldsymbol{a}_d} c_{d,a} = 1$)
$\boldsymbol{\phi}_k$	distribution over terms for topic k
$z_{d,i}$	topic assignment of $w_{d,i}$

Constants and hyperparameters

K	number of topics
ρ	symmetric Dirichlet hyperparameter for $\boldsymbol{\phi}_k$
σ_c^2	variance hyperparameter for author contributions $\boldsymbol{c}_d$
$\{\lambda^{(\beta)}, \alpha^{(\beta)}\}$, $\{\lambda^{(\eta)}, \alpha^{(\eta)}\}$, $\{\lambda^{(\kappa)}, \alpha^{(\kappa)}\}$	hyperparameters for priors of $\boldsymbol{\beta}, \boldsymbol{\eta}$, and $\log \kappa$ respectively

Table 3.1: Table of notation.

tion of documents $\mathcal{D}$ and their responses is:

$$\begin{aligned} \mathcal{L} = \log p(\boldsymbol{\beta}) &+ \sum_{d \in \mathcal{D}} \log p(\boldsymbol{c}_d) \\ &+ \sum_{d \in \mathcal{D}} \log p(y_d \mid \boldsymbol{\theta}_d, \boldsymbol{\beta}) + \log p(\boldsymbol{w}_d \mid \boldsymbol{\theta}_d) \\ &+ \sum_{a \in \mathcal{A}} \log p(\boldsymbol{\eta}_a) + \log p(\kappa_a) \\ &+ \sum_{d \in \mathcal{D}} \sum_{a \in \boldsymbol{a}_d} \log p(\boldsymbol{\theta}_d \mid \boldsymbol{\beta}, \boldsymbol{\eta}_a, \kappa_a, c_{d,a}) \end{aligned} \tag{3.5}$$

We adopt a Bayesian approach to parameter estimation. The generative story, including all priors, is as follows. Recall that $\mathcal{T}(\cdot,\cdot)$ denotes the time series prior discussed in §3.1.5. See also the plate diagram for the graphical model in Figure 3.1.

1. For each topic $k \in \{1, \ldots, K\}$:

   (a) Draw response coefficients $\boldsymbol{\beta}_k^{(\cdot)} \sim \mathcal{T}(\lambda^{(\beta)}, \alpha^{(\beta)})$ and term distribution $\boldsymbol{\phi}_k \sim$ Dirichlet(ρ).

   (b) For each author $a \in \mathcal{A}$, draw preference strengths for topic k over time, $\langle \eta_{a,k}^{(1)}$, …, $\eta_{a,k}^{(t)} \rangle \sim \mathcal{T}(\lambda^{(\eta)}, \alpha^{(\eta)})$.

2. For each author $a \in \mathcal{A}$, draw (transformed) tradeoff parameters $\langle \log \kappa_a^{(1)}$, …, $\log \kappa_a^{(T)} \rangle$ $\sim \mathcal{T}(\lambda^{(\kappa)}, \alpha^{(\kappa)})$.

3. For each timestep $t \in \{1, \ldots, T\}$, and each document $d \in \mathcal{D}_t$:

   (a) Draw author contributions $\boldsymbol{c}_d \sim$ Softmax$(\mathcal{N}(\mathbf{0}, \sigma_c^2 \mathbf{I}))$. This is known as a logistic-normal distribution (Aitchison and Shen 1980).

   (b) Draw d's topic distributions (this distribution is discussed further below):

$$\boldsymbol{\theta}_d \sim \mathcal{N}\left(\sum_{a \in \boldsymbol{a}_d} \kappa_a^{(t)} \boldsymbol{\beta}^{(t)} + c_{d,a} \boldsymbol{\eta}_a^{(t)}, \|\boldsymbol{c}_d\|_2^2 \mathbf{I}\right) \tag{3.6}$$

   (c) For each token $i \in \{1, \ldots, N_d\}$, draw topic $z_{d,i} \sim$ Categorical(Softmax$(\boldsymbol{\theta}_d)$) and term $w_{d,i} \sim$ Categorical$(\boldsymbol{\phi}_{z_{d,i}})$.

   (d) Draw response $y_d \sim \mathcal{N}\left(\boldsymbol{\beta}^{(t_d)\top} \boldsymbol{\theta}_d, 1\right)$; note that it collapses out ξ_d, which is drawn from a standard normal.

Eq. 3.6 captures the choice by authors $\boldsymbol{a}_d$ of a distribution over topics $\boldsymbol{\theta}_d$. Assuming that the $\boldsymbol{\epsilon}_{d,a}$s are i.i.d. and Gaussian, from Eq. 3.4, we get

$$\boldsymbol{\theta}_d = \sum_{a \in \boldsymbol{a}_d} \kappa_a \boldsymbol{\beta} + c_{d,a} \boldsymbol{\eta}_a + c_{d,a} \boldsymbol{\epsilon}_{d,a},$$

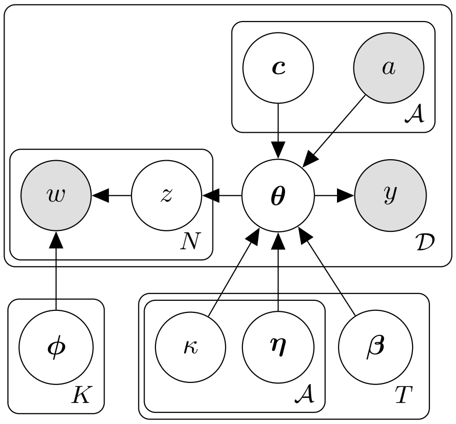


Figure 3.1: Plate diagram for author utility model. Hyperparameters and edges between consecutive time steps of $\boldsymbol{\beta}, \boldsymbol{\eta}$ and κ are omitted for clarity.

and the linear additive property of Gaussians gives us

$$\boldsymbol{\theta}_d \sim \mathcal{N}\left(\sum_{a \in \boldsymbol{a}_d} \kappa_a \boldsymbol{\beta} + c_{d,a}\boldsymbol{\eta}_a, \|\boldsymbol{c}_d\|_2^2 \mathbf{I}\right)$$

which is what we see in Eq. 3.6. From the above formulation, $\boldsymbol{\eta}_a$ can be seen as a deviation of author a's deviation from global trends $\boldsymbol{\beta}$.

In §3.1.2, we described a utility function for each author. The model we are estimating is similar to those estimated in discrete choice econometrics (McFadden 1974). We assumed that authors are utility maximizing (optimizing) and that their optimal topic distribution satisfies the first-order conditions (Eq. 3.4). However, we cannot observe the idiosyncratic component, $\boldsymbol{\epsilon}_{d,a}$, which is assumed to be Gaussian; as noted, this is known as a random utility model. Together, these assumptions give the structure of the distribution over topics in terms of (estimated) utility, which allows us to naturally incorporate the utility function into our probabilistic model in a familiar way. In Chapter 4, we will see alternative ways of formulating the utility function and incorporating them into a probabilistic model.

3.2 Learning and Inference

Exact inference in our model is intractable, so we resort to an approximate inference technique based on Monte Carlo EM (Wei and Tanner 1990). During the E-step, we perform Bayesian inference over latent parameters ($\boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\phi}$) using a Metropolis-Hastings within Gibbs algorithm (Tierney 1994), and in the M-step, we compute maximum *a posteriori* (MAP) estimates of $\boldsymbol{\beta}$ by directly optimizing the log-likelihood function. Since we are using conjugate priors for $\boldsymbol{\phi}$, we can integrate it out. We did not perform Bayesian posterior inference over $\boldsymbol{\beta}$ because the coupling of $\boldsymbol{\beta}$ would slow mixing of the MCMC chain. Also, an MAP point estimate of $\boldsymbol{\beta}$ is convenient if we want to inspect the parameters.

3.2.1 E-step

We sample each $\boldsymbol{\eta}_{a,j}$, for $j = 1 \ldots K$, and $\boldsymbol{\kappa}_a^{(\cdot)}$ blockwise across time steps using Metropolis-Hastings algorithm with a multivariate Gaussian proposal distribution. The equation for sampling $\boldsymbol{\eta}_{a,j}$ is

$$
\begin{aligned}
&p(\boldsymbol{\eta}_{a,j} \mid \boldsymbol{\eta}_{-(a,j)}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\kappa}, \boldsymbol{\beta}, \Lambda^{(\eta)}) \\
&\quad \propto \exp\left(-\frac{1}{2}\boldsymbol{\eta}_{a,j}\Lambda^{(\eta)}\boldsymbol{\eta}_{a,j}^{\top} - \sum_{\substack{t\in T \\ d\in D_t}} \frac{\left(\theta_{d,j} - \sum_{a'\in \boldsymbol{a}_d} \kappa_{a'}^{(t)}\beta_j^{(t)} + c_{d,a'}\eta_{a',j}^{(t)}\right)^2}{2\|\boldsymbol{c}_d\|_2^2} \right)
\end{aligned}
$$

and for sampling $\boldsymbol{\kappa}_a^{(\cdot)}$ is

$$
\begin{aligned}
&p(\boldsymbol{\kappa}_a \mid \boldsymbol{\kappa}_{-(a)}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\eta}, \boldsymbol{\beta}, \Lambda^{(\kappa)}) \\
&\quad \propto \exp\left(-\frac{1}{2}\log(\boldsymbol{\kappa}_a)\Lambda^{(\kappa)}\log(\boldsymbol{\kappa}_a^{\top}) - \sum_{\substack{t\in T \\ d\in D_t}} \frac{\|\boldsymbol{\theta}_d - \sum_{a'\in \boldsymbol{a}_d} \kappa_{a'}^{(t)}\boldsymbol{\beta}^{(t)} + c_{d,a'}\boldsymbol{\eta}_{a'}^{(t)}\|_2^2}{2\|\boldsymbol{c}_d\|_2^2} \right)
\end{aligned}
$$

$\Lambda^{(\eta)}$ and $\Lambda^{(\kappa)}$ are shorthands for the precision matrices $\Lambda(\lambda^{(\eta)}, \alpha^{(\eta)})$ and $\Lambda(\lambda^{(\kappa)}, \alpha^{(\kappa)})$ respectively. Likewise, $\boldsymbol{\theta}_d$ is sampled blockwise for each document with a multivariate

Gaussian distribution and likelihood

$$\begin{aligned}&p(\boldsymbol{\theta}_d \mid \boldsymbol{c}_d, \boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{\beta})\\ &\propto \exp\left(-\frac{(y_d - \boldsymbol{\beta}^{(t_d)\top}\boldsymbol{\theta}_d)^2}{2} - \frac{\|\boldsymbol{\theta}_d - \sum_{a\in\boldsymbol{a}_d} \kappa_a^{(t_d)}\boldsymbol{\beta}^{(t_d)} + c_{d,a}\boldsymbol{\eta}_a^{(t_d)}\|_2^2}{2\|\boldsymbol{c}_d\|_2^2}\right)\end{aligned}$$

For $\boldsymbol{c}_d$, we first sample each $\boldsymbol{c}_d$ from a multivariate Gaussian distribution, and applied a logistic transformation to map it onto the simplex. The likelihood for $\boldsymbol{c}_d$ is

$$\begin{aligned}&p(\boldsymbol{c}_d \mid \boldsymbol{\theta}_d, \boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{\beta})\\ &\propto \exp\left(-\frac{1}{2\sigma_c^2}\left\|\log\left(\frac{\boldsymbol{c}_d}{c_{d,|\boldsymbol{a}_d|}}\right)\right\|_2^2 - \frac{\|\boldsymbol{\theta}_d - \sum_{a\in\boldsymbol{a}_d} \kappa_a^{(t_d)}\boldsymbol{\beta}^{(t_d)} + c_{d,a}\boldsymbol{\eta}_a^{(t_d)}\|_2^2}{2\|\boldsymbol{c}_d\|_2^2}\right)\end{aligned}$$

where $c_{d,|\boldsymbol{a}_d|}$ denotes c for the last author on the document. The diagonal covariance matrix of the Gaussian proposal distributions are tuned independently to achieve a target acceptance rate of 15–45$\%$.

For $\boldsymbol{z}$, we integrate out $\boldsymbol{\phi}$ and sample each $z_{d,i}$ directly from

$$p(z_{d,i} = k \mid \boldsymbol{\theta}_d, \boldsymbol{\phi}_k) \propto \exp(\theta_{d,k})\frac{C_{k,w_{d,i}}^{-d,i} + \rho}{C_{k,\cdot}^{-d,i} + V\rho}$$

where $C_{k,w}^{-d,i}$ and $C_{k,\cdot}^{-d,i}$ are the number of times w is associated with topic k, and the number of tokens associated with topic k respectively.

We run the E-step Gibbs sampler to collect 3,500 samples, discarding the first 500 samples for burn-in and only saving samples at every third iteration.

3.2.2 M-step

We approximate the expectations of our latent variables using the samples collected during the E-step, and directly optimize $\boldsymbol{\beta}^{(t)}$ using L-BFGS (Liu and Nocedal 1989),[2] which

[2] We used libLBFGS, an open source C++ implementation (`https://github.com/chokkan/liblbfgs`).

requires a gradient. The gradient of the log-likelihood with respect to $\beta_j^{(t)}$ is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_j^{(t)}} &= -2\lambda^{(\beta)}\beta_j^{(t)} \\ &\quad - 2\lambda^{(\beta)}\alpha^{(\beta)}\mathbf{1}\{t > 1\}(\beta_j^{(t)} - \beta_j^{(t-1)}) \\ &\quad - 2\lambda^{(\beta)}\alpha^{(\beta)}\mathbf{1}\{t < T\}(\beta_j^{(t)} - \beta_j^{(t+1)}) \\ &\quad + 2\sum_{d\in\mathcal{D}_t} \theta_{d,j}(y_d - \beta_j^{(t)}\theta_{d,j}) \\ &\quad + 2\sum_{d\in\mathcal{D}_t} \kappa_d^{(t)}\left(\theta_{d,j} - \kappa_d^{(t)}\beta_j^{(t)} - \sum_{a\in\boldsymbol{a}_d}\frac{\eta_{a,j}^{(t)}}{|\boldsymbol{a}_d|}\right)\end{aligned}$$

where $\kappa_d^{(t)} = \frac{1}{|\boldsymbol{a}_d|}\sum_{a\in\boldsymbol{a}_d}\kappa_a^{(t)}$.

We ran L-BFGS until convergence[3] and slice sampled the hyperparameters $\lambda^{(\eta)}$, $\alpha^{(\eta)}$, $\lambda^{(\kappa)}$, $\alpha^{(\kappa)}$ (with vague priors) at the end of the M-step. We fix the symmetric Dirichlet hyperparameter $\rho = 1/V$, and tuned $\lambda^{(\beta)}, \alpha^{(\beta)}$ on a held-out developement dataset by grid search over $\{10^{-2}, 10^{-1}, 1, 10\}$. During initialization, we randomly set the topic assignments, while the other latent parameters are set to 0. We ran the model for 10 EM iterations.

3.2.3 Inference

During inference, we fix the model parameters and only sample $(\boldsymbol{\theta}, \boldsymbol{z})$ for each document. As in the E-step, we discard the first 500 samples, and save samples at every third iteration, until we have 500 posterior samples. In our experiments, we found the posterior samples to be reasonably stable after the initial burn in.

[3] Relative tolerance of 10^{-4}.

3.3 Experiments

In this section, we conduct experiments to evaluate the performance of our author utility model.

3.3.1 Data

The ACL Anthology Network Corpus contains 21,212 papers published in the field of computational linguistics between 1965 and 2013 and written by 17,792 authors. Additionally, the corpus provides metadata such as authors, venue and in-community citation networks. For our experiments, we focused on conference papers published between 1980 and 2010.[4] We tokenized the texts, tagged the tokens using the Stanford POS tagger (Toutanova et al. 2003), and extracted n-grams with tags that follow the simple (but effective) pattern of `(Adj|Noun)*` `Noun` (Justeson and Katz 1995), representing the dth document as a *bag of phrases* ($\boldsymbol{w}_d$). Note that phrases can also be unigrams. We pruned phrases that appear in $< 1\%$ or $> 95\%$ of the documents, obtaining a vocabulary of $V =$ 6,868 types. Authors with less than 3 papers were pruned. The pruned corpus contains 5,498 documents and 2,643,946 phrase tokens written by 5,575 authors. We let responses

$$y_d = \log(1 + \text{\# of incoming citations in 3 years})$$

For our experiments, we used 3 different random splits of our data (70% train, 20% test, and 10% development) and averaged quantities of interest. Furthermore, we remove an author from a paper in the development or test set if we have not seen him before in the training data.

[4] The conferences we included are: ACL, CoNLL, EACL, EMNLP, HLT, and NAACL. We ignored journal papers, as well as workshop papers, since they are characteristically different from conference papers.

3.3.2 Examples of Authors and Topics

Table 3.2 illustrates ten manually selected topics (out of 64) learned by the author utility model. Each topic is labeled with the top 10 words most likely to be generated conditioned on the topic ($\boldsymbol{\phi}_k$). For each topic, we compute an author's topic preference score:

$$\text{TPS}(a, k) = \sum_{d \in D_a} \eta_{a,k}^{(t_d)} [\text{Softmax}(\boldsymbol{\theta}_d)]_k \times y_d$$

where $\text{Softmax}(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x})}{\sum_i \exp(x_i)}$. The TPS scales the author's $\boldsymbol{\eta}$ preferences by the relative number of citations that the author received for the topic. This way, we can account for different $\boldsymbol{\eta}$s over time, and reduce variance due to authors who publish less frequently.[5] For each topic, the five authors with the highest TPS are displayed in the rightmost column of Table 3.2. These topics were among the roughly one third (out of 64) that seemed to coherently map to research topics within NLP. Some others corresponded to parts of a paper (e.g., explaining notation and formulae, experiments) or to stylistic groups (e.g., "rational words" including *rather*, *fact*, *clearly*, *argue*, *clear*, *perhaps*). Others were not interpretable to us.

3.3.3 Predicting Responses

We compare against two baselines for predicting in-community citations. Yogatama et al. (2011) is a strong baseline for predicting responses; they used n-gram features and metadata features in a generalized linear model with the time series prior discussed in §3.1.5.[6] We also compare against a version of our model without the author utility component. This equates to replacing Yogatama et al. (2011)'s features with LDA topic mixtures, and performing joint learning of the topics and citations; we therefore call it "TimeLDA." Without the time series component, TimeLDA would instantiate supervised LDA (Mcauliffe and

[5] The TPS is only a measure of an author's propensity to write papers in a specific topic area and is not meant to be a measure of an author's reputation in a particular research sub-field.

[6] For the ACL dataset, Yogatama et al. (2011)'s model predicts whether a paper will receive at least 1 citation within three years, while here, we train it to predict $\log(1 + \#\text{citations})$ instead.

Blei 2008). Figure 3.2 shows the mean absolute error (MAE) for the three models. With sufficiently many topics ($K \geq 16$), topic representations achieve lower error than surface features. Removing the author utility component from our model leads to better predictive performance. This is unsurprising, since our model forces $\boldsymbol{\beta}$ to explain both the responses (what is evaluated here) and the divergence between author preferences $\boldsymbol{\eta}_a$ and what is actually written. The utility model is nonetheless competitive with the Yogatama et al. (2011) baseline, while being able to model the expertise of authors and their trade-offs.

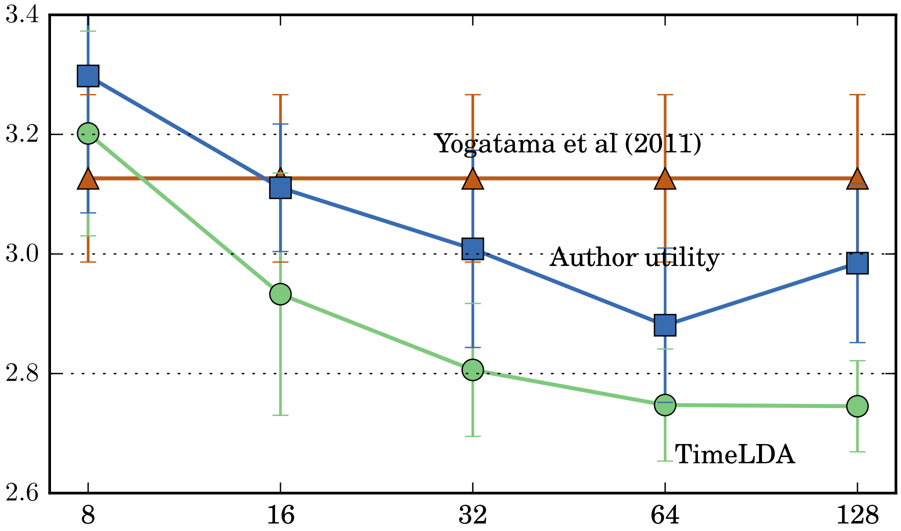


Figure 3.2: Mean absolute error (in citation counts) for predicted citation counts (y-axis) against the number of topics K (x-axis). Errors are in actual citation counts, while the models are trained with log counts. TimeLDA significantly outperforms Yogatama et al. (2011) for $K \geq 64$ (paired t-test, $p < 0.01$), while the differences between Yogatama et al. (2011) and author utility are not significant. The MAE is calculated over 3 random splits of the data with 809, 812, and 811 documents in the test set respectively.

3.3.4 Predicting Words

"Given a set of authors in a given year, what are they likely to write?" — we use perplexity as a proxy to measure the content predictive ability of our model. Perplexity on a test set is commonly used to quantify the generalization ability of probabilistic models and make comparisons among models over the same observation space. For a document $\boldsymbol{w}_d$ written by authors $\boldsymbol{a}_d$, perplexity is defined as

$$\text{perplexity}(\boldsymbol{w}_d \mid \boldsymbol{a}_d) = \exp\left(-\frac{\log p(\boldsymbol{w}_d \mid \boldsymbol{a}_d)}{N_d}\right)$$

and a lower perplexity indicates better generalization performance. Using S samples from the inference step, we can compute

$$p(\boldsymbol{w}_d \mid \boldsymbol{a}_d) = \frac{1}{S}\sum_{s=1}^{S}\prod_{i=1}^{N_d}\frac{1}{|\boldsymbol{a}_d|}\sum_{a\in\boldsymbol{a}_d,k}\theta^s_{d,k}\phi^s_{k,w_{di}}$$

where $\boldsymbol{\theta}^s$ is the sth sample of $\boldsymbol{\theta}$, and $\boldsymbol{\phi}^s$ is the topic-word distribution estimated from the sth sample of $\boldsymbol{z}$.

We compared our models against several baselines: TimeLDA, as described in §3.3.3, Author-Topic model of Rosen-Zvi et al. (2004), and a version of our author utility model that ignores temporal information ("–Time"), i.e., setting $T = 1$ and collapsing all timesteps. The AT model is similar to setting $\boldsymbol{\kappa}_a = 0$ for all authors, $\boldsymbol{c}_d = \frac{1}{|\boldsymbol{a}_d|}$, and using a Dirichlet prior instead of logistic normal on $\boldsymbol{\eta}_a$. For the AT model, we also included a version that incorporates temporal information; we instantiate different author topic distributions for each year and used the time series regularizer to tie the parameters together. This is very similar to our Author utility model, albeit with different priors and without the utility function. Figure 3.3 present the perplexity of these models at different values of K. We find that perplexity improves with the addition of the utility model as well as the temporal dynamics.

3.3.5 Exploration: Tradeoffs and Seniority

Recall that κ_a encodes author a's tradeoff between increasing citations (high κ_a) and writing papers on topics a prefers (low κ_a). We do not claim that individual κ_a values consistently represent authors' tradeoffs between citations and writing about preferred topics.

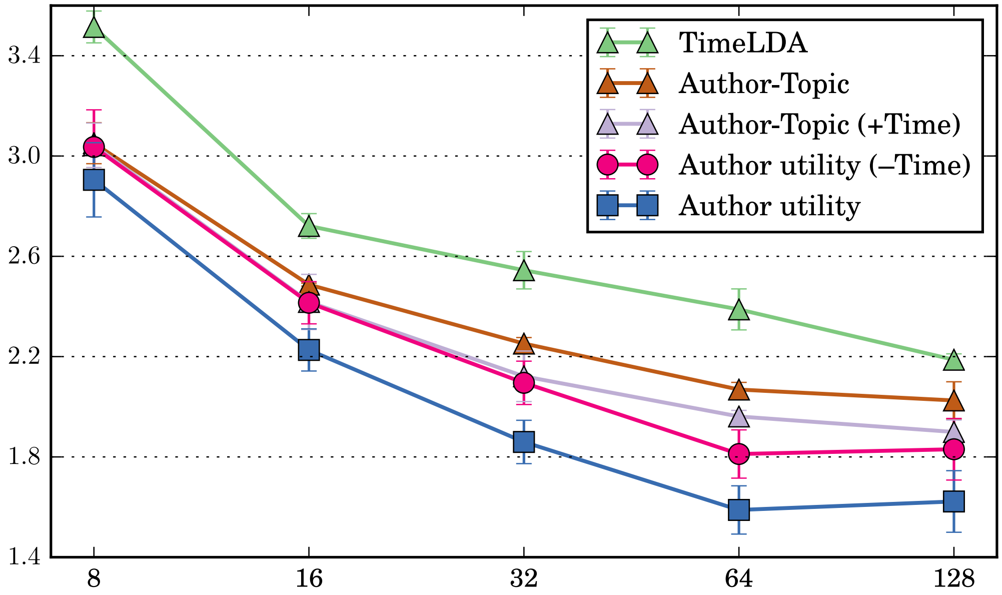


Figure 3.3: Held-out perplexity ($\times 10^3$, y-axis) with varying number of topics K (x-axis). The differences are significant between all models except "Author-Topic (+Time)" at $K \geq 64$ (paired t-test, $p < 0.01$). There are 523,381, 529,397, 533,792 phrase tokens in the random test sets.

We have noted a number of potentially confounding factors that affect authors' choices, for which our data do not allow us to control.

However, in aggregate, κ_a values can be explored in relation to other quantities. Given our model's posterior, one question we can ask is: do an author's tradeoffs tend to change over the course of her career? In Figure 3.4, we plot the median of κ (and 95% credible intervals) for authors at different "ages." Here, "age" is defined as the number of years since an author's first publication in this dataset.[7]

A general trend over the long term is observed: researchers appear to move from higher to lower κ_a. Statistically, there is significant dependence between κ of an author and her age; the Spearman's rank correlation coefficient is $\rho = -0.870$ with p-value $< 10^{-5}$. This finding is consistent with the idea that greater seniority brings increased and more

[7] This means that larger ages correspond to seniority, but smaller ages are a blend of junior researchers and researchers of any seniority new to this publication community.

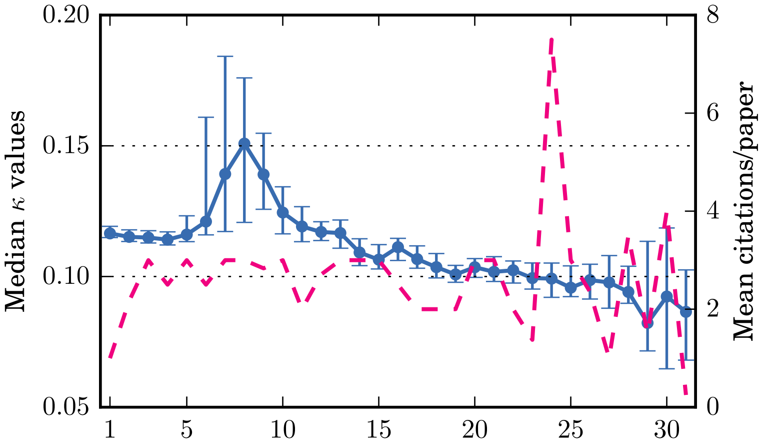


Figure 3.4: Plot of authors' median κ (blue, solid) and mean citation counts (magenta, dashed) against their academic age in this dataset (see text for explanation).

stable resources and greater freedom to pursue idiosyncratic interests with less concern about extrinsic payoff. It is also consistent with decreased flexibility or openness to shifting topics over time.

To illustrate the importance of our model in making these observations, we also plot the mean number of citations per paper published (across all authors) against their academic age (magenta lines). There is no clear statistical trend between the two variables ($\rho = -0.017$). This suggests that through κ, our model is able to pick up evidence of author's optimizing behaviors, which is not possible using simple citation counts.

There is a noticeable effect during years 5–10, in which κ tends to rise by around 40% and then fall back. (Note that the model maintains considerable uncertainty — wider intervals — about this effect.) Recall that, for a researcher trained within the field and whose primary publication venue is in the ACL community, our measure of age corresponds roughly to academic age. Years 5–10 would correspond to the later part of a Ph.D. program and early postgraduate life, when many researchers begin faculty careers. Insofar as it reflects a true effect, this rise and fall suggests a stage during which a researcher focuses more on writing papers that will attract citations. However, more in-depth study based on data that is not merely observational is required to quantify this effect and, if it persists under scrutiny, determine its cause.

The effect in year 24 of mean citations per paper (magenta line) can be attributed to well cited papers co-authored by senior researchers in the field who published very few papers in their 24th year.[8] Since there are relatively few authors[9] in the dataset at that academic age, there is more variance in mean citation counts.

3.4 Related Work

Previous work on modeling author interests mostly focused on characterizing authors by their style (Holmes and Forsyth 1995, *inter alia*),[10] through latent topic mixtures of documents they have co-authored (Rosen-Zvi et al. 2004) and their collaboration networks (Johri et al. 2011). Like our paper, the latter two are based on topic models, which have been popular for modeling the content of scientific articles. For instance, Gerrish and Blei (2010) measured scholarly impact using dynamic topic models, while Hall et al. (2008) analyzed the output of topic models to study the "history of ideas."

Predicting responses to scientific articles was explored in two shared tasks at KDD Cup 2003 (Brank and Leskovec 2003; McGovern et al. 2003) and by Yogatama et al. (2011), which served as a baseline for our experiments and whose time-series prior we used in our model. In Yogatama et al. (2011), the authors formulated the problem of predicting responses in the context of *forecasting*. In this view, the $\boldsymbol{\beta}$s can be seen as capturing the annual topical trends of the community; Yogatama et al. (2011)'s experiments showed that there is substantial power in using linear models with text features to forecast response in future articles. However, as their model and ours are first-order models (i.e., the time series regularizer only cares about $\boldsymbol{\beta}$ parameters at adjacent time steps), we will likely need higher

[8] The top 3 authors in terms of mean citations per paper in their 24th year are: Fernando Pereira in 2006 for McDonald and Pereira (2006, 52 citations), John Carroll in 2006 with 60 citations for both Briscoe and Carroll (2006) and Briscoe et al. (2006), and Ronald M. Kaplan in 2004 for Kaplan et al. (2004, 26 citations).

[9] There are 26 authors with academic age 24 and older.

[10] A closely related problem is that of authorship attribution. There has been extensive research on authorship attribution focusing mainly on learning "stylometric" features of authors; see Stamatatos (2009) for a detailed review.

order models for longer term trends.

Furthermore, there has been considerable research using topic models to predict (or recommend) citations (instead of aggregate counts), such as modeling link probabilities within the LDA framework (Cohn and Hofmann 2001; Erosheva et al. 2004; Kataria et al. 2010; Nallapati and Cohen 2008; Zhu et al. 2013) and augmenting topics with discriminative author features (Liu et al. 2009; Tanner and Charniak 2015).

We modeled both interests of authors and responses to their articles jointly, by assuming authors' text production is an expected utility-maximizing decision. This approach is similar to our earlier work (Sim et al. 2015), where authors are rational agents writing texts to maximize the chance of a favorable decision by a judicial court. In that study, we did not consider the unique preferences of each decision making agent, nor the extrinsic-intrinsic reward tradeoffs that these agents face when authoring a document.

Our utility model can also be viewed as a form of natural language generator, where we take into account the context of an author (i.e., his preferences, the tradeoff coefficient, and what is popular) to generate his document. This is related to natural language pragmatics, where text is influenced by context.[11] Hovy (1990) approached the problem of generating text under pragmatic circumstances from a planning and goal-orientation perspective, while Vogel et al. (2013) used multi-agent decision-theoretic models to show cooperative pragmatic behavior. Vogel et al. (2013)'s models suggest an interesting extension of ours for future work: modeling cooperation among co-authors and, perhaps, in the larger scientific discourse.

3.5 Conclusions

In this chapter, we presented a model of scientific authorship in which authors trade off between seeking citation from others and staying true to their individual preferences among

[11] The $\boldsymbol{\beta}$ vectors can be seen as a naïve representation of world knowledge that motivates an author to select content that reflects his behavioral preferences and intentions.

research topics. We find that topic modeling improves over state-of-the-art text regression models for predicting citation counts, and that the author utility model generalizes better than simpler models when predicting what a particular group of authors will write. Further, our models enable analysis of underlying strategic behaviors while mantaining comparable performance in predictive accuracy. When we analyzed our the results of our model on the ACL community, we find patterns that suggests interesting strategic behavior across a researcher's career.

More generally, the author utility model is a demonstration of our methodology for modeling the strategic choices of an author — by designing utility functions which we then incorporate into a Bayesian generative model. In the subsequent chapter, we use similar ideas to explore yet another example of strategic behavior, this time in the Supreme Court of the United States, where we seek to understand the legal agendas of different parties in the judiciary system.

Topic	**Top words**	**Authors**
“MT”	machine, translation, equation, decode, phrase, och, bleu, ney, bleu score, target language	Philipp Koehn, Chris Dyer, Qun Liu, Hermann Ney, David Chiang
“Empirical methods”	model, parameter, learn, iteration, maximize, prior, initialize, distribution, weight, crf	Noah Smith, Dan Klein, Percy Liang, John DeNero, Andrew McCallum
“Parsing”	parse, sentence, parser, accuracy, collins, dependency, tree, parse tree, head, charniak	Michael Collins, Joakim Nivre, Jens Nilsson, Dan Klein, Ryan McDonald
“Dialogue systems”	speak, speech, utterance, user, speaker, dialogue system, turn, act, recognition, transcription	Diane Litman, Marilyn Walker, Julia Hirschberg, Oliver Lemon, Amanda Stent
“NER”	name, entity, identify, person, location, list, organization, system, entity recognition, mention	Jenny Rose Finkel, Satoshi Sekine, Rion Snow, Christopher Manning, Abraham Ittycheriah
“Semantics”	argument, verb, predicate, syntactic, relation, semantic role, annotated, frame, assign	Martha Palmer, Alessandro Moschitti, Daniel Jurafsky, Sanda Harabagiu, Mirella Lapata
“Lexical semantics”	wordnet, noun, sense, concept, context, sens, relation, meaning, pair, disambiguate	Rion Snow, Rob Koeling, Eneko Agirre, Ido Dagan, Patrick Pantel
“Tagging & chunking”	method, sentence, propose, japanese, noun phrase, extract, table, analyze, precision, technology	Yuji Matsumoto, Hitoshi Isahara, Junichi Tsujii, Sadao Kurohashi, Kentaro Torisawa
“Coreference”	mention, instance, create, approach, report, due, text, pair, exist, system	Vincent Ng, Aria Haghighi, Xiaofeng Yang, Claire Cardie, Pascal Denis
“Sentiment classification”	classify, label, accuracy, positive, classification, annotated, annotator, classifier, review, negative	Janyce Wiebe, Soo Min Kim, Eduard Hovy, Carmen Banea, Ryan McDonald

Table 3.2: Top words from selected topics and authors with preferences in those topics. We manually labeled each of these topics.

Chapter 4

The U.S. Supreme Court and its Strategic Friends

> Language is the central tool of our trade. You know, when we're looking at a statute, trying to figure out what it means, we're relying on the language. When we're construing the Constitution, we're looking at words. Those are the building blocks of the law. And so if we're not fastidious, as you put it, with language, it dilutes the effectiveness and clarity of the law.
>
> *Chief Justice John G. Roberts* (Kimble 2010)

The work described in this chapter is a cumulation of research efforts published in Sim et al. (2015) and another paper that is currently under review.

The Supreme Court of the United States (SCOTUS) is the highest court in the American judicial system; its decisions have far-reaching effects. While the ideological tendencies of SCOTUS' nine justices are widely discussed by press and public, there is a formal mechanism by which organized interest groups can lobby the court on a given case. These groups are known as **amici curiae**[1] and the textual artifacts they author — known as amicus briefs

[1] *Amicus curiae* is Latin for "friends of the court." Hereafter, we use *amicus* in singular and *amici* in plural to refer to these interested third parties. It is common for several amici to co-author a single brief, which we

— reveal explicit attempts to sway justices one way or the other. In recent years, amicus briefs are increasingly being employed as a lobbying tool to influence the Court's decision-making process (Franze and Anderson 2015; Kearney and Merrill 2000). Taken alongside voting records and other textual artifacts that characterize a case, amicus briefs provide a fascinating setting for empirical study of influence through language. As such, we take the perspective of an amicus, proposing a probabilistic model of the various parties to a case that accounts for the amicus' goals.

Our model of SCOTUS is considerably more comprehensive than past work in political science, which has focus primarily on **ideal point** models that use votes as evidence. Text has been incorporated more recently as a way of making such models more interpretable, but without changing the fundamental assumptions (Lauderdale and Clark 2014). Although the influence of amici has been studied extensively by legal scholars (Collins 2008), we are the first to incorporate them into ideal points analysis. In §4.1, we propose two different approaches for incorporating amicus briefs into the ideal points model. Drawing on decision theory, we then posit amici as rational agents seeking to maximize their expected utility by framing an amicus brief's arguments to influence justices toward a favorable outcome (§4.3). We derive appropriate inference and parameter estimation procedures (§4.5).

Our experiments (§4.6) show that the new approach offers substantial gains in vote prediction accuracy. More importantly, we show how the model can be used to answer questions such as:

- How susceptible are different justices to influence by amici? (§4.6.3)
- What would have happened if some or all amicus briefs were not filed? (§4.7.2)
- How might an amicus have changed her brief to obtain a better outcome? (§4.7.2)
- How effective were amici on each side of a case? (§4.7.3)

Since we characterize the amicus brief as a (probabilistic) function of the case parameters, our model-based approach could also be used to ask how the amici would have altered

account for in our model.

their briefs given different merits facts or a different panel of justices. Although we focus on SCOTUS, our model is applicable to any setting where textual evidence for competing goals is available alongside behavioral response.

SCOTUS Terminology. SCOTUS reviews the decisions of lower courts and (less commonly) resolves disputes between states.[2] At a given time it typically consists of 9 justices.[3] The primary means to petition SCOTUS to review a case is to ask it to grant a writ of certiorari; this happens in about 1.5% of 7,000 petitions a year. SCOTUS is under no obligation to hear a case. In a typical case that is heard by the Court, the **petitioner** writes a brief putting forward her legal argument; the **respondent** (the other party) then files a brief. These, together with a round of responses to each other's initial briefs, are collectively known as **merits briefs**. **Amicus briefs** — further arguments and recommendations on either side — may be filed by groups with an interest (but not a direct stake) in the outcome, with the Court's permission. After oral arguments, which are not necessarily allotted for every case, conclude, the justices vote and author one or more opinions. Typically, a justice from the majority is assigned to author the rationale for the Court's decision; others may write separate concurring or dissenting opinions.

4.1 Ideal Point Voting Models

Ideal point (IP) models are a mainstay in quantitative political science, often applied to voting records to place voters (lawmakers, justices, etc.) in a continuous space. A justice's "ideal point" is a latent variable positioning her in this space. To keep the discussion self contained, we begin with classical models of votes alone and build up toward our novel contributions.

[2] Details about the procedures and rules of the SCOTUS can be found at `http://www.uscourts.gov`.

[3] Unless a governmental gridlock prevents the appointment of a replacement justice in the event of death/retirement of a current justice.

4.1.1 Unidimensional Ideal Points

The simplest model for judicial votes is a unidimensional IP model (Martin and Quinn 2002), which posits an IP $\psi_j \in \mathbb{R}$ for each justice j.[4] Often the ψ_j values are interpreted as positions along a liberal-conservative ideological spectrum. Each case i is represented by **popularity** (a_i) and **polarity** (b_i) parameters.[5] A probabilistic view of the unidimensional IP model is that justice j votes in favor of case i's petitioner with probability

$$p(v_{i,j} = \text{petitioner} \mid \psi_j, a_i, b_i) = \sigma\left(a_i + \psi_j b_i\right)$$

where $\sigma(x) = \frac{\exp(x)}{1+\exp(x)}$ is the logistic function. When the popularity parameter a_i is high enough, every justice is more likely to favor the petitioner. The polarity b_i captures the importance of the justice's ideology ψ_j: more polarizing cases (i.e., $|b_i| \gg 0$) push justice j more strongly to the side of the petitioner (if b_i has the same sign as ψ_j) or the respondent (otherwise). Figure 4.1a illustrates the graphical model view of the Martin and Quinn (2002) unidimensional IP model.

The unidimensional IP model has been extended to multiple dimensions using the same formulation (Jackman 2001; Martin and Quinn 2001). While multidimensional IP models recover dimensions that maximize statistical fit, they conflate many substantive dimensions of opinion and policy, making it difficult to interpret additional dimensions.[6] Indeed, such embeddings are ignorant of the issues at stake, or any content of the case, and they cannot generalize to new cases.

[4] Martin and Quinn (2002) describe a dynamic unidimensional IP model where justice IPs vary over time. In this work, we assume each justice's IP is fixed over time, for simplicity.

[5] This model is also known as a two parameter logistic model in item response theory literature (Fox 2010), where a_i is "difficulty" and b_i is "discrimination."

[6] A seminal finding of Poole and Rosenthal (1985) is that two dimensions, corresponding to left-right ideology and geographical latitude, explain most of the variance in U.S. Congressional votes.

4.1.2 Issues and Ideal Points ("Model L")

Lauderdale and Clark (2014) incorporate text as evidence and infer dimensions of IP that are grounded in "topical" space. They build on latent Dirichlet allocation (Blei et al. 2003b), a popular model of latent topics or themes in text corpora. In their model, each case i is embedded as $\boldsymbol{\theta}_i$ in a D-dimensional simplex; the dth dimension $\theta_{i,d}$ corresponds to the proportion of case i that is about issue (or, in LDA terminology, topic) d. The probability of justice j's vote is given by

$$p(v_{i,j} = \text{petitioner} \mid \boldsymbol{\psi}_j, \boldsymbol{\theta}_i, a_i, b_i) = \sigma\left(a_i + \boldsymbol{\psi}_j^\top (b_i \boldsymbol{\theta}_i)\right)$$

where $\psi_{j,d}$ is an *issue-specific* position for justice j. Therefore, the relative degree that each dimension predicts the vote outcome is determined by the text's mixture proportions, resulting in the issue-specific IP $\boldsymbol{\psi}_j^\top \boldsymbol{\theta}_i$. In their work, they inferred the mixture proportions from justices' opinions, although one can similarly use merits briefs, appeals court opinions, or any other texts that serve as evidence for inferring the issues of a case.

Lauderdale and Clark (2014) found that incorporating textual data in this manner[7] addresses the labeling problem for multidimensional models, and is especially useful for small voting bodies (e.g., SCOTUS), where estimating multidimensional models is difficult due to few observations and variation of preferences across issues.

The plate diagram for model L can be found in Figure 4.1b.

4.1.3 Amici and Ideal Points ("Model A")

The merits briefs describe the issues and facts of the case. It is our hypothesis that amicus briefs serve to "frame" the facts and, potentially, influence the case outcome. Collins (2008) argued that these organized interest groups play a significant role in shaping jus-

[7] Of course, LDA is not the only way to "embed" a case in a simplex. One can take advantage of expert categorization of case issues. For example, Gerrish and Blei (2012b) used bill labels as supervision to infer the proportions of issues.

tices' choices and Corley et al. (2013) have observed that justices systematically incorporate language from amicus briefs to enhance their ability to make effective law and policy. Public interest groups, such as the American Civil Liberties Union (ACLU) and Citizens United, frequently advocate their positions on any case that impinges on their goals. These briefs can provide valuable assistance to the Court in its deliberation; for example, they can present an argument not found in the merits.[8]

When filing amicus briefs, amici are required to identify the side they are supporting (or if neither). However, it may not always be trivial to tell which side the amici are on as these intentions are found in the brief text and are not expressed consistently. We solve this by training a classifier on hand-labeled data (§4.5.4).

We propose that amici represent an attempt to shift the position of the case by emphasizing some issues more strongly or framing the case distinctly from the perspectives given in the merits briefs. The effective position of a case, previously $b_i\boldsymbol{\theta}_i$, is in our model $b_i\boldsymbol{\theta}_i + c_i^{\text{p}}\boldsymbol{\Delta}_i^{\text{p}} + c_i^{\text{r}}\boldsymbol{\Delta}_i^{\text{r}}$, where c_i^{p} and c_i^{r} are the **amicus polarities** for briefs on the side of the petitioner and respondent. $\boldsymbol{\Delta}_i^{\text{p}}$ and $\boldsymbol{\Delta}_i^{\text{r}}$ are the mean issue proportions of the amicus briefs on the side of the petitioner and respondent, respectively. Our amici-augmented IP model is:

$$\begin{aligned} &p(v_{i,j} = \text{petitioner} \mid \psi_j, \boldsymbol{\theta}_i, \boldsymbol{\Delta}_i, a_i, b_i, c_i^{\text{p}}, c_i^{\text{r}}) \\ &= \sigma\left(a_i + \boldsymbol{\psi}_j^\top \left(b_i\boldsymbol{\theta}_i + c_i^{\text{p}}\boldsymbol{\Delta}_i^{\text{p}} + c_i^{\text{r}}\boldsymbol{\Delta}_i^{\text{r}}\right)\right) \end{aligned} \tag{4.1}$$

In this model, the vote-specific IP is influenced by two forms of text: legal arguments put forth by the parties involved (merits briefs, embedded in $\boldsymbol{\theta}_i$), and by the amici curiae (amicus briefs, embedded in $\boldsymbol{\Delta}_i^{\{\text{p,r}\}}$), both of which are rescaled independently by the case discrimination parameters to generate the vote probability. When either $|c_i^{\text{p}}|$ or $|c_i^{\text{r}}|$ is large (relative to a_i and b_i), the vote is determined by the contents of the amicus briefs.

By letting $\boldsymbol{\Delta}_i^{s}$ be the average mixture proportions inferred from text of briefs sup-

[8] On occasion, SCOTUS may adopt a position not advanced by either side, but instead urged solely by an amicus brief. Some notable cases are: *Mapp v. Ohio*, 367 U.S. 643, 646 (1961) and more recently, *Turner v. Rogers*, 131 S. Ct. 2507 (2011).

porting side s, we implicitly assume that briefs supporting the same side share a single parameter, and individual briefs on one side influence the vote-specific IP equally. While Lynch (2004) and others have argued that some amici are more effective (i.e., influence on justices' votes varies across amicus authors), this model captures the collective effect of amicus briefs and is simple. In the sequel, we introduce variations that take into account the effectiveness of individual amicus.

Generative story for model A. With appropriate priors on the latent variables, the generative story of model A is:

1. For each topic $t \in \{1, \ldots, T\}$, draw topic-word distributions $\boldsymbol{\phi}_t \sim \text{Dir}(\beta)$.
2. Draw justice IP off-diagonal covariance, $\rho \sim \text{Uniform}(0, 1)$.
3. For each justice $j \in \mathcal{J}$, draw justice IP $\boldsymbol{\psi}_j \sim \mathcal{N}(\mathbf{0}, \sigma_J^2 \mathbf{I} + \rho \mathbf{1})$.[9]
4. For each case $i \in \mathcal{C}$:
   (a) Draw case parameters $a_i, b_i, c_i^{\text{p}}, c_i^{\text{r}} \sim \mathcal{N}(0, \sigma_C^2)$.
   (b) Draw topic proportions for merits briefs $\boldsymbol{\theta}_i \sim \text{Dir}(\alpha)$.
   (c) For each word $w_{i,n}^{(m)}$ in the merits briefs, draw topic indicators $z_{i,n}^{(m)} \sim \text{Categorical}(\boldsymbol{\theta}_i)$ and $w_{i,n}^{(m)} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{i,n}^{(m)}})$.
   (d) For each amicus brief indexed by k:
      i. Draw topic proportions $\boldsymbol{\Delta}_{i,k}$ according to a distribution discussed in §4.3 .
      ii. For each word $w_{i,n}^{(a)}$ in the brief, draw topic indicators $z_{i,k,n}^{(a)} \sim \text{Categorical}(\boldsymbol{\Delta}_{i,k})$ and $w_{i,k,n}^{(a)} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{i,k,n}^{(a)}})$.
   (e) For each participating justice $j \in \mathcal{J}_i$, draw vote $v_{i,j}$ according to Eq. 4.1.

[9] The positive off-diagonal elements of the covariance matrix for justice IPs (ψ_j) orient the issue-specific dimensions in the same direction (i.e., with conservatives at the same end) and provide shrinkage of IP in each dimension to their common mean across dimensions (Lauderdale and Clark 2014).

We discuss the setting of hyperparameters (i.e., α, β, σs) in §4.5.3. The full likelihood of the model is:

$$
\begin{aligned}
&\mathcal{L}^{(\mathrm{A})}(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\psi}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{\Delta}, \boldsymbol{z}, \boldsymbol{\phi}, \rho \mid \alpha, \beta, \sigma_C, \sigma_J) \\
&\propto \prod_{t=1}^{T} p(\boldsymbol{\phi}_t \mid \beta) \\
&\times p(\rho) \prod_{j \in \mathcal{J}} p(\boldsymbol{\psi}_j \mid \sigma_J, \rho) \\
&\times \prod_{i \in \mathcal{C}} p(a_i, b_i, c_i^{\mathrm{p}}, c_i^{\mathrm{r}} \mid \sigma_C) \\
&\times \prod_{i \in \mathcal{C}} p(\boldsymbol{\theta}_i \mid \alpha) \prod_{n} p(z_{i,n}^{(m)} \mid \boldsymbol{\theta}_i) p(w_{i,n}^{(m)} \mid \boldsymbol{\phi}_{z_{i,n}^{(m)}}) \\
&\times \prod_{i \in \mathcal{C}} \prod_{k \in \mathcal{A}_i} p(\boldsymbol{\Delta}_{i,k}) \prod_{n} p(z_{i,k,n}^{(a)} \mid \boldsymbol{\Delta}_{i,k}) p(w_{i,k,n}^{(a)} \mid \boldsymbol{\phi}_{z_{i,k,n}^{(a)}}) \\
&\times \prod_{i \in \mathcal{C}} \prod_{j \in \mathcal{J}} p(v_{i,j} \mid \psi_j, \boldsymbol{\theta}_i, \boldsymbol{\Delta}_i, a_i, b_i, c_i^{\mathrm{p}}, c_i^{\mathrm{r}})
\end{aligned}
$$

where $\mathcal{A}_i$ is a notational shorthand for the set of amicus briefs filed in case i. Also, the distribution over $\boldsymbol{\Delta}$s will be discussed in §4.3.

The plate diagram for model A is presented in Figure 4.1c.

4.1.4 Persuasive Amici Ideal Points ("Model P")

Here, we propose a new model which considers amici as individual actors. Starting from Eq. 4.1, we consider two additional variables: each amicus e's **persuasiveness** ($\pi_e > 0$) and each justice j's **influenceability** ($\chi_j > 0$).

$$
\begin{aligned}
&p(v_{i,j} = \text{petitioner} \mid \boldsymbol{\psi}_j, \chi_j, \boldsymbol{\theta}_i, \boldsymbol{\Delta}_i, a_i, b_i, \boldsymbol{\pi}) \\
&= \sigma\left(a_i + b_i \boldsymbol{\psi}_j^{\top}\left(\boldsymbol{\theta}_i + \frac{\chi_j}{|\mathcal{A}_i|} \sum_{k \in \mathcal{A}_i} \bar{\pi}_{i,k} \boldsymbol{\Delta}_{i,k}\right)\right)
\end{aligned}
\tag{4.2}
$$

where $\bar{\pi}_{i,k} = \frac{\sum_{e \in \mathcal{E}_{i,k}} \pi_e}{|\mathcal{E}_{i,k}|}$ is the average of their π-values, with $\mathcal{E}_{i,k}$ denoting the set of entities who co-authored the kth amicus brief for case i.

Intuitively, a larger value of χ_j will shift the case IP more towards the contents of the amicus briefs, thus making the justice seem more "influenced" by amicus. Likewise, briefs co-authored by groups of amici who are more effective (i.e., larger $\bar{\pi}_{i,k}$), will "frame" the case towards their biases. Therefore, χ_j provides an additional degree of freedom for the model to capture the "scale" of influence of amicus brief on justice j, while ψ_j can be seen as j's IP vector projection of the resulting topics.

Unlike Eq. 4.1, here we eschew the amicus polarity parameters (c_i) and instead rely on the influenceability and persuasiveness parameters. Also, we note that model A in §4.1.3 is a special case where $\chi_j = 1$ and each case has polarity parameters on each side; no information is shared across briefs written by the same amicus-entity for different cases.

Generative story for model P. With appropriate priors on the latent variables, the generative story of model P is:

1. For each topic $t \in \{1, \ldots, T\}$, draw topic-word distributions $\boldsymbol{\phi}_t \sim \text{Dir}(\beta)$.
2. Draw justice IP off-diagonal covariance, $\rho \sim \text{Uniform}(0, 1)$.
3. For each justice $j \in \mathcal{J}$, draw justice IP $\boldsymbol{\psi}_j \sim \mathcal{N}(\mathbf{0}, \sigma_J^2\mathbf{I} + \rho\mathbf{1})$[10] and influenceability $\chi_j \sim \log\mathcal{N}(\mathbf{0}, \sigma_I^2\mathbf{I})$.
4. For each amicus-entity $e \in \mathcal{E}$, draw its persuasiveness $\pi_e \sim \log\mathcal{N}(\mathbf{0}, \sigma_P^2\mathbf{I})$.
5. For each case $i \in \mathcal{C}$:
   (a) Draw case parameters $a_i, b_i \sim \mathcal{N}(0, \sigma_C^2)$.
   (b) Draw topic proportions for merits $\boldsymbol{\theta}_i \sim \text{Dir}(\alpha)$.
   (c) For each word $w_{i,n}^{(m)}$ in the merits briefs, draw topic indicators $z_{i,n}^{(m)} \sim \text{Categorical}(\boldsymbol{\theta}_i)$

[10] See footnote 9 in §4.1.3 regarding the shrinkage parameter ρ.

and $w_{i,n}^{(m)} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{i,n}^{(m)}})$.

(d) For each amicus brief indexed by k:

i. Draw topic proportions $\boldsymbol{\Delta}_{i,k}$ according to a distribution discussed in §4.3 .

ii. For each word $w_{i,n}^{(a)}$ in the brief, draw topic indicators $z_{i,k,n}^{(a)} \sim \text{Categorical}(\boldsymbol{\Delta}_{i,k})$ and $w_{i,k,n}^{(a)} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{i,k,n}^{(a)}})$.

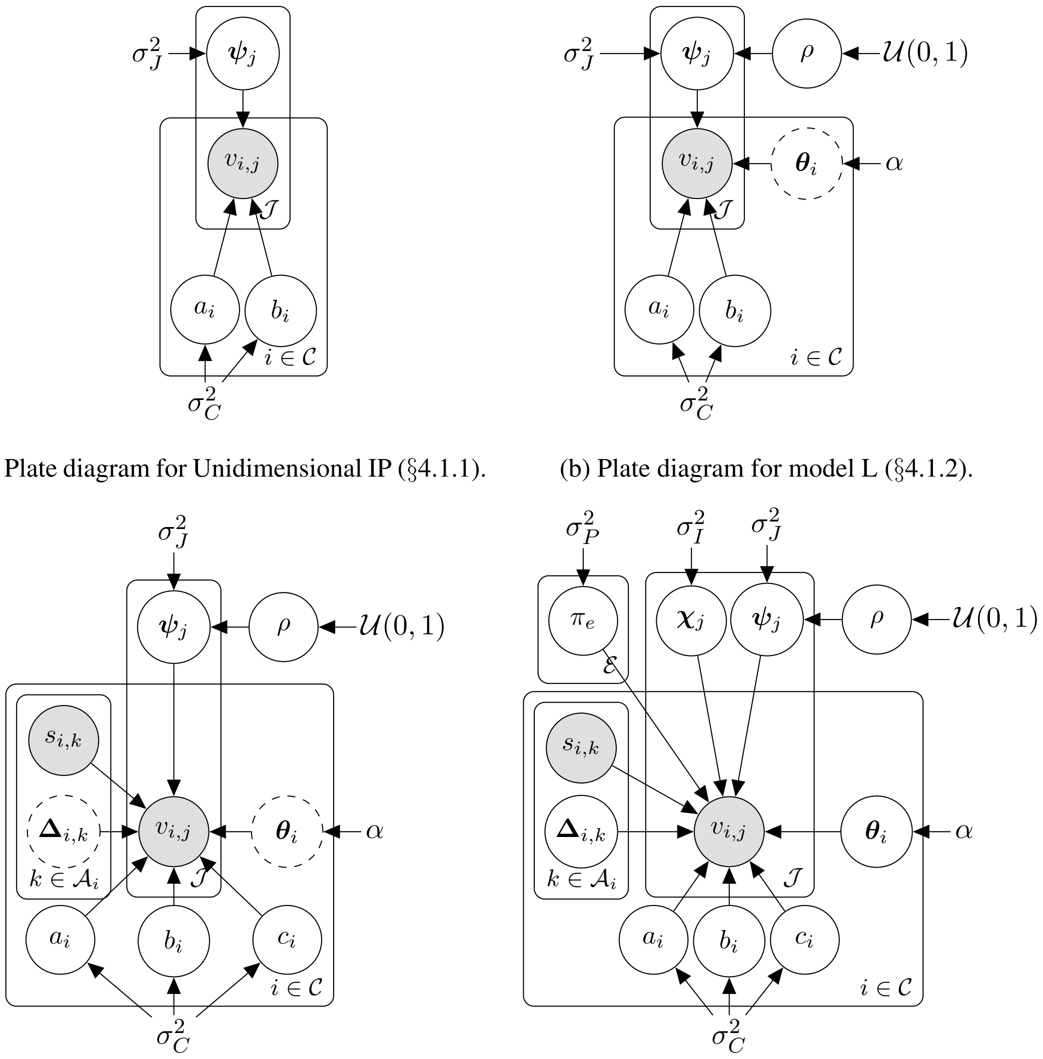
(e) For each participating justice $j \in \mathcal{J}_i$, draw vote $v_{i,j}$ according to Eq. 4.2.

We discuss the setting of hyperparameters (i.e., α, β, σs) in §4.5.3. The full likelihood of the model is:

$$
\begin{aligned}
&\mathcal{L}^{(\mathrm{P})}(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\psi}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{\Delta}, \boldsymbol{\chi}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\phi}, \rho \mid \alpha, \beta, \sigma_C, \sigma_I, \sigma_J, \sigma_P) \\
&\propto \prod_{t=1}^{T} p(\boldsymbol{\phi}_t \mid \beta) \\
&\times p(\rho) \prod_{j \in \mathcal{J}} p(\boldsymbol{\psi}_j \mid \sigma_J^2) p(\chi_j \mid \sigma_I^2) \\
&\times \prod_{e \in \mathcal{E}} p(\pi_e \mid \sigma_P^2) \\
&\times \prod_{i \in \mathcal{C}} p(a_i, b_i \mid \sigma_C^2) \\
&\times \prod_{i \in \mathcal{C}} p(\boldsymbol{\theta}_i \mid \alpha) \prod_{n} p(z_{i,n}^{(m)} \mid \boldsymbol{\theta}_i) p(w_{i,n}^{(m)} \mid \phi_{z_{i,n}^{(m)}}) \\
&\times \prod_{i \in \mathcal{C}} \prod_{k \in \mathcal{A}_i} p(\boldsymbol{\Delta}_{i,k}) \prod_{n} p(z_{i,k,n}^{(a)} \mid \boldsymbol{\Delta}_{i,k}) p(w_{i,k,n}^{(a)} \mid \phi_{z_{i,k,n}^{(a)}}) \\
&\times \prod_{i \in \mathcal{C}} \prod_{j \in \mathcal{J}_i} p(v_{i,j} \mid \boldsymbol{\psi}_j, a_i, b_i, \boldsymbol{\theta}_i, \boldsymbol{\Delta}_i, \chi_j, \boldsymbol{\pi})
\end{aligned}
$$

where $\mathcal{A}_i$ is a notational shorthand for the set of amicus briefs filed in case i and the distribution over $\boldsymbol{\Delta}$s will be discussed in detail in §4.3.

The plate diagram for voting model P is presented in Figure 4.1d.



(a) Plate diagram for Unidimensional IP (§4.1.1).

(b) Plate diagram for model L (§4.1.2).

(c) Plate diagram for model A (§4.1.3).

(d) Plate diagram for model P (§4.1.4).

Figure 4.1: Plate diagrams for voting models. $\mathcal{J}$, $\mathcal{C}$, $\mathcal{A}_i$, and $\mathcal{E}$ are the sets of justices, cases, amicus briefs (for case i), and amicus entities, respectively. The mixture proportion nodes (dashed) for model L and model A are fixed in our estimation procedure. Note that the prior on amicus briefs, $\mathbf{\Delta}$ as well as LDA, is not shown.

4.2 Modeling Opinions ("Model O")

In most SCOTUS cases, a justice is assigned to author a majority opinion, and justices voting in the majority "join" the opinion. According to Wikipedia (2016)'s entry on the procedures of SCOTUS,

> In Court, the *justice assigned to write the majority opinion will produce and circulate a draft opinion to the other justices*. Once the draft opinion has been reviewed, the remaining Justices may recommend changes to the opinion. Whether these changes are accommodated depends on the legal philosophy of the drafters as well as on how strong a majority the opinion garnered at conference. A justice may instead simply join the opinion at that point without comment.

Justices may author additional opinions concurring or dissenting with the majority, and they may choose to join concurring and dissenting opinions written by others. Here, we extend the IP model of votes to generate the opinions of a case; this marks the third major extension beyond issues IP model of Lauderdale and Clark (2014).

SCOTUS justices often incorporate language from merits and amicus briefs into their opinions (Collins et al. 2015; Ditzler 2011; Feldman 2016a,b). While amicus briefs are not usually used directly in legal analyses, the background and technical information they provide are often quoted in opinions. As such, we model opinions as a mixture of its justice-authors' topic preferences, topic proportions of the merits briefs ($\boldsymbol{\theta}$), and topic proportions of the amicus briefs ($\boldsymbol{\Delta}$). This can also be viewed as an author-topic model (Rosen-Zvi et al. 2004) where justices, litigants, and groups of amici are all effective authors. To accomplish this, we introduce an explicit switching variable x for each word, which selects between the different sources of topics, to capture the mixture proportions.

To simplify matters, we concatenate all opinions supporting the same side into a single document, i.e., opinions where justices dissent from the majority are concatenated together, and those where justices concur with the majority are concatenated with the majority opinion. However, we note that concurring opinions often contain perspectives that are different from the majority opinion and by concatenating them, we may lose some information about individual justices' styles or preferences.

Building on the generative model for votes, model P, the generative story for each case i's two opinions-documents is:

6. For each justice $j \in \mathcal{J}$, draw topics $\mathbf{\Gamma}_j \sim \text{Dir}(\alpha)$.
7. For each case $i \in \mathcal{C}$:
   (a) For each side $s \in \{\text{petitioner}, \text{respondent}\}$, draw "author"-mixing proportions:

$$\boldsymbol{\tau}_i^s \sim \text{Dir}\left(\begin{bmatrix} p(v_{i,1} = s) \\ \vdots \\ p(v_{i,|\mathcal{J}|} = s) \\ 1 \\ 1 \end{bmatrix}\right) \tag{4.3}$$

   where the last two dimensions are for choosing topics from the merits and amicus briefs, respectively.[11] Intuitively, our model assumes that opinions will incorporate more language from justices who agree with it.

   (b) For each side $s \in \{\text{petitioner}, \text{respondent}\}$ and each word $w_{i,s,n}^{(o)}$ in the opinion for side s,

      i. Draw $x_{i,s,n} \sim \text{Categorical}(\boldsymbol{\tau}_i^s)$.

      ii. If $x_{i,s,n} \in \mathcal{J}_i$, draw $z_{i,s,n}^{(o)} \sim \text{Categorical}(\mathbf{\Gamma}_{x_{i,s,n}})$, the justice's topic distribution.

      iii. If $x_{i,s,n} = \text{merits}$, draw $z_{i,s,n}^{(o)} \sim \text{Categorical}(\boldsymbol{\theta}_i)$, the merits topic distribution.

      iv. If $x_{i,s,n} = \text{amici}$, draw $z_{i,s,n}^{(o)} \sim \text{Categorical}(\mathbf{\Delta}_i^s)$, side s's amicus briefs

[11] In cases where there are less than nine justices voting, the size of $\boldsymbol{\tau}_i^{\text{p}}$ and $\boldsymbol{\tau}_i^{\text{r}}$ may be smaller.

mean topic distribution.

v. Draw word $w^{(o)}_{i,s,n} \sim \text{Categorical}(\boldsymbol{\phi}_{z^{(o)}_{i,s,n}})$.

Unlike in the Court, where an opinion is mainly authored by a single justice, all the participating justices contribute to an opinion in our generative story, with different proportions. This approach simplifies the computational model and reflects the closed-door nature of discussions held by justices prior to writing their opinions. Our model assumes that justices debate together, and that the arguments are reflected in the final opinions. In future work, we might extend the model to infer an authoring process that separates an initial author from "joiners."

The full likelihood of the opinions model is:

$$\begin{aligned}
&\mathcal{L}^{(\mathrm{O})}(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\psi}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{\Delta}, \boldsymbol{\chi}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\tau}, \boldsymbol{\Gamma}, \boldsymbol{\phi}, \rho \mid \alpha, \beta, \sigma_C, \sigma_I, \sigma_J, \sigma_P) \\
&\propto \mathcal{L}^{(P)}(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\psi}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{\Delta}, \boldsymbol{\chi}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\phi}, \rho \mid \alpha, \beta, \sigma_C, \sigma_I, \sigma_J, \sigma_P) \\
&\times \prod_{j \in \mathcal{J}} p(\boldsymbol{\Gamma}_j \mid \alpha) \\
&\times \prod_{i \in \mathcal{C}} p(\tau^{\mathrm{p}}_i, \tau^{\mathrm{r}}_i \mid \boldsymbol{v}, \boldsymbol{\psi}, a_i, b_i, \boldsymbol{\theta}_i, \boldsymbol{\Delta}_i, \boldsymbol{\chi}, \boldsymbol{\pi}) \\
&\times \prod_{i \in \mathcal{C}} \prod_{s \in \{\mathrm{p,r}\}} \prod_{n} p(x_{i,s,n} \mid \boldsymbol{\tau}^s_i)\, p(z^{(o)}_{i,s,n} \mid x_{i,s,n}, \boldsymbol{\Gamma}, \boldsymbol{\theta}_i, \boldsymbol{\Delta}^s_i)\, p(w^{(o)}_{i,s,n} \mid \boldsymbol{\phi}_{z^{(o)}_{i,s,n}})
\end{aligned}$$

In this chapter, we build the opinions model on top of model P, although conceivably, the modular nature of graphical models allows us to use the opinions model with any probabilistic model of votes.

4.3 Amici as Agents

In §4.1, the IP models focus on justices' positions embedded in a continuous space. However, we want to account for the fact that amici are rational and purposeful decision makers who write briefs to influence the outcome of a case; this assumption leads to the design of the distribution over $\boldsymbol{\Delta}$ (generative model step 4(d)i in model A (§4.1.3) and step 5(d)i in

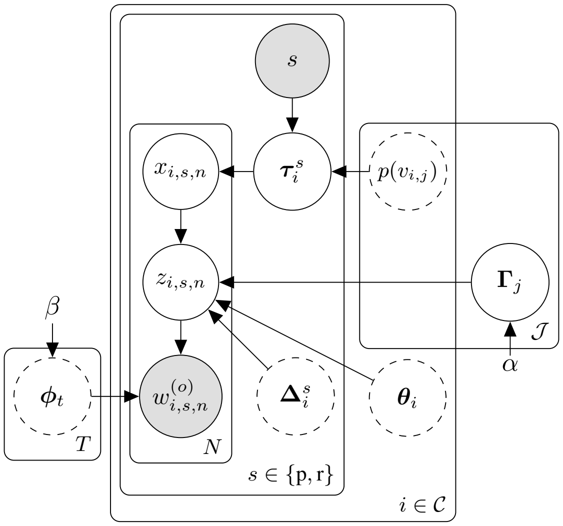


Figure 4.2: Plate diagrams for model O. The dashed nodes are fixed in our estimation procedure for opinions. Note that the dependencies on model P are not shown.

model P (§4.1.4)).

When writing a brief $\boldsymbol{\Delta}$, an amicus seeks to increase the response to her brief (i.e., votes), while keeping her costs low. We encode her objectives as a utility function, which she aims to maximize with respect to the decision variable $\boldsymbol{\Delta}$:

$$U(\boldsymbol{\Delta}) = R(\boldsymbol{\Delta}) - C(\boldsymbol{\Delta})$$

where $R(\cdot)$ is the extrinsic response (reward) that an amicus gets from filing brief $\boldsymbol{\Delta}$ and $C(\cdot)$ is the "cost" of filing the brief; dependency on other latent variables is notationally suppressed. When authoring her brief, we assume that the amicus writer has knowledge of the justices (IP and topic preferences), case parameters, and merits, but not the other amici participating in the case.[12]

Amicus curiae are motivated to position themselves (through their briefs) in such a way as to improve the likelihood that their arguments will persuade SCOTUS justices. This

[12] Capturing strategic amici agents (a petitioner amicus choosing brief topics considering a respondent amicus' brief) would require a complicated game theoretical model and, we conjecture, would require a much richer representation of policy and goals. That idea is left for future research.

is reflected in the way a justice votes, or through the language of the opinions. Hence, we investigate two response functions.

4.3.1 The Vote Seeking Amicus

Suppose we have an amicus curiae supporting side s (e.g., petitioner), which is presided by a set of justices, $\mathcal{J}$. The amicus is interested in getting votes in favor of her side, that is, $v_j{=}s$. Thus, we assume that she has a simple evaluation function over the outcome of votes $v_1, \ldots, v_9$,

$$u(v_1, v_2, \ldots, v_9) = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \mathbb{I}(v_j = s), \tag{4.4}$$

where $\mathbb{I}$ is the indicator function. The outcome of the case is uncertain, so the amicus' objective will consider her *expected* reward:

$$R^{\text{vote}}(\boldsymbol{\Delta}) = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} p(v_j = s \mid \ldots), \tag{4.5}$$

This is her **vote seeking reward**.

Note that we use an evaluation function in Eq. 4.4 that is linear in votes for simplicity. The scale of the function is unimportant (expected utility is invariant to affine transformations). However, we leave for future work other specifications of the evaluation function; for example a function that places more emphasis on the majority vote outcome.

4.3.2 The Opinion Seeking Amicus

Comparato (2003) suggests two main objectives for amici curiae: policy and maintenance goals. To achieve their policy goals, amici curiae are motivated to position themselves in such a way as to improve the likelihood that the arguments they provide will be used by SCOTUS justices. Besides pursuing their policy goals, there is a need for amici to maintain (and grow) their membership levels —– SCOTUS is a highly visible avenue for them to publicize themselves. Unlike the vote seeking amicus, success in the Court is not limited to favorable judicial outcomes (i.e., votes); their goals may still be served through participation, where they can demonstrate to current (and potential) members their active presence in particular forums and policy domains.

Thus, an alternative utility function is to maximize the (topical) similarity between her brief and the Court's opinion(s) siding with s,

$$R^{\text{opinion}}(\boldsymbol{\Delta}) = 1 - H^2(\boldsymbol{\Delta}, \boldsymbol{\Omega}^s), \tag{4.6}$$

where $H^2(P, Q) = \frac{1}{2}\|\sqrt{P} - \sqrt{Q}\|_2^2$ is the squared Hellinger (1909) distance between two distributions, and $\boldsymbol{\Omega}^s$ is the expected topic mixture under the model assumptions in §4.2. $\boldsymbol{\Omega}^s$ has the closed form:

$$\boldsymbol{\Omega}^s = \begin{bmatrix} \boldsymbol{\Gamma}_1 & \cdots & \boldsymbol{\Gamma}_{|\mathcal{J}|} & \boldsymbol{\theta} & \boldsymbol{\Delta} \end{bmatrix} \frac{\gamma^s(\boldsymbol{v})}{\|\gamma^s(\boldsymbol{v})\|_1},$$

where

$$\gamma^s(\boldsymbol{v}) = \begin{bmatrix} p(v_{i,1} = s) \\ \vdots \\ p(v_{i,|\mathcal{J}|} = s) \\ 1 \\ 1 \end{bmatrix}$$

is the same as Eq. 4.3. The expectation is simply a weighted average of the different topic mixtures from justices and documents. In short, the amicus gains utility by accurately predicting the expected opinion, thereby gaining publicity and demonstrating to members, donors, potential clients, and others that the language of the highly visible SCOTUS opinion was influenced. Therefore, Eq 4.6 is an amicus' **opinion seeking reward**.

Both Eqs. 4.5 and 4.6 reward amici when justices "agree" with them, for different definitions of agreement.

4.3.3 Cost of Writing

In addition to the policy objectives of an amicus, we need to characterize her "technology" (or "budget") set. We do this by specifying a cost function, C, that is increasing in difference between $\boldsymbol{\Delta}$ and the "facts" in $\boldsymbol{\theta}$:

$$\begin{aligned} C^{\ell_2}(\boldsymbol{\Delta}) &= \frac{\xi}{2}\|\boldsymbol{\Delta} - \boldsymbol{\theta}\|_2^2 \\ C^{H}(\boldsymbol{\Delta}) &= \xi H^2(\boldsymbol{\Delta}, \boldsymbol{\theta}) \end{aligned}$$

are two possible cost functions for the different IP vote models introduced in §4.1.3 and §4.1.4 respectively. $\xi > 0$ is a hyperparameter controlling the cost (relative to the reward function).

Both functions capture the notion that amicus briefs cannot be arbitrary text; there is disutility or effort required to carefully frame a case, or the monetary cost of hiring legal counsel. The key assumption here is that framing is costly, while simply matching the merits is easy (and presumably unnecessary). Note the role of the cost function is analogous to regularization in other contexts.

4.3.4 Random Utility Models

The outcome of the case is uncertain. In this chapter, we consider three forms *expected* utility for the amicus:

$$\begin{aligned} U^{(\mathrm{A})}(\boldsymbol{\Delta}) &= R^{\mathrm{vote}}(\boldsymbol{\Delta}) - C^{\ell_2}(\boldsymbol{\Delta}) \\ U^{(\mathrm{P})}(\boldsymbol{\Delta}) &= R^{\mathrm{vote}}(\boldsymbol{\Delta}) - C^{H}(\boldsymbol{\Delta}) \\ U^{(\mathrm{O})}(\boldsymbol{\Delta}) &= R^{\mathrm{opinion}}(\boldsymbol{\Delta}) - C^{H}(\boldsymbol{\Delta}) \end{aligned}$$

While both utility functions, $U^{(\mathrm{A})}$ and $U^{(\mathrm{P})}$ represent the vote seeking amici, they differ in two ways: different vote probabilities and cost functions.

Specifically, $U^{(\text{A})}$ is based on model A, which is

$$U^{(\text{A})}(\boldsymbol{\Delta}) = \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} \sigma\left(a + \boldsymbol{\psi}_j^\top \left(b\boldsymbol{\theta} + c^s\boldsymbol{\Delta}\right)\right) - \frac{\xi}{2}\|\boldsymbol{\Delta} - \boldsymbol{\theta}\|_2^2 \tag{4.7}$$

for an amicus filing a brief supporting side s. This is the utility function used in Sim et al. (2015).

On the other hand, $U^{(\text{P})}$ is based on model P,

$$U^{(\text{P})}(\boldsymbol{\Delta}) = \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} \sigma\left(a + b\boldsymbol{\psi}_j^\top \left(\boldsymbol{\theta} + \chi_j\pi\boldsymbol{\Delta}\right)\right) - \xi H^2(\boldsymbol{\Delta}, \boldsymbol{\theta}) \tag{4.8}$$

and contains the justice influenceability and amici persuasiveness variable described in §4.1.4. For completeness, $U^{(\text{O})}(\boldsymbol{\Delta})$, which is based on the opinions model, is

$$U^{(\text{O})}(\boldsymbol{\Delta}) = 1 - H^2(\boldsymbol{\Delta}, \boldsymbol{\Omega}^s) - \xi H^2(\boldsymbol{\Delta}, \boldsymbol{\theta}) \tag{4.9}$$

is the utility function for an amicus filing a brief supporting side s.

Recall our assumption that amici are purposeful writers whose briefs are optimized for their utility function. In an ideal setting, the $\boldsymbol{\Delta}$ which we observe will be utility maximizing. Thus, there are several conceivable ways to incorporate amici's utility functions into our estimation of justices' IP. We could maximize the likelihood and impose the constraint on $\boldsymbol{\Delta}$ that solve our expected utility optimization either directly or by checking the first-order conditions. This is reminiscent of learning frameworks where constraints are placed on the posterior distributions (Chang et al. 2007; Ganchev et al. 2010; McCallum et al. 2007). However, the nonlinear nature of our expectations makes it difficult to optimize and characterize the constrained distribution.

Instead, we can view such (soft) constraints as imposing a prior on $\boldsymbol{\Delta}$:

$$p_{\text{util}}(\boldsymbol{\Delta}) \propto \exp \eta U(\boldsymbol{\Delta}) \tag{4.10}$$

where our functional equations for utility imply $-1 \leq U(\cdot) \leq 1$.[13] η is a hyperparameter tuned using cross validation. The behavior which we observe (i.e., the amicus' topic mixture proportions) has a likelihood that is proportional to utility.

This approach is known as a **random utility model** in the econometrics discrete-choice literature (McFadden 1974). Random utility models relax the precision of the optimization by assuming that agent preferences also contain an idiosyncratic random component that is unobserved to us.

4.4 Data and Pre-processing

We used data from 30 terms of the Court from 1985–2014 (Spaeth et al. 2015)[14] using texts from LexisNexis.[15] We concatenate each case's merits briefs from both parties to form a single document, where the text is used to infer the representation of the case in topical space ($\boldsymbol{\theta}$; i.e., merits briefs are treated as "facts of the case"). In model A, each amicus brief is treated as a single document. Likewise in model O, opinions supporting the same side of the case (i.e., majority and concurring vs. dissents) were concatenated to form a single document. For our experiments with model L, model A, and model P, we did not make use of case opinions because opinions are written after votes are cast, tainting the data for predictive modeling.

The persuasive amici models, model P and model O, require identifying the authors of amicus briefs. Hence, we created regular expression rules to identify and standardize amicus authors from the header of briefs. We filtered amici who have participated in less than 5 briefs[16] and merged regional chapters of amicus organizations together (i.e., "ACLU

[13] Note, were the utility negative, the amici would have chosen not to write a the brief.

[14] The unit of analysis is the case citation, and we select cases where the type of decision equals 1 (orally argued cases with signed opinions), 5 (cases with equally divided vote), 6 (orally argued per curiam cases), or 7 (judgements of the Court). In addition, we dropped cases where the winning side was not clear (i.e., coded as "favorable disposition for petitioning party unclear").

[15] LexisNexis (`http://www.lexisnexis.com`) is an online subscription-based provider of legal content.

[16] Briefs which have no authors because of the filtering process are removed from our dataset. This occurred in about 24% of amicus briefs.

Cases	2,643
Votes	23,465
Merits briefs	16,416
Amicus briefs	16,303
Opinions	4,187
Amici	4,550
Phrases	18,207,326

Table 4.1: Statistics of our corpus after pre-processing. Cases are from the 1985–2014 terms.

of Kansas" and "ACLU of Kentucky" are both labeled "ACLU"). On the other hand, we separated labeled amicus briefs by the U.S. Solicitor General according to the presidential administration when the brief is filed (i.e., an amicus brief filed during Barack Obama's administration will be labeled "USSG-Obama"). The top five amici curiae by number of briefs filed in our dataset are American Civil Liberties Union (463), State of Utah (376), National Association of Criminal Defense Lawyers (359), State of Montana (330), and Chamber of Commerce (326).

We tokenized all the texts and part-of-speech tagged the tokens using spaCy.[17] We extract n-grams with tags that follow the simple (but effective) pattern `(Adj|Cardinal|Noun)+ Noun` (Justeson and Katz 1995), representing each document as a "bag of phrases,". We filter phrases that appear less than 100 times or in more than 8,500 documents, obtaining a final set of 48,589 phrase types. Table 4.1 summarizes details of our corpus.

4.5 Learning and Inference

The models we described above can be estimated within a Bayesian framework. However, for computational reasons, we decoupled the estimation of parts of the models from each other. Below, we describe in detail the estimation procedures for the different models.

[17] spaCy is a Python NLP library. Its part-of-speech tagger is based on the averaged perceptron tagger following Collins (2002) but with Brown cluster features as described by Koo et al. (2008), and using greedy decoding. It is available at `https://spacy.io/`.

4.5.1 Parameter Estimation for Model L and Model A

We decoupled the estimation of the topic mixture parameters as a stage separate from the IP parameters. This approach follows Lauderdale and Clark (2014), who argued for its conceptual simplicity: the text data define the rotation of a multidimensional preference space, while the second stage estimates the locations in that space. We found in preliminary experiments that similar issue dimensions result from joint vs. stage-wise inference, but that the latter is much more computationally efficient. Using LDA, $\boldsymbol{\theta}$ (and, where relevant, $\boldsymbol{\Delta}$) are estimated, then fixed to their posterior means while solving for justice parameters $\boldsymbol{\psi}$ and case parameters $\boldsymbol{a}$'s, $\boldsymbol{b}$'s, and $\boldsymbol{c}$'s.[18]

For the second stage, we used Metropolis within Gibbs, a hybrid MCMC algorithm, to sample the latent parameters from their posterior distributions (Tierney 1994). We sampled the case parameters a_i, b_i, and $c_i^{\mathrm{p,\,r}}$ for each case i and $\boldsymbol{\psi}_j$ for each justice j blockwise from a multivariate Gaussian proposal distribution, tuning the diagonal covariance matrix to a target acceptance rate of between 15–45%. Likewise, ρ is sampled from a univariate Gaussian proposal, with its variance tuned similarly. We discarded samples from the first 1,500 iterations (burn-in) and kept every 10th subsequent sample to compute the posterior mean. In total, we performed 3,000 iterations over the data.

4.5.2 Parameter Estimation for Model P and Model O

Here, we decoupled the estimation of the votes model from the opinions model; we first estimate the parameters for the votes model and hold them fixed while we estimate the new latent variables in the opinions model. In our preliminary experiments, we found that estimating parameters for both votes and opinions jointly led to slow mixing and poor predictive performance. Separating the estimation procedure into two stages allows the model to find better parameters for the votes model, which are then fed into the opinions model as priors through the vote probabilities.

[18] We used the parallel C++ implementation of fast Gibbs sampling of LDA (Liu et al. 2011) for our implementation of model L and model A.

Unlike before, we first initialize the topic mixtures ($\boldsymbol{\theta}$, $\boldsymbol{\Delta}$) and topic-word distributions ($\boldsymbol{\phi}$) of the model using latent Dirichlet allocation.[19] Then, we used Metropolis within Gibbs, to sample the latent parameters from their posterior distributions (Tierney 1994). For the Metropolis-Hastings proposal distributions, we used a Gaussian for the case parameters a, b, and justice IPs $\boldsymbol{\psi}$, log-normal distributions for χ and π, and logistic-normal distribution for the variables on the simplex $\boldsymbol{\theta}, \boldsymbol{\Delta}, \boldsymbol{\tau}$, and $\boldsymbol{\Gamma}$. We tuned the hyperparameters of the proposal distributions at each iteration to achieve a target acceptance rate of 15–45%. In each iteration, we sampled the latent variables $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{\Delta}, \boldsymbol{\chi}$, and $\boldsymbol{\pi}$ in turn using the Metropolis-Hastings algorithm (Hastings 1970). We discarded samples from the first 1,500 iterations (burn-in) and kept every 10th subsequent sample to compute the posterior mean. In total, we performed 3,000 iterations over the data.

After estimating the parameters for the vote model, we held all the parameters of the vote model fixed and sampled the opinion model parameters $\langle \boldsymbol{\Gamma}, \boldsymbol{\tau}, \boldsymbol{x}, \boldsymbol{z}^{(o)} \rangle$. We discarded samples from the first 2,500 iterations (burn-in) and keeping every 10th subsequent sample to compute the posterior mean. In total, we performed 5,000 iterations over the opinions model.

Our random utility models can also be described through a similar generative story. Instead of drawing amicus briefs ($\boldsymbol{\Delta}$) from a Dirichlet, they are drawn from the expected utility distribution (Eq. 4.10). Importantly, note that $\boldsymbol{\Delta}$ here serves as direct evidence for the justice and case parameters, rather than influencing them through v-structures. Thus, for the models with utility functions, we use the same MCMC approach in sampling the latent variables, but this time including the expected utility term for each brief in the likelihood function (Eq. 4.7 for model A, Eq. 4.8 for model P, and Eq. 4.9 for model O).

[19] We used the online variational Bayes algorithm (Hoffman et al. 2010) implementation found in Python library `scikit-learn` (Pedregosa et al. 2011) for model P and model O.

4.5.3 Hyperparameter Settings

For priors on the IP latent variables, we follow the same settings used by Lauderdale and Clark (2014), setting priors on case parameters, $\boldsymbol{\sigma}_C^2$, to 4.0, and justice IPs component-wise variance, σ_J^2 to 1.0. For model P, we experimented with $\sigma_I^2, \sigma_P^2 \in \{0.25, 0.5, 1, 2\}$ to find the best parameters using 5-fold cross validation on accuracy and perplexity. Likewise, we experimented with a range of η values in $\{0.125, 0.25, 0.5, 1, 2, 4\}$ for the utility models. Table 4.2 presents the hyperparameters for our final models.

Description	Symbol	Model L	Model A	Model P & O
No. of topics	T	30	30	128
Document-topic priors ($\boldsymbol{\theta}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}$)	α	0.1	0.1	$\frac{1}{128}$
Topic-word ($\boldsymbol{\phi}$) prior	β	0.001	0.001	0.001
Justice IP ($\boldsymbol{\psi}$) diagonal variance	σ_J^2	1.0	1.0	1.0
Case parameters (a, b, c) variance	σ_C^2	4.0	4.0	4.0
Justice influenceability (χ) scale	σ_I^2	-	-	0.5
Amicus persuasiveness (π) scale	σ_P^2	-	-	1.0
Utility cost weight	ξ	-	1.0	1.0
Model A and model P utility function weight	$\eta^{(\mathrm{A})}, \eta^{(\mathrm{P})}$	-	1.0	1.0
Model O utility function weight	$\eta^{(\mathrm{O})}$	-	-	2.0

Table 4.2: Final hyperparameter settings for our model.

4.5.4 Amicus Briefs Side Classification

The amicus briefs in the dataset were not explicitly labeled with the side that they support, and manually labeling each brief would be a tedious endeavor. In Sim et al. (2015), we built a classifier to automatically label the briefs with the side the amici are supporting, taking advantage of cues in the brief content that *strongly* signal the side that the amici is supporting (e.g., "in support of petitioner" and "affirm the judgement"). We manually labeled 1,241 randomly selected amicus briefs with its side (petitioner, respondent, neither), and trained a logistic regression classifier[20] using lexical and formatting features. We

[20] We used `scikit-learn`'s logistic regression classifier implementation (Pedregosa et al. 2011).

identified 5 sections which are common across almost all briefs and used them as features. The feature templates for our classifier are: $\langle w \rangle$, $\langle \text{title}, w \rangle$, $\langle \text{counsel}, w \rangle$, $\langle \text{introduction}, w \rangle$, $\langle \text{statement}, w \rangle$, $\langle \text{conclusion}, w \rangle$, where w can be any unigram, bigram, or trigram.

We evaluated the performance of our classifier using 5 random splits, with 50% of our data for training, 30% for testing, and 20% for the development set. We tuned the ℓ_1-regularization weights on our dev set over the range of coefficients $\{0.5, 1, 2, 4, 8, 16\}$. The average accuracy of our classifier is 79.1%. Limiting our evaluation to instances whose posterior probability after classification is greater than 0.8, we obtain 90.0% accuracy and recall of 52.1%.

4.6 Quantitative Experiments

In this section, we evaluate the performance of our novel contributions.

4.6.1 Predicting Votes

We quantify the performance of our vote model using 5-fold cross validation and on predicting future votes from past votes. Due to the specification of IP models, the probability of a vote is a logistic function of the vote-specific IP, which is a symmetric function implying that justice j's probability of voting towards the petitioner will be the same as if she voted for the respondent when we negate the vote-specific IP. Thus, we need the case parameters of new cases to predict the direction of the votes.

Gerrish and Blei (2011) accomplished this by using regression on legislative text to predict the case parameters (a, b).[21] Here, we follow a similar approach, fitting ridge re-

[21] An alternative method around the identifiability issue is to anchor the sides of a case by its liberal-conservative ideology instead of labeling the sides petitioner vs. respondent. This would require us to build a classifier to automatically label which side the litigants are on. We conducted preliminary experiments to study the feasibility of such a classifier and the details can be found in appendix A.

Gerrish and Blei (2011)'s method of using directly predicting case parameters is similar; the ideological direction of a case can be inferred from the signs of the case parameters (e,g., positive is conservative). However, we find that estimating case parameters directly using regression provides us with more information about the case beyond the ideological directions. Firstly, the sign and magnitude of the case parameters do not necessargression models on the merits brief topic mixtures $\boldsymbol{\theta}$ to predict the case parameters[22] for each case. We tuned the parameters of the regression using 5-fold cross-validation on the training data.

On the held-out test cases, we sampled the mixture proportions for the merits and amicus briefs directly using latent Dirichlet allocation with parameters learned while fitting our vote model. With the parameters from our fitted vote model and ridge regression, we can predict the votes of every justice for every case.

We compared the performance of our models with two strong baselines: (i) a majority classifier which always predict a vote for the petitioners, and (ii) a random forest trained on case-centric metadata coded by Spaeth et al. (2015) to make predictions on how justices would vote (Breiman 2001; Katz et al. 2014).[23] Table 4.3 shows performance on vote prediction.

We evaluated the models using 5-fold cross validation, as well as on forecasting votes in 2013 and 2014 (trained using data from 1985 to the preceding year). In general, models incorporating utility functions outperformed their counterparts, while the improvement in accuracy of model P over model A is small; most likely because both models are very similar, the main difference being the parametrization of amicus briefs. Apart from the 2013 test set, our utility models outperformed the baseline approaches. In the 2013 test set, the distribution of votes is significantly skewed towards the petitioner (compared to the training data), which resulted in the most frequent class classifier performing much better than everything else.

ily reflect the liberal-conservative positioning of a case. The model does not inherently know the ideological leaning of each side and instead it is inferred from the data; in our Supreme Court setting, it happens to correlate with our understanding of the political spectrum in the United States. Secondly, each case parameter has different effects on the vote probability corresponding to the nature of the different brief types. Further, the magnitude of the case parameters signal their relative importance to the vote probability (i.e., larger values pushes the vote towards one side more). Lastly, it is also convenient that the model would be able to directly return the vote without having to map between sides and ideological directions.

[22] model A has case parameters a, b, and $c^{\mathrm{p,r}}$, while model P's case parameters are a and b.

[23] We used the random forest implementation in scikit (Pedregosa et al. 2011) with 128 trees in a forest, maximum tree depth of 16, and using the default settings for all other model hyperparameters. We selected the hyperparameters (number of trees and maximum tree depth) based on computational time and 5-fold cross validation average performance.

Model	5-fold	2013	2014
Most frequent	59.7 ± 0.4	**69.4**	65.0
Random forest	65.1 ± 0.5	64.8 ± 0.1	63.3 ± 0.04
Unidimensional IP (§4.1.1)	62.5 ± 0.7	61.9 ± 0.1	61.6 ± 0.1
Model L (§4.1.2)	64.0 ± 0.2	63.6 ± 0.1	63.9 ± 0.1
Model A (§4.1.3)	65.8 ± 0.4	65.1 ± 0.2	65.6 ± 0.3
Model P (§4.1.4)	66.1 ± 0.3	65.5 ± 0.2	66.0 ± 0.2
Model A with $U^{(\mathrm{A})}$	67.5 ± 0.3	65.8 ± 0.1	66.1 ± 0.2
Model P with $U^{(\mathrm{P})}$	**68.5** ± 0.4	66.4 ± 0.1	**67.2** ± 0.3

Table 4.3: Accuracy of vote prediction (%). $U^{(\mathrm{A})}$ (Eq. 4.7) and $U^{(\mathrm{P})}$ (Eq. 4.8) are utility functions we specified in §4.3.4. There are 70 cases (625 votes) and 69 cases (619 votes) in the 2013 and 2014 test sets, respectively. For 2013 and 2014 test sets, we trained the model 5 times to obtain the mean and standard deviation of the prediction accuracy (all the models except "Most frequent" contain some degree of randomness in their learning algorithm).

4.6.2 Predicting Opinions

We also estimated model O using the opinion seeking utility function in Eq. 4.9. As mentioned earlier in §4.5.2, we first estimate a vote model, then hold the parameters fixed while we sampled the opinion model parameters $\langle \boldsymbol{\Gamma}, \boldsymbol{\tau}, \boldsymbol{x}, \boldsymbol{z}^{(o)} \rangle$. When estimating the initial vote model, we experimented with model P both with and without the utility function in $U^{(\mathrm{P})}$ (Eq. 4.8).

Perplexity on a test set is commonly used to quantify the generalization ability of probabilistic models and make comparisons among models over the same observation space. Thus, we use perplexity as a proxy to measure the opinion content predictive ability of our model. For a case with opinion $\boldsymbol{w}^{(o)}$ supporting side s, the perplexity is defined as,

$$\text{perplexity}(\boldsymbol{w}^{(o)} \mid s) = \exp\left(-\frac{\log p(\boldsymbol{w} \mid s, \ldots)}{N}\right)$$

where N is the number of tokens in the opinion and a lower perplexity indicates better generalization performance. The likelihood term can be approximated using samples from the inference step.

Table 4.4 shows the perplexity of model O on opinions in the different test sets. We compared against using initial vote models that do not include $U^{(\mathrm{P})}$ to evaluate the sensitivity of the opinion model to vote model's parameters. Additionally, we compared against two baselines trained on just the opinions: one using LDA[24] and another using the author-topic model (Rosen-Zvi et al. 2004). For the author-topic model, we treat each opinion as being "authored" by the participating justices, a pseudo-author representing the litigants which is shared between opinions in a case, and a unique amicus author for each side.

Our model with $U^{(\mathrm{O})}$ achieves better generalization performance than the simpler baselines, while we do not see significant differences in whether the first stage vote models use $U^{(\mathrm{P})}$. This is not surprising since the vote model's results are similar with or without utility and it influences the opinion model indirectly through priors and $U^{(\mathrm{O})}$.

In model O, the latent variable $\mathbf{\Gamma}_j$ captures the proportion of topics that justice j is likely to contribute to an opinion. When j has a high probability of voting for a particular side, our informed prior increases the likelihood that j's topics will be selected for words in the opinion. While $\mathbf{\Gamma}_j$ serves a similar purpose to $\boldsymbol{\psi}_j$ in characterizing j through her ideological positions, $\boldsymbol{\psi}_j$ relies on votes and gives us a "direction" of j's ideological standing, whereas $\mathbf{\Gamma}_j$ is estimated from text produced by the justices and only gives us the "magnitude" of her tendency to author on a particular issue. In Table 4.5, we identify the top topics in $\mathbf{\Gamma}_j$ by considering the deviation from the mean of all justice's $\mathbf{\Gamma}$, i.e., $\Gamma_{j,k} - \frac{1}{|\mathcal{J}|}\sum_j \Gamma_{j,k}$.

4.6.3 Justice Influenceability

In model P, the latent variable χ_j measures the relative effect of amicus briefs on justice j's vote IP; when χ_j is large, justice j's vote probability is affected by amicus briefs more. Since χ_j is shared between all cases that a justice participates in, χ_j should correspond to how much they value amicus briefs. Some justices, such as Scalia, are known to be dubious of amicus briefs, preferring to leave the task of reading these briefs to their law clerks, who

[24] We used Python `scikit-learn`'s LDA module (Pedregosa et al. 2011) which implements the online variational Bayes algorithm Hoffman et al. 2010.

Model	5-fold	2013	2014
LDA	2.86 ± 0.07	2.67 ± 0.02	2.63 ± 0.01
Author-Topic	2.62 ± 0.13	2.36 ± 0.03	2.25 ± 0.05
Model O	$2.45 \pm 0.11^*$	$2.27 \pm 0.03^*$	$2.11 \pm 0.02^*$
	$2.43 \pm 0.14^{\dagger}$	$2.26 \pm 0.04^{\dagger}$	$2.13 \pm 0.04^{\dagger}$
Model O with $U^{(\mathrm{O})}$	$\mathbf{2.07}\pm 0.11^*$	$1.98 \pm 0.01^*$	$\mathbf{1.94}\pm 0.05^*$
	$2.10 \pm 0.14^{\dagger}$	$\mathbf{1.91}\pm 0.06^{\dagger}$	$1.96 \pm 0.08^{\dagger}$

Table 4.4: Perplexity of Court's opinions ($\times 10^3$). † indicates that the model is initialized with model P and $U^{(\mathrm{P})}$ while * indicates it is initialized with model P only. A lower perplexity indicates better generalization performance. There are 30,133 phrases (98 opinions) and 23,706 phrases (109 opinions) in the 2013 and 2014 test set, respectively. For 2013 and 2014 test sets, we trained the model 5 times to obtain the mean and standard deviation of the perplexity.

will pick out any notable briefs for them. In fact, the late Justice Scalia once remarked (Amick 2009):

> *Don't re-plow the ground that you expect the parties to plow unless you expect the parties to plow with a particularly dull plow.*
>
> — Justice Antonin Scalia (1936–2016)

when referring to the numerous repetitive amicus briefs that are filed. Therefore, we will expect Scalia to have a smaller χ than other justices.

In Table 4.6, we compare the χ values of justices with how often they cite an amicus brief in any opinion they wrote (Franze and Anderson 2015). We find the χ values estimated by our model are consistent with our expectations.

4.7 Qualitative Analysis

Our models enable us to address interesting exploratory analyses and hypotheses generation. In this section, we will illustrate several such analyses in detail.

The topics and justices' IPs learned from our corpus using model A can be found in appendix B and C respectively. Since we used $T = 128$ topics for model P, we do not

John G. Roberts	
32	speech, first amendment, free speech, message, expression
61	eeoc, title vii, discrimination, woman, civil rights act
52	sec, fraud, security, investor, section ##b
Antonin Scalia	
94	42 USC 1983, qualified immunity, immunity, official, section ####
57	president, senate, executive, article, framer
80	class, settlement, rule ##, class action, r civ
Anthony M. Kennedy	
57	president, senate, executive, article, framer
94	42 USC 1983, qualified immunity, immunity, official, section ####
15	plea, trial counsel, strickland, magistrate, guilty plea
Clarence Thomas	
5	federal government, framer, commerce, commerce clause, lopez
32	speech, first amendment, free speech, message, expression
72	due process, liberty, fourteenth amendment, hearing, forfeiture
Ruth B. Ginsburg	
61	eeoc, title vii, discrimination, woman, civil rights act
80	class, settlement, rule ##, class action, r civ
96	taxpayer, bank, corporation, fund, irs
Stephen Breyer	
96	taxpayer, bank, corporation, fund, irs
61	eeoc, title vii, discrimination, woman, civil rights act
15	plea, trial counsel, strickland, magistrate, guilty plea
Samuel A. Alito	
32	speech, first amendment, free speech, message, expression
61	eeoc, title vii, discrimination, woman, civil rights act
52	sec, fraud, security, investor, section ##b
Sonia Sotomayor	
22	sentence, offense, release, guidelines, guideline
23	legislature, voter, race, 42 USC 1973, minority voter
52	sec, fraud, security, investor, section ##b
Elena Kagan	
34	candidate, buckley, 424 US 1, contribution, fec
96	taxpayer, bank, corporation, fund, irs
105	fda, drug, manufacturer, product, federal law

Table 4.5: Top 3 topics contributed to Court opinions by recent justices ($\mathbf{\Gamma}$).

Justice	χ_j	Citation rate (%)
Sonia Sotomayor	1.590	45
Elena Kagan	0.714	40
Stephen G. Breyer	0.637	38
Ruth B. Ginsburg	0.515	41
John G. Roberts	0.495	42
Anthony M. Kennedy	0.468	42
Samuel A. Aalito	0.286	27
Antonin Scalia	0.268	22
Clarence Thomas	0.162	25

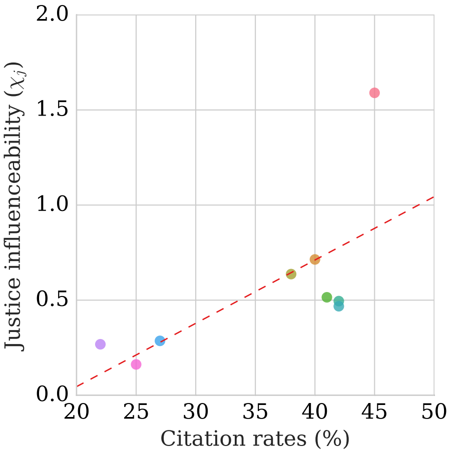


Table 4.6: Justice χ values and their average amicus citation rates between 2010–2015, provided by Franze and Anderson (2015). The adjacent figure presents the same findings in a linear plot; the Spearman's ρ between χ_j and citation rates is 0.678.

present the entire set of topics and justices IPs learned using model P.

In the following sections (§4.7.1–4.7.2), we use model A with utility function $U^{(\mathrm{A})}$ for our analyses and examples although the same can be done with our other models as well.

4.7.1 Post Hoc Analysis of Votes

On a case level, we can tease apart the relative contribution each textual component to a justice's decision by analyzing the case parameters learned by our utility models. By zeroing out the various case parameters of model A, and plotting them, we can visualize the different impact that each type of text has on a justice's vote-specific IP.

For example, Figure 4.3 shows the vote-specific IP estimates of justices for the 2011 term death penalty case *Maples v. Thomas*, 132 S. Ct. 912 (2012).[25] The issues-only IPs are computed by zeroing out both the amicus polarity parameters (c^{p} and c^{r}). On the other hand, the IP due to amicus briefs supporting Maples is computed by zeroing out only c^{r}

[25] The Court ruled that the petitioner and death row inmate, Cory Maples, should get another opportunity to appeal his death sentence because of his lawyers' unannounced and unauthorized abandonment.

(and zeroing out c^{p} for briefs supporting Thomas).

We observe that the issues-only IPs are aligned with each justice's (widely known) ideological stance on the issue of capital punishment. For instance, the issues-only IPs of Thomas, Scalia, Alito, and Roberts, the strong conservative bloc, favor the respondents (that Maples should not be awarded relief); so did Kennedy, who is widely recognized as the swing justice. When effects of all amicus briefs are taken into account, the justices' IPs shift toward Maples with varying magnitudes, with the result reflecting the actual ruling (7–2 with Thomas and Scalia dissenting).

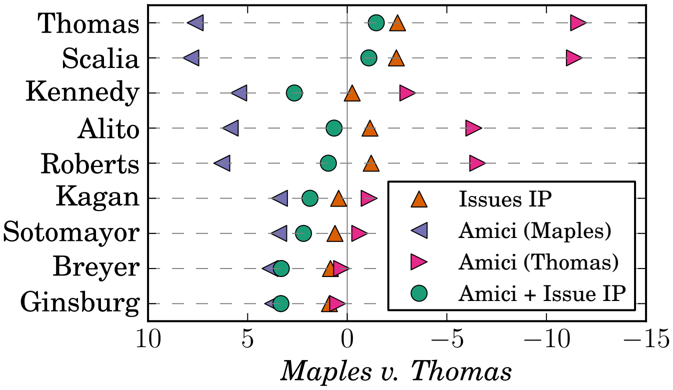


Figure 4.3: Vote-specific IP estimates decomposed into different influences on each justice's vote on *Maples v. Thomas*. The numbers on the x-axis represent the log-odds of a vote for the petitioners. Therefore, an IP towards the left (right) indicates higher probability of vote that is favorable to Maples (Thomas).

4.7.2 Counterfactual Analyses

The study of counterfactuals has engaged the interest of researchers in a wide range of domains such as philosophy (Goodman 1947), psychology (Fillenbaum 1974), and social sciences (Cowan and Foray 2002; Fearon 1991). Counterfactual conditionals also form the basis of experimental methods for establishing causality in the natural and social sciences, e.g., whether a medical treatment leads to a cure.[26] In our setting, our data is observational (i.e., we have no control over the parties in the Supreme Court), we cannot perform

[26] Pearl (2009)'s framework for defining counterfactuals gives a clear account of connections between causal models and probabilistic models.

experiments to answer the “what if” questions. While there are techniques for estimating causal effects from observational data (i.e., identifying close substitutes, randomization, or statistical adjustments, see Gelman and Hill (2007, chapter 9) and Bottou et al. (2013)), we take a model-based predictive approach to answering counterfactual questions; we use our model (i.e., model A) to simulate potential outcomes by modifying the variables.

However, there are limits to making predictions using our models which should be made explicit. Firstly, our predictions are model-dependent, they are made based on the assumptions we make about the conditional dependence between variables as well as their distributional properties. Problems such as incorrect model specification, endogenity, noisy data, etc will hinder the validity of our predictions. The second limit to our approach is that our estimation and inference methods are approximate. These reasons suggest that we should take caution when interpreting the results of our counterfactual analyses. We should keep in mind that the analyses below are exploratory and view them as a form of hypotheses generation.

As an illustration, we consider *National Federation of Independent Business (NFIB) v. Sebelius (HHS)*, 132 S. Ct. 2566 (2011), a landmark 2011 case in which the Court upheld Congress’s power to enact most provisions of the Affordable Care Act (ACA).[27] The case attracted much attention, including a record 136 amicus briefs, of which 76 of these briefs are used in our dataset. 58 (of the 76) were automatically classified as supporting NFIB.

In the merits briefs, the topics discussed revolve around *interstate commerce* and the *individual mandate* (see Figure 4.4 for the dominant topic proportions), while there is an interesting disparity in topics between briefs supporting NFIB and HHS. Notably, amici supporting NFIB are found, on average, to use language concerning *individual mandate*, while amici supporting HHS tend to focus more on topics related to *interstate commerce*. This is commensurate with the main arguments put forth by the litigants, where NFIB

[27] Commonly known as Obamacare, the case concerns Congress’s authority to enact provisions in the ACA, and the Health Care and Education Reconciliation Act (HCERA), including a requirement for most Americans to have health insurance by 2014.

was concerned about the overreach of the government in imposing an individual mandate, while HHS argued that healthcare regulation by Congress falls under the Commerce Clause. During test time, our model was most uncertain about Roberts and Kennedy, and wrong about both (Figure 4.6a).

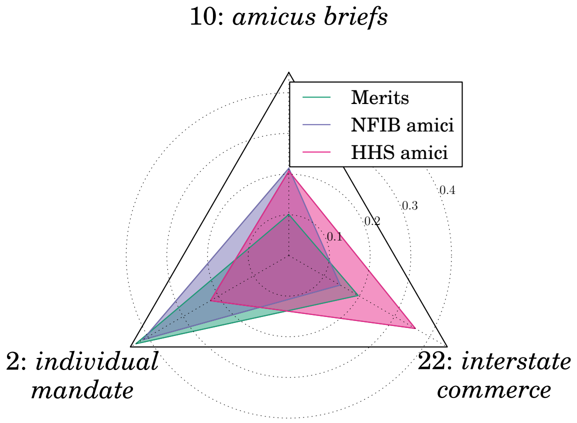


Figure 4.4: Top 3 topic proportions for merits briefs and amicus briefs. The amicus briefs topic is a collection of terms commonly associated with amicus briefs.

Choosing Sides

The first type of counterfactual analysis that we introduce is: "What if no (or only one side's) amicus briefs were submitted in the ACA case?" To answer it, we hold the case out of the training set and attempt to predict the votes under the hypothetical circumstances with the random utility model. Figure 4.6a shows the resulting IP of hypothetical situations where no amicus briefs were filed, or when only briefs supporting one side are filed.

If no amicus briefs were filed, the model expects that all but Kagan and Sotomayor would favor NFIB, but with uncertainty. With the inclusion of the amicus briefs supporting NFIB, the model becomes more confident that the conservative bloc of the court would vote in favor of NFIB (except for Alito). Interestingly, the model anticipates that the same briefs will turn the liberals *away*. In contrast, the briefs on HHS' side have more success in swaying the case in their favor, especially the crucial swing vote of Kennedy (although it turned out that Kennedy sided with the conservative bloc, and Roberts emerged as the deciding vote in HHS favor). Consequently, the model can provide insights about judicial decisions, while postulating different hypothetical situations.

Choosing What to Write

Another counterfactual analysis we can perform, more useful from the viewpoint of the amicus, is: “How should an amicus frame arguments to best achieve her goals?” In the context of our model, such an amicus would like to choose the topic mixture $\boldsymbol{\Delta}$ to maximize her expected utility (Eq. 4.10). Ideally, one would compute such a topic mixture by maximizing over both $\boldsymbol{\Delta}$ and vote outcome $\boldsymbol{v}$, while integrating over the case parameters. We resort to a cheaper approximation: analyzing the filer’s expected utility curve over two particular topic dimensions: the *individual mandate* and *interstate commerce* topics. That is, we compute the expected utility curve faced by a single amicus as we vary the topic proportions of *individual mandate* and *interstate commerce* topics over multiples of 0.1. Figure 4.5 presents the expected utility curve. The model expects an amicus on NFIB’s side to get more votes and hence, higher utility, as the model expects justices to be in favor of NFIB prior to amici influence (see Figure 4.6a).

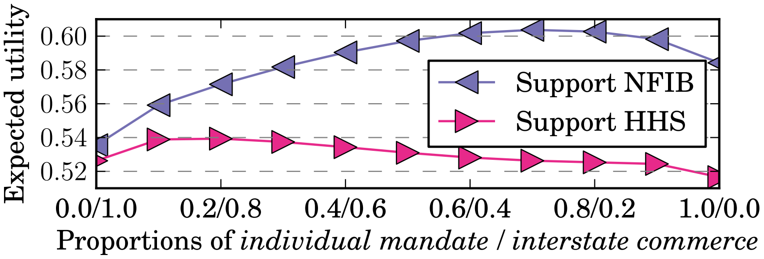
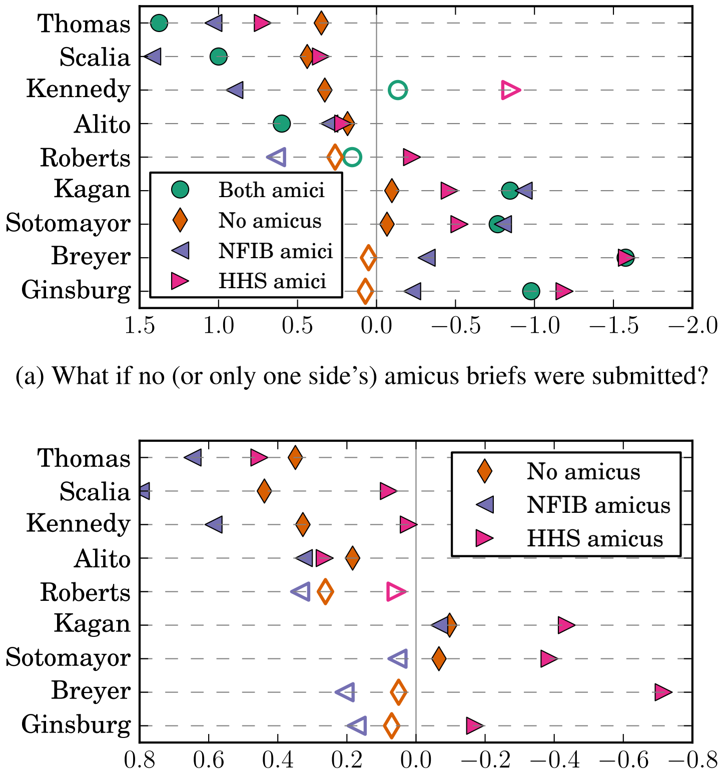


Figure 4.5: Expected utility when varying between proportions of *individual mandate* and *interstate commerce* topics. During our experiment, we set the proportions of inactive topics to 10^{-8} instead of 0. We note that interior solutions for the expected utility optimization only exist when proportions are $\in (0, 1)$.

Consequently, the hypothetical amicus who supports NFIB can expect to maximize their expected utility (5.2 votes at a cost of 0.21) by “spending” about 70% of their text on



(a) What if no (or only one side's) amicus briefs were submitted?

(b) What if a single amicus files an "optimally" written brief?

Figure 4.6: Counterfactual analyses for *National Federation of Independent Business (NFIB) v. Sebelius (HHS)*. Hollow markers denote that the prediction differed from the actual outcome.

individual mandate. On the other hand, the best that an amicus supporting HHS can do is to write a brief that is 80% about *interstate commerce*, and garner 4.7 votes at a cost of 0.31. We plot the justices' predicted IPs in Figure 4.6b using these "best" proportions.

The "best" proportions IPs are different (sometimes worse) from that in Figure 4.6a because in Figure 4.6a, there are multiple amici influencing the case parameters (through their utility functions) and other topics are present which will sway the justices. From the perspective of an amicus supporting HHS, the two closest swing votes in the case are Roberts and Kennedy; we know *a posteriori* that Roberts sided with HHS.

4.7.3 Amici Persuasiveness

Recall in model P, the latent variable π_e captures the model's belief about amicus e's brief's effect on the case IP, which we call "persuasiveness." A large π_e indicates that across the dataset, e exerts a larger effect on the case IPs, that is, according to our model, she has a larger impact on the Court's decision than other amicus. Figure 4.7 is a swarm plot illustrating the distribution of π values for different types of amicus writers.

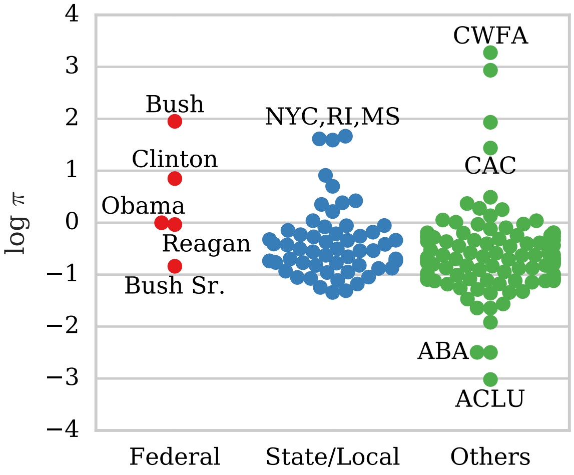


Figure 4.7: Amici "persuasiveness" by organization type. *Federal* refers to different presidential administration's federal government (and represented by the U.S. Solicitor General) and *State/Local* refers to state and local governments. The abbreviated amici are New York City (NYC), Rhode Island (RI), Mississippi (MS), Concerned Women For America (CWFA), Constitution Accountability Center (CAC), American Bar Association (ABA), and American Civil Liberties Union (ACLU).

Our model infers that governmental offices tend to have larger π values than private organizations, especially the U.S. Solicitor General. The average π for *Federal*, *State/Local* and *Others* are 2.35, 1.11, and 0.929 respectively. In fact, Lynch (2004) found through interviews with SCOTUS law clerks that "amicus briefs from the solicitor general are 'head and shoulders' above the rest, and are often considered more carefully than party briefs."

Another interesting observation from Figure 4.7 is the low π value for ACLU and ABA, despite being prolific amicus brief filers. While it is tempting to say that amici with low π values are ineffective, we find that there is almost no correlation between π and the proportion of cases where they were on the winning side.[28] Note that our model does not assume that a "persuasive" amicus tends to win. Instead, an amicus with large π will impact the case IP most, and thus explain a justice's vote or opinion (even dissenting) more than the other components in a case.

Insofar as π explains a vote, we must exercise caution; it is possible that the amicus played no role in the decision-making process and the values of π_e simply reflect our modeling assumptions and/or artifacts of the data. Without entering the minds of SCOTUS justices, or at least observing their closed-door deliberations, it is difficult to measure the influence of amicus briefs on justices' decisions.

4.8 Related Work

Poole and Rosenthal (1985) introduced the IP model, using roll call data to infer latent positions of lawmakers. Since then, many varieties of IP models have been proposed for different voting scenarios: IP models for SCOTUS (Martin and Quinn 2002), multidimensional IP models for Congressional voting (Clinton et al. 2004; Heckman and Snyder 1996), grounding multidimensional IP models using topics learned from text of Congressional bills (Gerrish and Blei 2012b) and SCOTUS opinions (Lauderdale and Clark 2014). Segal-Cover scores (Segal and Cover 1989), obtained by manual coding of pre-confirmation news articles, are another popular method for characterizing behaviors of justices.

Amici have been studied extensively, especially their influence on SCOTUS (Caldeira and Wright 1988; Collins 2008; Corley et al. 2013; Kearney and Merrill 2000). Particularly, Collins (2007) found that justices can be influenced by persuasive argumentation presented by organized interests. These studies focus on ideology metadata (liberal/conservative slant of amici, justices, decisions, etc.), disregarding the rich signals encoded in the text of these

[28] The Spearman's ρ between π and the proportion of winning sides is -0.0549. On average, an amicus supports the winning side in 55% of cases. For the ACLU, ABA, CAC, and CWFA, the proportions are 44%, 50%, 47%, and 50% respectively.

briefs, whereas we use text as evidence of utility maximizing behavior to study the influence of amici curiae. Hansford (2004) studied how amici decide whether to participate in a case, finding that amici participate in situations where justices are "information poor" or where cases allow for "high visibility."

We view amicus briefs as "purposeful" texts, where authors are writing to maximize their utility function. It is related to Gentzkow and Shapiro (2010)'s work on modeling the purposeful "slant" of profit-maximizing newspapers looking to gain circulation from consumers with a preference for such slant, and Jelveh et al. (2015)'s work modeling economists who choose research topics to maximize certain career outcomes. More generally, extensive literature in econometrics estimates structural utility-based decisions (Berry et al. 1995, *inter alia*).

In addition to work on IP models, authorship (Li et al. 2013) and historical (Wang et al. 2012) analysis has been done on SCOTUS opinions, and oral argument transcripts have been used to study power relationships (Danescu-Niculescu-Mizil et al. 2012; Prabhakaran et al. 2013) and pragmatics (Goldwasser and Daumé III 2014).

4.9 Conclusion

Our model makes several simplifying assumptions: (i) it ignores the effects of other amici on a single amicus' writing; (ii) amici are treated modularly, with a multiplicative effect and no consideration of diminishing returns or temporal ordering; (iii) the cost function does not capture the intricacies of legal writing style (i.e., choice of citations, artful language, etc.); (iv) the utility functions does not fully capture the agenda of each individual amicus. Despite these simplifications, we have shown that our models are useful for quantitative analysis and hypothesis generation in support of substantive research on the judiciary.

We presented a random utility model of the Supreme Court that is more comprehensive than earlier work. We incorporated amicus briefs and opinions, considered individual amicus' persuasiveness and their motivations through two different types of utility functions. Through our novel contributions to ideal point models, we can now study the influence of amicus briefs on Supreme Court votes as well as infer and compare the relative effectiveness of individual amicus. In the domain of SCOTUS, this leads to improved vote prediction performance, as the model captures the structure of amicus briefs better than simpler treatments of the text. Moreover, our opinions model and opinion utility function achieved better generalization performance than simpler methods as well.

More importantly, we have used our model to address interesting counterfactual questions. Were some amicus briefs not filed, or had they been written differently, or had the facts of the case been presented differently, or had different justices presided, our approach can estimate the resulting outcomes.

In this chapter, we see once again that random utility models for persuasive text are similar to a classical generative model and can be estimated using familiar algorithms. The key distinction is that persuasive text is modeled as a function of the addressee and the particulars of the matter about which she is being convinced; authors are agents seeking to maximize their expected utility in a given scenario.

Chapter 5

Conclusion and Future Directions

> Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world.
>
> *Albert Einstein*

In this thesis, we explored a novel approach for text modeling, from the perspective that text is both *purposeful* and *strategic*. We had two goals throughout the thesis: (i) to develop methods to model authors and their motivations through their textual artifacts, and (ii) to use our models to perform in-depth exploration of the data and examine a variety of hypotheses. We built on the machinery of probabilistic models and decision theory to develop *random utility models* that capture our intuitions about authors and their motivations. Probabilistic models enable us to encode our assumptions of statistical relationships between authors' latent attributes, their text, and the audience's responses. They also provide a convenient, modular framework where we can easily incorporate utility functions to describe the author's motivations. We presented instantiations of these random utility models in the domain of scientific publishing and the judiciary, and empirically demonstrated the soundness of our proposed methods. We obtained better modeling performance than simpler treatments of text, and we illustrated several examples of in-depth analyses of our data.

In recent years, there is an increasing emphasis on data-driven empirical methods in the social sciences as a result of ubiquitous availability of digitized data and affordable computational power. A substantial share of this data is text and thus NLP presents an opportunity for quantitative social science researchers. This thesis takes steps toward tackling some of the challenges of this cross-disciplinary research program. In the sequel, we will elaborate on subsequent steps and interesting directions that deserve further pursuit.

Richer models of utility. The utility functions we proposed in Chapters 3 and 4 were simple functions where the response is usually singular. Needless to say, this is an oversimplification; authors are often balancing between multiple desired responses. For instance, in the Supreme Court, amici are interested in both getting votes and seeking judicial validation for their arguments. The extent to which they care about each type of response also varies, just like in Chapter 3, where some scientists are more interested in seeking citations than others. Further, it is conceivable to have response functions that are exogenous to the model, for instance, changes in donations or membership for an amicus organization after filing a brief.[1] Moreover, there are different ways response functions can be combined together in a utility function — be it a simple linear combination or a product combination. Therefore, there is more work to be done.

As we have hinted in Footnote 1 of §3.1.3, collaborative efforts between authors may be more than the sum of its parts. An NLP researcher and a social scientist with different expertise might derive more utility from collaborating together than each on her own. Through the lens of utility functions, we can investigate the co-authorship behaviors of authors through their textual outputs. This would be useful as a tool to identify potential collaborative opportunities for researchers in today's increasingly multidisciplinary research environment.

In the SCOTUS setting, there are multiple amici competing for the justice's vote, yet we did not consider the game theoretic aspect of the amici (§4.3 Footnote 12). While figuring out the game theoretic equilibrium of competing amici would be an ardous endeavor,

[1] While the response function is exogenous to the model, there may be confounding variables affecting both.

a potential starting point would be to design a reward function for an amicus that considers the aggregate response of his opponents, i.e., he gains utility by focusing his efforts on countering the arguments in all his opponents' briefs.

Richer representations of text. In this thesis, we represented documents as bags of n-grams and phrases, in the process losing much of the rich contextual knowledge encoded in the linguistics structures (e.g., syntax and semantics). While topic modeling may be able to recover high level semantic structures within a document, we would like richer representations of text that can lead to deeper understanding of how language is used for influence. Cano-Basave and He (2016) studied the impact of persuasive argumentation in political debates through semantic frames (Fillmore 1982) and rich linguistics features, while Bamman and Smith (2015) estimated political orientation of authors by extracting fine-grained political statements from news websites. On the other hand, argumentation mining is a growing research area at the intersection of natural language processing and argumentation theory (Eemeren and Grootendorst 2004); techniques from argumentation mining will enable us to model the argumentation process in a document, i.e. rhetorical or argumentative relationships between propositions. Palau and Moens (2009) presents an introduction of argumentation mining and related computational problems. The above are just some examples of how linguistically motivated representations can be useful for answering social science questions. It also presents a direction for us to improve the our models.

The Supreme Court, along with many other textual datasets in the political arena, provide an ideal setting for investigating the phenomenon of framing.[2] Framing is an aspect of language that is intertwined with persuasion and influence; if we have a richer representation of policy issues, we will be able to learn how authors (e.g., lawyers, politicians, journalists) use language for influence and persuasion. In this regard, we can use the Media Frames Corpus (Card et al. 2015), which provides a framework to understand and analyze

[2] Entman (1993) proposes that "to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."

framing around policy issues.

Tools and applications. We anticipate the work described in our thesis to be useful for building tools that aid researchers in exploring *strategic* text collections. Through an interactive user interface, users can easily visualize the texts and the inferred latent variables. We developed such an interface for the CLIP model.[3] Our website presents a graphical view of candidate speeches and the posterior probabilities that our model learned. It will be convenient for political scientists to interpret the results of the model which in turn improves transparency and allows for external validation. Additionally, such an interface should also allow user to specify inputs and query the model. For instance, a tool based on our model of the Supreme Court will make it convenient for users to query the model for counterfactuals and visualize the vote probabilities of justices (e.g., Figure 4.6b).

Latent variable neural networks for social science. In recent years, there has been a tsunami of neural network research being published at NLP conferences (Manning 2015). Neural networks employ dense distributed representations that can capture a range of linguistic phenomena, ranging from meaning of individual words to long range contexts in sentences and beyond (Le and Mikolov 2014; Turian et al. 2010). While neural models often yield very accurate predictive models, they are a black-box; that is, studying their structure won't give us any insights to our problem, and model interpretability is an important aspect in computational social science. On the other hand, graphical models are flexible; we can easily specify our assumptions and prior knowledge without sacrificing interpretability. By combining both approaches in a hybrid architecture, for example, we can take advantage of neural networks to learn distributed representations in places where we have a lot of data (e.g., language models) and latent variables where we want interpretability (e.g., justice and amicus behavior). Hybrid models involving both neural networks and latent variables are a very recent development (Ji et al. 2016; Kingma and Welling 2014) and is a promising direction for future work in building better models for exploratory analysis

[3] http://www.cs.cmu.edu/~ark/CLIP/results.html

in social science.

Appendices

Appendix A

Classifying Supreme Court Briefs By Ideology

The experiments described in this chapter was conducted in collaboration with Nelson Liu.

A.1 Introduction

In the specification of ideal point models, the probability of a vote is based on the two parameter logistic model (Swaminathan and Gifford 1985):

$$p(v = \text{petitioner} \mid a, b, \theta) = \frac{1}{1 + \exp\left(-a - b\theta\right)}$$

where θ is the vote-specific ideal point. As a result, the probability of vote for petitioner will be the same as that for the respondent if we negate θ. Also, since the petitioner and respondent labels are arbitrary, we need a way to ground the IP model for vote prediction. In Poole and Rosenthal (1985), they find that the liberal-conservative ideology spectrum explains most of the variance in U.S. Congressional votes. Hence, in this chapter, we describe experiments on classifying Supreme Court briefs by its political ideology.

A.2 Data

Our dataset consists of 44,528 merits and amicus briefs from 70 terms of the Court from 1946 – 2015 (see Chapter 4 and Sim et al. (2015)). Additionally, each brief is associated with case metadata from Spaeth et al. (2015), which contains manual annotations on the ideological "direction" of each side: whether it was a conservative, liberal, or unknown.[1]

We treat each brief as a single document and tokenized the brief contents based on word boundaries (i.e., the `\b` anchor in regular expressions). Punctuation tokens are ignored and word tokens are lowercased.

A.3 Features

We implemented the feature extraction and classification pipeline entirely in scikit-learn (Pedregosa et al. 2011). For features, we extracted just the unigram and bigrams and used feature hashing to vectorize our features (Ganchev and Dredze 2008). The size of our feature vector is set to 2^{20}.

We used scikit's implementation of logistic regression using the stochastic average gradient descent solver (Schmidt et al. 2013) for classification.

A.4 Experiments

We evaluated classification performance using 5-fold cross validation and performed grid search over the following hyperparameter configurations:

1. Regularizer type: ℓ_1 and ℓ_2 penalty

[1] In our experiments, we ignored briefs whose ideological direction is labeled "Unknown".

2. Regularization constant[2]: $C \in \{2^{10}, 2^{12}, 2^{14}, 2^{16}, 2^{18}, 2^{20}, 2^{22}, 2^{24}, 2^{26}, 2^{28}\}$

Figure A.1 presents the classification performance over the different hyperparameter configurations. The best cross validation accuracy of 0.715 was obtained with ℓ_1-penalty and $C = 2^{24}$. ℓ_1-penalty tend to do better than ℓ_2-penalty with our features on our dataset.

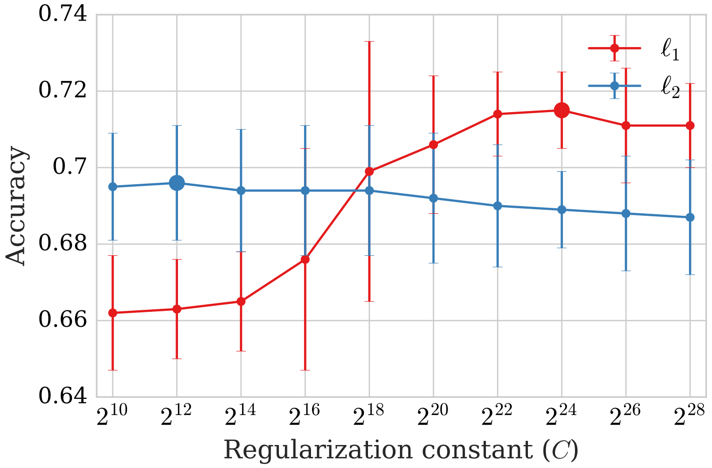


Figure A.1: Classification accuracy over different regularization constants. Smaller C values denote stronger regularization and returns sparser models.

A.5 Conclusion

We implemented a simple linear classifier for political ideology of briefs in SCOTUS. This classifier can potentially be used to predict the political leanings of briefs in existing cases, and be used as a way to ground IP models for vote prediction. From our experiments, we see that classifying briefs by their ideology is not a trivial task. In future work, we will incorporate complex linguistic features (Moschitti and Basili 2004) and ideological cue words (Chapter 2). Additionally, we can experiment with non-linear approaches such as neural network models (Goldberg 2015), which have been shown to do well in text

[2] The regularization constant is inversely proportionate to the penalty weight, λ, i.e., $C = \frac{1}{\lambda}$. Smaller C values denote stronger regularization and returns sparser models.

categorization tasks.

Appendix B

Topics for Model A (Amici IP)

Table B.1 lists the topics and top phrases estimated from our dataset using LDA with 30 topics.

#	Topic	Top phrases
1	Criminal procedure (1)	reasonable doubt, supervised release, grand jury, prior conviction, plea agreement, controlled substance, guilty plea, double jeopardy clause, sixth amendment, jury trial
2	Employment	erisa plan, plan administrator, employee benefit plan, insurance company, pension plan, health care, plan participant, individual mandate, fiduciary duty, health insurance
3	Due process	due process clause, equal protection clause, fundamental right, domestic violence, equal protection, state interest, d e, assisted suicide, controlled substance, rational basis
4	Indians	m r, m s, indian tribe, tribal court, indian country, fifth amendment, miranda warning, indian affair, vice president, tribal member

5	Economic activity	attorney fee, limitation period, hobbs act, security law, rule 10b, actual damage, racketeering activity, fiduciary duty, loss causation, security exchange act
6	Bankruptcy law	bankruptcy court, bankruptcy code, 1996 act, state commission, telecommunication service, network element, eighth circuit, new entrant, pole attachment, communication act
7	Voting rights	voting right, minority voter, j app, voting right act, covered jurisdiction, fifteenth amendment, redistricting plan, political process, political subdivision, minority group
8	First amendment	first amendment right, commercial speech, strict scrutiny, cable operator, free speech, first amendment protection, protected speech, child pornography, government interest, public forum
9	Taxation	interstate commerce, commerce clause, state tax, tax court, gross income, internal revenue code, income tax, dormant commerce clause, state taxation, sale tax
10	Amicus briefs	national association, amicus brief, vast majority, brief amicus curia, large number, wide range, recent year, public policy, wide variety, washington dc
11	Labor management	north carolina, collective bargaining agreement, confrontation clause, sta t, north platte river, collective bargaining, inland lake, laramie river, labor organization, re v
12	Civil action	class action, class member, injunctive relief, final judgment, federal claim, civil action, preliminary injunction, class certification, civil procedure, subject matter jurisdiction
13	Civil rights	title vii, title vi, civil right act, age discrimination, sexual harassment, old worker, major life activity, reasonable accommodation, prima facie case, disparate impact

14	State sovereign	sovereign immunity, eleventh amendment, state official, absolute immunity, false claim, private party, 42 usc §1983, state sovereign immunity, eleventh amendment immunity, federal employee
15	Federal administrations	federal agency, statutory construction, plain meaning, other provision, statutory text, dc circuit, sub §(a), fiscal year, senate report, agency action
16	Interstate relations	special master, new mexico, prejudgment interest, arkansas river, rt vol, comp act, new jersey, elli island, john martin reservoir, video game
17	Court of Appeals	eleventh circuit, sixth circuit, circuit court, fourth circuit, oral argument, tenth circuit, further proceeding, appeal decision, instant case, defendant motion
18	Fourth amendment	fourth amendment, probable cause, arbitration agreement, police officer, national bank, search warrant, exclusionary rule, arbitration clause, reasonable suspicion, law enforcement officer
19	Eighth amendment	eighth amendment, sex offender, prison official, facto clause, copyright act, copyright owner, unusual punishment, public domain, liberty interest, public safety
20	International law	international law, foreign state, vienna convention, human right, foreign country, foreign government, jones act, united kingdom, native hawaiian, foreign nation
21	Equal protection clause	peremptory challenge, law school, equal protection clause, strict scrutiny, high education, racial discrimination, prima facie case, school district, consent decree, compelling interest
22	Commerce clause	interstate commerce, commerce clause, local government, political subdivision, state regulation, supremacy clause, federal regulation, tobacco product, tenth amendment, federal fund

23	Immigration law	judicial review, immigration law, final order, removal proceeding, immigration judge, due process clause, deportation proceeding, administrative remedy, compliance order, time limit
24	Death penalty	death penalty, habeas corpus, reasonable doubt, trial judge, death sentence, ineffective assistance, direct appeal, defense counsel, new rule, mitigating evidence
25	Environmental issues	navigable water, clean water act, colorado river, special master, project act, public land, fill material, water right, point source, lake mead
26	Establishment clause	establishment clause, school district, public school, private school, religious school, ten commandment, boy scout, religious belief, religious organization, free exercise clause
27	Patent law	federal circuit, patent law, prior art, subject matter, expert testimony, lanham act, hazardous substance, patent system, patent act, new drug
28	Antitrust law	antitrust law, sherman act, contr act, market power, postal service, joint venture, natural gas, high price, public utility, interstate commerce
29	Election law	political party, taking clause, private property, property owner, fifth amendment, independent expenditure, federal election, property right, contribution limit, general election
30	Criminal procedure (2)	punitive damage, habeas corpus, second amendment, punitive damage award, enemy combatant, military commission, compensatory damage, state farm, new trial, due process clause

Table B.1: Topics and top-10 phrases estimated from briefs using LDA. We manually annotated each topic with a label.

Appendix C

Justices’ Ideal Points for AmiciIP

The ideal points of justices vary depending on the issues. We present the justices’ ideal points for each of the 30 topics in Figure C.1.

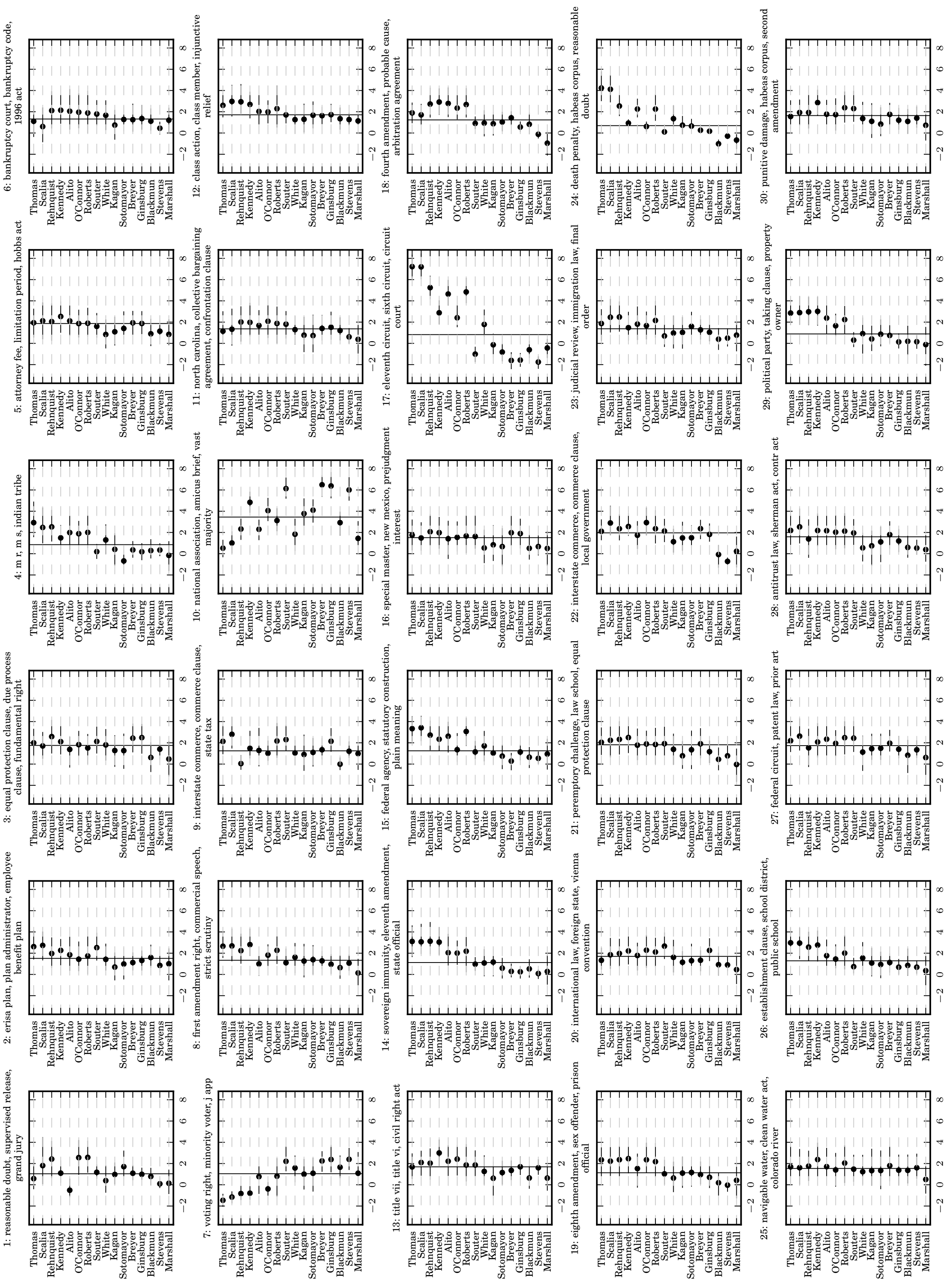


Figure C.1: Justices' ideal points by topics. Solid vertical line denotes median ideal point of the justices.

References

Abelson, Robert P. and J. Douglas Carroll (1965). "Computer Simulation of Individual Belief Systems". In: *American Behavioral Scientist* 8.9, pp. 24–30. URL: http://abs.sagepub.com/content/8/9/24.extract (cit. on p. 34).

Aitchison, John and Shir-Ming Shen (1980). "Logistic-Normal Distributions: Some Properties and Uses". In: *Biometrika* 67.2, pp. 261–272. URL: http://www.jstor.org/stable/2335470 (cit. on p. 45).

Amick, Kimberly (2009). *California Appelate Law Blog: Justice Scalia on Amicus Briefs . . . and Plows*. Online; accessed 13-June-2016. URL: http://www.caappellatelaw.com/2009/01/articles/on-being-a-lawyer/justice-scalia-on-amicus-briefs-and-plows/ (cit. on p. 88).

Anand, Pranav, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor (2011). "Cats Rule and Dogs Drool!: Classifying Stance in Online Debate". In: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. WASSA '11. Portland, OR, USA: Association for Computational Linguistics, pp. 1–9. URL: http://dl.acm.org/citation.cfm?id=2107653.2107654 (cit. on p. 35).

Anderson, Katharine A. (2012). "Specialists and Generalists: Equilibrium Skill Acquisition Decisions in Problem-Solving Populations". In: *Journal of Economic Behavior & Organization* 84.1, pp. 463–473. URL: http://www.sciencedirect.com/science/article/pii/S0167268112001552 (cit. on p. 41).

Andrew, Galen and Jianfeng Gao (2007). “Scalable Training of L_1-regularized Log-linear Models”. In: *Proceedings of the 24th International Conference on Machine Learning*. ICML ’07. Corvalis, OR, USA: ACM, pp. 33–40. URL: `http://dl.acm.org/citation.cfm?id=1273501` (cit. on p. 14).

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen (2014a). “Gender identity and lexical variation in social media”. In: *Journal of Sociolinguistics* 18.2, pp. 135–160. DOI: `10.1111/josl.12080` (cit. on p. 2).

Bamman, David and Noah A. Smith (2015). “Open Extraction of Fine-Grained Political Statements”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. EMNLP ’15. Lisbon, Portugal: Association for Computational Linguistics, pp. 76–85. URL: `https://aclweb.org/anthology/D/D15/D15-1008` (cit. on p. 102).

Bamman, David, Ted Underwood, and Noah A. Smith (2014b). “A Bayesian Mixed Effects Model of Literary Character”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 370–379. URL: `http://www.aclweb.org/anthology/P/P14/P14-1035` (cit. on p. 26).

Beal, Matthew J. (2003). “Variational Algorithms for Approximate Bayesian Inference”. PhD thesis. London, UK: Gatsby Computational Neuroscience Unit, University College London. URL: `http://www.cse.buffalo.edu/faculty/mbeal/thesis/index.html` (cit. on p. 14).

Berry, Steven, James Levinsohn, and Ariel Pakes (1995). “Automobile Prices in Market Equilibrium”. In: *Econometrica* 63.4, pp. 841–890. URL: `http://www.jstor.org/stable/2171802` (cit. on p. 98).

Black, Duncan (1948). “On the Rationale of Group Decision-making”. In: *Journal of Political Economy* 56.1, pp. 23–34. URL: `http://www.jstor.org/stable/1825026` (cit. on p. 9).

Blei, David M. and John D. Lafferty (2007). “A Correlated Topic Model of Science”. In: *The Annals of Applied Statistics*, pp. 17–35. URL: http://www.jstor.org/stable/4537420 (cit. on p. 42).

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003a). “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3, pp. 993–1022. URL: http://dl.acm.org/citation.cfm?id=944919.944937 (cit. on p. 42).

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003b). “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3, pp. 993–1022. URL: http://dl.acm.org/citation.cfm?id=944919.944937 (cit. on p. 64).

Bottou, Léon, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson (2013). “Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising”. In: *Journal of Machine Learning Research* 14, pp. 3207–3260. URL: http://jmlr.org/papers/v14/bottou13a.html (cit. on p. 92).

Box, George E. P. (1979). “Robustness in the Strategy of Scientific Model Building”. In: *Robustness in Statistics*. Ed. by Robert L. Launer and Graham N. Wilkinson. Academic Press, pp. 201–236. ISBN: 978-0-12-438150-6. URL: http://www.sciencedirect.com/science/article/pii/B9780124381506500182 (cit. on p. 1).

Brank, Janez and Jure Leskovec (2003). “The Download Estimation Task on KDD Cup 2003”. In: *SIGKDD Explorations Newsletter* 5.2, pp. 160–162. URL: http://doi.acm.org/10.1145/980972.980997 (cit. on p. 56).

Breiman, Leo (2001). “Random Forests”. In: *Machine Learning* 45.1, pp. 5–32. URL: http://link.springer.com/article/10.1023%2FA%3A1010933404324 (cit. on p. 85).

Brin, Sergey and Lawrence Page (1998). “The Anatomy of a Large-scale Hypertextual Web Search Engine”. In: *Computer Networks and ISDN Systems* 30.1, pp. 107–117. URL: http://dl.acm.org/citation.cfm?id=297827 (cit. on p. 23).

Briscoe, Ted and John Carroll (2006). “Evaluating the Accuracy of an Unlexicalized Statistical Parser on the PARC DepBank”. In: *Proceedings of the COLING/ACL 2006 Main*

Conference Poster Sessions. COLING-ACL ’06. Sydney, Australia: Association for Computational Linguistics, pp. 41–48. URL: http://dl.acm.org/citation.cfm?id=1273073.1273079 (cit. on p. 56).

Briscoe, Ted, John Carroll, and Rebecca Watson (2006). “The Second Release of the RASP System”. In: *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. COLING-ACL ’06. Sydney, Australia: Association for Computational Linguistics, pp. 77–80. URL: http://dl.acm.org/citation.cfm?id=1225423 (cit. on p. 56).

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer (1993). “The Mathematics of Statistical Machine Translation: Parameter Estimation”. In: *Computational Linguistics* 19.2, pp. 263–311. URL: http://dl.acm.org/citation.cfm?id=972470.972474 (cit. on p. 15).

Caldeira, Gregory A. and John R. Wright (1988). “Organized Interests and Agenda Setting in the U.S. Supreme Court”. In: *American Political Science Review* 82 (04), pp. 1109–1127. URL: http://journals.cambridge.org/article_S0003055400196352 (cit. on p. 97).

Cano-Basave, Amparo E. and Yulan He (2016). “A Study of the Impact of Persuasive Argumentation in Political Debates”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA, USA: Association for Computational Linguistics, pp. 1405–1413. URL: http://www.aclweb.org/anthology/N16-1166 (cit. on p. 102).

Carbonell, Jaime G. (1978). “POLITICS: Automated Ideological Reasoning”. In: *Cognitive Science* 2.1, pp. 27–51. URL: http://www.sciencedirect.com/science/article/pii/S0364021378800603 (cit. on p. 34).

Card, Dallas, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith (2015). “The Media Frames Corpus: Annotations of Frames Across Issues”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. ACL ’15. Beijing, China: Association for Computational Linguistics, pp. 438–444. URL: http://www.aclweb.org/anthology/P15-2072 (cit. on p. 102).

Chang, Ming-Wei, Lev Ratinov, and Dan Roth (2007). “Guiding Semi-Supervision with Constraint-Driven Learning”. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic, pp. 280–287. URL: http://www.aclweb.org/anthology/P07-1036 (cit. on p. 78).

Charteris-Black, Jonathan (2005). *Politicians and Rhetoric: The Persuasive Power of Metaphor*. Palgrave Macmillan UK. URL: http://www.palgrave.com/us/book/9780230251649 (cit. on p. 8).

Clinton, Joshua, Simon Jackman, and Douglas Rivers (2004). “The Statistical Analysis of Roll Call Data”. In: *American Political Science Review* 98.2, pp. 355–370. URL: http://dx.doi.org/10.1017/S0003055404001194 (cit. on pp. 19, 35, 97).

Cohn, David A. and Thomas Hofmann (2001). “The Missing Link – A Probabilistic Model of Document Content and Hypertext Connectivity”. In: *Advances in Neural Information Processing Systems 13*. Ed. by Todd K. Leen, Thomas G. Dietterich, and Volker Tresp. MIT Press, pp. 430–436. URL: http://papers.nips.cc/paper/1846-the-missing-link-a-probabilistic-model-of-document-content-and-hypertext-connectivity (cit. on p. 57).

Collins, Michael (2002). “Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms”. In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*. EMNLP ’02. Philadelphia, PA, USA: Association for Computational Linguistics, pp. 1–8. URL: http://dl.acm.org/citation.cfm?id=1118694 (cit. on p. 80).

Collins, Paul M. (2007). “Lobbyists before the U.S. Supreme Court: Investigating the Influence of Amicus Curiae Briefs”. In: *Political Research Quarterly* 60.1, pp. 55–70. URL: http://prq.sagepub.com/content/60/1/55.abstract (cit. on p. 97).

Collins, Paul M (2008). *Friends of the Supreme Court: Interest Groups and Judicial Decision Making*. Oxford University Press. URL: http://www.psci.unt.edu/~pmcollins/FOSC.htm (cit. on pp. 61, 64, 97).

Collins, Paul M., Pamela C. Corley, and Jesse Hamner (2015). “The Influence of Amicus Curiae Briefs on U.S. Supreme Court Opinion Content”. In: *Law & Society Review* 49.4, pp. 917–944. URL: http://onlinelibrary.wiley.com/doi/10.1111/lasr.12166/abstract (cit. on p. 71).

Collins-Thompson, Kevyn and Jamie Callan (2005). “Query Expansion Using Random Walk Models”. In: *Proceedings of CIKM*. CIKM ’05. Bremen, Germany: ACM, pp. 704–711. URL: http://dl.acm.org/citation.cfm?doid=1099554.1099727 (cit. on p. 23).

Comparato, Scott Alson (2003). *Amici Curiae and Strategic Behavior in State Supreme Courts*. Greenwood Publishing Group (cit. on p. 75).

Conover, Michael D., Bruno Goncalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer (2011). “Predicting the Political Alignment of Twitter Users”. In: *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pp. 192–199. DOI: 10.1109/PASSAT/SocialCom.2011.34 (cit. on p. 2).

Converse, Philip E. (2006). “The Nature of Belief Systems in Mass Publics (1964)”. In: *Critical Review: A Journal of Politics and Society* 18.1–3. Originally published in David E. Apter, ed., Ideology and Its Discontents (New York: The Free Press of Glencoe)., pp. 1–74. URL: http://www.tandfonline.com/doi/abs/10.1080/08913810608443650 (cit. on p. 34).

Cooper, Gregory F. (1990). “The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks”. In: *Artificial Intelligence* 42.2, pp. 393–405. URL: http://www.sciencedirect.com/science/article/pii/000437029090060D (cit. on p. 4).

Corley, Pamela, Paul M Collins, and Jesse Hamner (2013). “The Influence of Amicus Curiae Briefs on U.S. Supreme Court Opinion Content”. In: *APSA 2013 Annual Meeting Paper* (cit. on pp. 65, 97).

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory*. Wiley-Interscience. URL: http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471241954.html (cit. on p. 16).

Cowan, Robin and Dominique Foray (2002). "Evolutionary Economics and the Counterfactual Threat: On the Nature and Role of Counterfactual History as an Empirical Tool in Economics". In: *Journal of Evolutionary Economics* 12.5, pp. 539–562. URL: http://link.springer.com/article/10.1007%2Fs00191-002-0134-8 (cit. on p. 91).

Danescu-Niculescu-Mizil, Cristian, Lillian Lee, Bo Pang, and Jon Kleinberg (2012). "Echoes of Power: Language Effects and Power Differences in Social Interaction". In: *Proceedings of the 21st International Conference on World Wide Web*. WWW '12. Lyon, France, pp. 699–708. URL: http://dl.acm.org/citation.cfm?id=2187931 (cit. on p. 98).

Deirmeier, Daniel, Jean-Francois Godbout, Bei Yu, and Stefan Kaufmann (2012). "Language and Ideology in Congress". In: *British Journal of Political Science* 42.1, pp. 31–55. URL: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=8444227 (cit. on p. 8).

Ditzler, Megan Ann (2011). "Language Overlap Between Solicitor General Amicus Curiae and Supreme Court Majority Opinions: An Analysis". MA thesis. Southern Illinois University Carbondale. URL: http://opensiuc.lib.siu.edu/theses/651/ (cit. on p. 71).

Downs, Anthony (1957). "An Economic Theory of Political Action in a Democracy". In: *Journal of Political Economy* 65.2, pp. 135–150. URL: http://www.jstor.org/stable/1827369 (cit. on p. 9).

Eemeren, Frans H. van and Rob Grootendorst (2004). *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. New York, NY, USA: Cambridge University Press. ISBN: 978-0-51161-638-9. URL: http://ebooks.cambridge.org/ebook.jsf?bid=CBO9780511616389 (cit. on p. 102).

Efron, Miles (2004). “Cultural Orientation: Classifying Subjective Documents by Cociation Analysis”. In: *AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*. URL: http://www.aaai.org/Library/Symposia/Fall/2004/fs04-07-007.php (cit. on p. 35).

Eisenstein, Jacob, Amr Ahmed, and Eric P Xing (2011). “Sparse Additive Generative Models Of Text”. In: *Proceedings of the 28th International Conference on Machine Learning*. Ed. by Lise Getoor and Tobias Scheffer. ICML ’11. New York, NY, USA: ACM. URL: http://repository.cmu.edu/machine_learning/210/ (cit. on pp. 13, 14).

Eisenstein, Jacob, Brendan O’Connor, Noah A. Smith, and Eric P. Xing (2010). “A Latent Variable Model for Geographic Lexical Variation”. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP ’10. Cambridge, MA, USA: Association for Computational Linguistics, pp. 1277–1287. URL: http://dl.acm.org/citation.cfm?id=1870658.1870782 (cit. on p. 2).

Entman, Robert M (1993). “Framing: Toward Clarification of a Fractured Paradigm”. In: *Journal of Communication* 43.4, pp. 51–58. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1460-2466.1993.tb01304.x/abstract (cit. on p. 102).

Erosheva, Elena, Stephen Fienberg, and John Lafferty (2004). “Mixed-membership Models of Scientific Publications”. In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5220–5227. URL: http://www.pnas.org/content/101/suppl_1/5220.abstract (cit. on p. 57).

Fader, Anthony, Dragomir R. Radev, Michael H. Crespin, Burt L. Monroe, Kevin M. Quinn, and Michael Colaresi (2007). “MavenRank: Identifying Influential Members of the US Senate Using Lexical Centrality”. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. EMNLP ’07. Prague, Czech Republic: Association for Computational Linguistics, pp. 658–666. URL: http://www.aclweb.org/anthology/D/D07/D07-1069 (cit. on pp. 8, 35).

Fearon, James D. (1991). "Counterfactuals and Hypothesis Testing in Political Science". In: *World Politics* 43.2, pp. 169–195. URL: http://www.jstor.org/stable/2010470 (cit. on p. 91).

Feldman, Adam (2016a). "A Brief Assessment of Supreme Court Opinion Language, 1946–2013". In: *Mississippi Law Journal* 85. Forthcoming. URL: http://ssrn.com/abstract=2574451 (cit. on p. 71).

Feldman, Adam (2016b). "All Copying Is Not Created Equal: Examining Supreme Court Opinions' Borrowed Language". In: *Journal of Appellate Practice and Process* 17. Forthcoming. URL: http://ssrn.com/abstract=2679625 (cit. on p. 71).

Figueiredo, Mario A. T. (2003). "Adaptive Sparseness for Supervised Learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.9, pp. 1150–1159. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1227989 (cit. on p. 14).

Fillenbaum, Samuel (1974). "Information Amplified: Memory for Counterfactual Conditionals". In: *Journal of Experimental Psychology* 102.1, pp. 44–49. URL: http://psycnet.apa.org/journals/xge/102/1/44/ (cit. on p. 91).

Fillmore, Charles J. (1982). "Frame Semantics". In: *Linguistics in the Morning Calm*. Seoul, South Korea: Hanshin Publishing Co., pp. 111–137 (cit. on p. 102).

Fortuna, Blaz, Carolina Galleguillos, and Nello Cristianini (2009). "Detecting the Bias in Media with Statistical Learning Methods". In: *Text Mining: Classification, Clustering, and Applications*. Ed. by Ashok N. Srivastava and Mehran Sahami. Chapman & Hall. Chap. 2, pp. 27–50. URL: http://www.crcnetbase.com/doi/abs/10.1201/9781420059458.ch2 (cit. on p. 35).

Fox, Jean-Paul (2010). *Bayesian Item Response Modeling: Theory and Applications*. Statistics for Social and Behavioral Sciences. Springer. URL: http://www.springer.com/us/book/9781441907417 (cit. on p. 63).

Franze, Anthony J. and R. Reeves Anderson (2015). "Record Breaking Term for Amicus Curiae in Supreme Court Reflects New Norm". In: *National Law Journal* Supreme Court

Brief. URL: http://www.nationallawjournal.com/supremecourtbrief/id=1202735095655/ (cit. on pp. 61, 88, 90).

Ganchev, Kuzman and Mark Dredze (2008). “Small Statistical Models by Random Feature Mixing”. In: *Proceedings of the ACL-08: HLT Workshop on Mobile Language Processing*. Columbus, OH, USA: Association for Computational Linguistics, pp. 19–20. URL: http://www.aclweb.org/anthology/W/W08/W08-0804 (cit. on p. 107).

Ganchev, Kuzman, João Graça, Jennifer Gillenwater, and Ben Taskar (2010). “Posterior Regularization for Structured Latent Variable Models”. In: *Journal of Machine Learning Research* 11, pp. 2001–2049. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1756006.1859918 (cit. on p. 78).

Gelman, Andrew (2013). “Preregistration of Studies and Mock Reports”. In: *Political Analysis* 21.1, pp. 40–41. URL: http://pan.oxfordjournals.org/content/21/1/40.short (cit. on p. 26).

Gelman, Andrew and Jennifer Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. New York, NY, USA: Cambridge University Press. URL: http://www.stat.columbia.edu/~gelman/arm/ (cit. on p. 92).

Geman, Stuart and Donald Geman (1984). “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6, pp. 721–741. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4767596 (cit. on p. 4).

Gentzkow, Matthew and Jesse M. Shapiro (2005). *Media Bias and Reputation*. Working Paper 11664. National Bureau of Economic Research. URL: http://www.nber.org/papers/w11664 (cit. on p. 35).

Gentzkow, Matthew and Jesse M. Shapiro (2010). “What Drives Media Slant? Evidence From U.S. Daily Newspapers”. In: *Econometrica* 78.1, pp. 35–71. URL: http://onlinelibrary.wiley.com/doi/10.3982/ECTA7195/abstract (cit. on pp. 1, 35, 98).

Gerrish, Sean and David Blei (2011). “Predicting Legislative Roll Calls from Text”. In: *Proceedings of the 28th International Conference on Machine Learning*. ICML ’11. Bellevue, WA, USA: ACM, pp. 489–496. URL: http://www.icml-2011.org/papers/333_icmlpaper.pdf (cit. on pp. 2, 8, 35, 84).

Gerrish, Sean and David M. Blei (2010). “A Language-based Approach to Measuring Scholarly Impact”. In: *Proceedings of the 27th International Conference on Machine Learning*. Ed. by Johannes Fürnkranz and Thorsten Joachims. ICML ’10. Haifa, Israel: Omnipress, pp. 375–382. URL: http://www.icml2010.org/papers/384.pdf (cit. on p. 56).

Gerrish, Sean and David M. Blei (2012a). “How They Vote: Issue-Adjusted Models of Legislative Behavior”. In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., pp. 2753–2761. URL: https://papers.nips.cc/paper/4715-how-they-vote-issue-adjusted-models-of-legislative-behavior (cit. on p. 35).

Gerrish, Sean and David M. Blei (2012b). “How They Vote: Issue-Adjusted Models of Legislative Behavior”. In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger. NIPS ’12. Curran Associates, Inc, pp. 2753–2761. URL: https://papers.nips.cc/paper/4715-how-they-vote-issue-adjusted-models-of-legislative-behavior (cit. on pp. 64, 97).

Goldberg, Yoav (2015). “A Primer on Neural Network Models for Natural Language Processing”. In: *ArXiv e-prints*. eprint: 1510.00726 (cs.CL). URL: http://arxiv.org/abs/1510.00726 (cit. on p. 108).

Goldwasser, Dan and Hal Daumé III (2014). ““I Object!” Modeling Latent Pragmatic Effects in Courtroom Dialogues”. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. EACL ’14. Gothenburg, Sweden: Association for Computational Linguistics, pp. 655–663. URL: http://www.aclweb.org/anthology/E14-1069 (cit. on p. 98).

Goodman, Nelson (1947). “The Problem of Counterfactual Conditionals”. In: *Journal of Philosophy* 44.5, pp. 113–128. URL: https://www.jstor.org/stable/2019988 (cit. on p. 91).

Greene, Stephan and Philip Resnik (2009). “More Than Words: Syntactic Packaging and Implicit Sentiment”. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL ’09. Boulder, CO, USA: Association for Computational Linguistics, pp. 503–511. URL: http://dl.acm.org/citation.cfm?id=1620754.1620827 (cit. on p. 35).

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*. EMNLP ’08. Honolulu, HI, USA: Association for Computational Linguistics, pp. 363–371. URL: http://dl.acm.org/citation.cfm?id=1613715.1613763 (cit. on p. 56).

Hansford, Thomas G. (2004). “Information Provision, Organizational Constraints, and the Decision to Submit an Amicus Curiae Brief in a U.S. Supreme Court Case”. In: *Political Research Quarterly* 57.2, pp. 219–230. URL: https://www.jstor.org/stable/3219866 (cit. on p. 98).

Hart, Roderick P. (2009). *Campaign Talk: Why Elections Are Good for Us*. Princeton University Press. URL: http://press.princeton.edu/titles/6797.html (cit. on p. 8).

Hart, Roderick P., Jay P. Childers, and Colene J. Lind (2013). *Political Tone: How Leaders Talk and Why*. University of Chicago Press. URL: http://press.uchicago.edu/ucp/books/book/chicago/P/bo15233236.html (cit. on p. 8).

Hastings, W. K. (1970). “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. In: *Biometrika* 57.1, pp. 97–109. URL: http://www.jstor.org/stable/2334940 (cit. on pp. 5, 82).

Heckman, James J. and James M. Snyder Jr. (1996). *Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Leg-*

islators. Working Paper 5785. National Bureau of Economic Research. URL: http://www.nber.org/papers/w5785 (cit. on p. 97).

Hellinger, Ernst D. (1909). “Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen”. In: *Journal Für Die Reine Und Angewandte Mathematik (Crelle’s Journal)* 1909.136, pp. 210–271. URL: http://dx.doi.org/10.1515/crll.1909.136.210 (cit. on p. 76).

Hillard, Dustin, Stephen Purpura, and John Wilkerson (2008). “Computer-assisted Topic Classification for Mixed-methods Social Science Research”. In: *Journal of Information Technology & Politics* 4.4, pp. 31–46. URL: http://www.tandfonline.com/doi/abs/10.1080/19331680801975367 (cit. on p. 9).

Hoffman, Matthew, Francis R. Bach, and David M. Blei (2010). “Online Learning for Latent Dirichlet Allocation”. In: *Advances in Neural Information Processing Systems 23*. Ed. by John D. Lafferty, Chris K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta. NIPS ’10. Vancouver, BC, Canada: Curran Associates, Inc., pp. 856–864. URL: http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf (cit. on pp. 82, 87).

Hofmann, Thomas (1999). In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’99. Berkeley, CA, USA: ACM, pp. 50–57. ISBN: 1-58113-096-1. URL: http://doi.acm.org/10.1145/312624.312649 (cit. on p. 42).

Holmes, David I. and Richard S. Forsyth (1995). “The Federalist Revisited: New Directions in Authorship Attribution”. In: *Literary and Linguistic Computing* 10.2, pp. 111–127. URL: http://llc.oxfordjournals.org/content/10/2/111.abstract (cit. on p. 56).

Hotelling, Harold (1929). “Stability in Competition”. In: *The Economic Journal* 39.153, pp. 41–57. URL: http://www.jstor.org/stable/2224214 (cit. on p. 9).

Hovy, Eduard H. (1990). “Pragmatics and Natural Language Generation”. In: *Artificial Intelligence* 43.2, pp. 153–197. URL: http://www.sciencedirect.com/science/article/pii/000437029090084D (cit. on p. 57).

Jackman, Simon (2001). “Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking”. In: *Political Analysis* 9.3, pp. 227–241. URL: http://pan.oxfordjournals.org/content/9/3/227 (cit. on pp. 35, 63).

Jelveh, Zubin, Bruce Kogut, and Suresh Naidu (2015). “Political Language in Economics”. In: *Columbia Business School Research Paper Series* 14.57. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2535453 (cit. on p. 98).

Ji, Yangfeng, Gholamreza Haffari, and Jacob Eisenstein (2016). “A Latent Variable Recurrent Neural Network for Discourse-Driven Language Models”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA, USA: Association for Computational Linguistics, pp. 332–342. URL: http://www.aclweb.org/anthology/N16-1037 (cit. on p. 103).

Johri, Nikhil, Daniel Ramage, Daniel A. McFarland, and Daniel Jurafsky (2011). “A Study of Academic Collaboration in Computational Linguistics with Latent Mixtures of Authors”. In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. LaTeCH ’11. Portland, OR, USA: Association for Computational Linguistics, pp. 124–132. ISBN: 9781937284046. URL: http://dl.acm.org/citation.cfm?id=2107636.2107652 (cit. on p. 56).

Jordan, Michael I., Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul (1999). “An Introduction to Variational Methods for Graphical Models”. In: *Machine Learning* 37.2, pp. 183–233. URL: http://dx.doi.org/10.1023/A:1007665907178 (cit. on p. 4).

Joshi, Mahesh, Dipanjan Das, Kevin Gimpel, and Noah A. Smith (2010). “Movie Reviews and Revenues: An Experiment in Text Regression”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT ’10. Los Angeles, CA, USA: Association for Computational Linguistics, pp. 293–296. URL: http://dl.acm.org/citation.cfm?id=1857999.1858037 (cit. on p. 2).

Justeson, John S. and Slava M. Katz (1995). “Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text”. In: *Natural Language Engineering* 1 (01), pp. 9–27. URL: http://journals.cambridge.org/article_S1351324900000048 (cit. on pp. 50, 80).

Kaplan, Ron, Stefan Riezler, Tracy H King, John T Maxwell III, Alex Vasserman, and Richard Crouch (2004). “Speed and Accuracy in Shallow and Deep Stochastic Parsing”. In: *Proceedings of the 2nd Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. HLT-NAACL ’04. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 97–104 (cit. on p. 56).

Kataria, Saurabh, Prasenjit Mitra, and Sumit Bhatia (2010). “Utilizing Context in Generative Bayesian Models for Linked Corpus”. In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI ’10. Atlanta, GA, USA: The AAAI Press, pp. 1340–1345. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1883 (cit. on p. 57).

Katz, Daniel Martin, Michael James Bommarito, and Josh Blackman (2014). *Predicting the Behavior of the Supreme Court of the United States: A General Approach*. URL: http://ssrn.com/abstract=2463244 (cit. on p. 85).

Kearney, Joseph D and Thomas W Merrill (2000). “The Influence of Amicus Curiae Briefs on the Supreme Court”. In: *University of Pennsylvania Law Review*, pp. 743–855. URL: http://scholarship.law.upenn.edu/penn_law_review/vol148/iss3/2/ (cit. on pp. 61, 97).

Kimble, Joseph, ed. (2010). *The Scribes Journal of Legal Writing: Interviews with United States Supreme Court Justices*. Vol. 13. Lansing, MI, USA: Scribes – The American Society of Legal Writers. URL: http://legaltimes.typepad.com/files/garner-transcripts-1.pdf (cit. on p. 60).

Kingma, Diederik P. and Max Welling (2014). “Auto-Encoding Variational Bayes”. In: *Proceedings of International Conference on Learning Representations 2014*. ICLR ’14.

Banff, AB, Canada. URL: https://arxiv.org/abs/1312.6114/ (cit. on p. 103).

Kogan, Shimon, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith (2009). “Predicting Risk from Financial Reports with Regression”. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL ’09. Boulder, CO, USA: Association for Computational Linguistics, pp. 272–280. URL: http://dl.acm.org/citation.cfm?id=1620754.1620794 (cit. on p. 3).

Koller, Daphne and Nir Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press. ISBN: 9780262013192 (cit. on p. 4).

Koo, Terry, Xavier Carreras, and Michael Collins (2008). “Simple Semi-supervised Dependency Parsing”. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Columbus, OH, USA: Association for Computational Linguistics, pp. 595–603. URL: http://www.aclweb.org/anthology/P/P08/P08-1068 (cit. on p. 80).

Lange, Kenneth and Janet S. Sinsheimer (1993). “Normal/Independent Distributions and Their Applications in Robust Regression”. In: *Journal of Computational and Graphical Statistics* 2.2, pp. 175–198. URL: http://www.jstor.org/stable/1390698 (cit. on p. 14).

Lauderdale, Benjamin E. and Tom S. Clark (2014). “Scaling Politically Meaningful Dimensions Using Texts and Votes”. In: *American Journal of Political Science* 58.3, pp. 754–771. URL: http://dx.doi.org/10.1111/ajps.12085 (cit. on pp. 61, 64, 66, 71, 81, 83, 97).

Laver, Michael, Kenneth Benoit, and John Garry (2003). “Extracting Policy Positions from Political Texts Using Words as Data”. In: *The American Political Science Review* 97.2, pp. 311–331. URL: http://www.jstor.org/stable/3118211 (cit. on pp. 10, 35).

Le, Quoc and Tomas Mikolov (2014). “Distributed Representations of Sentences and Documents”. In: *Proceedings of the 31st International Conference on Machine Learning*.

Ed. by Tony Jebara and Eric P. Xing. ICML ’14. Beijing, China: JMLR, pp. 1188–1196. URL: http://jmlr.org/proceedings/papers/v32/le14.pdf (cit. on p. 103).

Li, William, Pablo Azar, David Larochelle, Phil Hill, James Cox, Robert C. Berwick, and Andrew W. Lo (2013). “Using Algorithmic Attribution Techniques to Determine Authorship in Unsigned Judicial Opinions”. In: *Stanford Technology Law Review*, pp. 503–534. URL: https://journals.law.stanford.edu/stanford-technology-law-review/online/using-algorithmic-attribution-techniques-determine-authorship-unsigned-judicial-opinions (cit. on p. 98).

Lin, Wei-Hao, Eric Xing, and Alexander Hauptmann (2008). “A Joint Topic and Perspective Model for Ideological Discourse”. In: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases - Part II*. ECML PKDD ’08. Antwerp, Belgium: Springer-Verlag, pp. 17–32. URL: http://dl.acm.org/citation.cfm?id=1432002 (cit. on p. 35).

Liu, Dong C. and Jorge Nocedal (1989). “On the Limited-memory BFGS Method for Large Scale Optimization”. In: *Mathematical Programming* 45.1-3, pp. 503–528. URL: http://link.springer.com/article/10.1007%2FBF01589116 (cit. on p. 48).

Liu, Yan, Alexandru Niculescu-Mizil, and Wojciech Gryc (2009). “Topic-link LDA: Joint Models of Topic and Author Community”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML ’09. Montreal, QC, Canada: ACM, pp. 665–672. URL: http://doi.acm.org/10.1145/1553374.1553460 (cit. on p. 57).

Liu, Zhiyuan, Yuzhou Zhang, Edward Y. Chang, and Maosong Sun (2011). “PLDA+: Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing”. In: *ACM Transactions on Intelligent Systems and Technology* 2.3. Software available at http://code.google.com/p/plda., 26:1–26:18. URL: http://doi.acm.org/10.1145/1961189.1961198 (cit. on p. 81).

Londregan, John (1999). “Estimating Legislators’ Preferred Points”. In: *Political Analysis* 8.1, pp. 35–56. URL: http://pan.oxfordjournals.org/content/8/1/35.abstract (cit. on p. 35).

Luntz, Frank I. (2007). *Words That Work: It’s Not What You Say, It’s What People Hear*. Hyperion Books. ISBN: 9781401302597. URL: https://books.google.com/books?id=rTcWRFYFQoYC (cit. on p. 13).

Lynch, Kelly J (2004). “Best Friends – Supreme Court Law Clerks on Effective Amicus Curiae Briefs”. In: *Journal of Law & Politics* 20, p. 33. URL: https://litigation-essentials.lexisnexis.com/webcd/app?action=DocumentDisplay&crawlid=1&doctype=cite&docid=20+J.+L.+%26+Politics+33&srctype=smi&srcid=3B15&key=40ba1bd2e55ea29c6deb8d6b45a67fae (cit. on pp. 66, 96).

Manning, Christopher D. (2015). “Computational Linguistics and Deep Learning”. In: *Compututational Linguistics* 41.4, pp. 701–707. URL: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00239 (cit. on p. 103).

Martin, Andrew D. and Kevin M. Quinn (2001). “Estimating Latent Structures of Voting for Micro-Committees with Application to the U.S. Supreme Court”. In: *Proceedings of the Annual Meeting of the Midwest Political Science Association*. MPSA ’01. Chicago, IL, USA (cit. on p. 63).

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”. In: *Political Analysis* 10.2, pp. 134–153. URL: http://pan.oxfordjournals.org/content/10/2/134 (cit. on pp. 35, 63, 97).

Mcauliffe, Jon D. and David M. Blei (2008). “Supervised Topic Models”. In: *Advances in Neural Information Processing Systems 20*. Ed. by John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis. Curran Associates, Inc., pp. 121–128. URL: http://papers.nips.cc/paper/3328-supervised-topic-models.pdf (cit. on p. 51).

McCallum, Andrew, Gideon Mann, and Gregory Druck (2007). *Generalized Expectation Criteria*. Technical Report UM-CS-2007-60. Amherst, MA, USA: University of Massachusetts. URL: https://people.cs.umass.edu/~mccallum/papers/ge08note.pdf (cit. on p. 78).

McDonald, Ryan and Fernando Pereira (2006). "Online Learning of Approximate Dependency Parsing Algorithms". In: *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 6. EACL '06, pp. 81–88 (cit. on p. 56).

McFadden, Daniel (1974). "Conditional Logit Analysis of Qualitative Choice Behavior". In: *Frontiers in Econometrics*. Ed. by Paul Zarembka. Academic Press. Chap. 4, pp. 105–142. URL: http://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf (cit. on pp. 40, 46, 79).

McGovern, Amy, Lisa Friedland, Michael Hay, Brian Gallagher, Andrew Fast, Jennifer Neville, and David Jensen (2003). "Exploiting Relational Structure to Understand Publication Patterns in High-energy Physics". In: *SIGKDD Exploration Newsletter* 5.2, pp. 165–172. URL: http://doi.acm.org/10.1145/980972.980999 (cit. on p. 56).

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953). "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092. URL: http://scitation.aip.org/content/aip/journal/jcp/21/6/10.1063/1.1699114 (cit. on p. 5).

Mihalcea, Rada (2005). "Unsupervised Large-vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling". In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. EMNLP '05. Vancouver, BC, Canada: Association for Computational Linguistics, pp. 411–418. URL: http://dl.acm.org/citation.cfm?id=1220627 (cit. on p. 23).

Monogan, James E. (2013). "A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections". In: *Political Analysis* 21.1, pp. 21–37. URL:

http://pan.oxfordjournals.org/content/21/1/21.abstract (cit. on p. 26).

Monroe, Burt L. and Ko Maeda (2004). *Talk's cheap: Text-based Estimation of Rhetorical Ideal-points*. Presented at the Annual Meeting of the Society for Political Methodology (cit. on p. 10).

Moschitti, Alessandro and Roberto Basili (2004). "Complex Linguistic Features for Text Classification: A Comprehensive Study". In: *Advances in Information Retrieval: 26th European Conference on IR Research (ECIR '04), Sunderland, UK*. Ed. by Sharon McDonald and John Tait. Vol. 2997. Lecture Notes in Computer Science. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, pp. 181–196. ISBN: 978-3-540-24752-4. URL: http://link.springer.com/chapter/10.1007%2F978-3-540-24752-4_14 (cit. on p. 108).

Mullen, Tony and Robert Malouf (2006). "A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse". In: *AAAI Symposium on Computational Approaches to Analysing Weblogs*, pp. 159–162. URL: http://www-rohan.sdsu.edu/~malouf/pubs/aaai-politics.pdf (cit. on p. 35).

Nallapati, Ramesh and William W. Cohen (2008). "Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs". In: *Proceedings of the Second International Conference on Weblogs and Social Media*. ICWSM '08. Seattle, WA, USA: The AAAI Press, pp. 84–92. URL: http://www.aaai.org/Papers/ICWSM/2008/ICWSM08-018.pdf (cit. on p. 57).

Palau, Raquel Mochales and Marie-Francine Moens (2009). "Argumentation Mining: The Detection, Classification and Structure of Arguments in Text". In: *Proceedings of the 12th International Conference on Artificial Intelligence and Law*. ICAIL '09. Barcelona, Spain: ACM, pp. 98–107. ISBN: 978-1-60558-597-0. URL: http://dl.acm.org/citation.cfm?doid=1568234.1568246 (cit. on p. 102).

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. 2nd. Cambridge University Press. URL: http://bayes.cs.ucla.edu/BOOK-2K/ (cit. on p. 91).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12. Available at `http://scikit-learn.org/`., pp. 2825–2830 (cit. on pp. 82, 83, 85, 87, 107).

Petrović, Saša, Miles Osborne, and Victor Lavrenko (2011). “RT to Win! Predicting Message Propagation in Twitter”. In: Barcelona, Spain. URL: `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2754/3209` (cit. on p. 2).

Poole, Keith T. and Howard Rosenthal (1985). “A Spatial Model for Legislative Roll Call Analysis”. In: *American Journal of Political Science* 29.2, pp. 357–384. URL: `http://www.jstor.org/stable/2111172` (cit. on pp. 19, 35, 63, 97, 106).

Poole, Keith T. and Howard Rosenthal (2000). *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press (cit. on p. 19).

Prabhakaran, Vinodkumar, Ajita John, and Dorée D. Seligmann (2013). “Who Had the Upper Hand? Ranking Participants of Interactions Based on Their Relative Power”. In: *Proceedings of the 6th International Joint Conference on Natural Language Processing*. IJCNLP ’2013. Nagoya, Japan: Association for Computational Linguistics, pp. 365–373. URL: `http://www.ofuturescholar.com/paperpage?docid=2248383` (cit. on p. 98).

Radev, Dragomir R., Pradeep Muthukrishnan, and Vahed Qazvinian (2013). “The ACL Anthology Network Corpus”. In: *Language Resources and Evaluation* 47.4. Data available at `http://clair.eecs.umich.edu/aan/`, pp. 919–944. URL: `http://link.springer.com/article/10.1007%2Fs10579-012-9211-2#page-1` (cit. on p. 38).

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth (2004). “The Author-Topic Model for Authors and Documents”. In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. UAI ’04. Banff, Canada: AUAI Press, pp. 487–494. ISBN: 0-9749039-0-6. URL: `http://dl.acm.org/citation.cfm?id=1036843.1036902` (cit. on pp. 42, 53, 56, 71, 87).

Sack, Warren (1994). “Actor-role Analysis: Ideology, Point of View, and the News”. MA thesis. Cambridge, MA, USA: Massachusetts Institute of Technology. URL: http://danm.ucsc.edu/~wsack/Writings/wsack-narrative-perspective.pdf (cit. on p. 34).

Schmidt, M., N. Le Roux, and F. Bach (2013). “Minimizing Finite Sums with the Stochastic Average Gradient”. In: *ArXiv e-prints*. eprint: 1309.2388 (math.OC). URL: http://arxiv.org/abs/1309.2388 (cit. on p. 107).

Schneider, Karl-Michael (2005). “Weighted Average Pointwise Mutual Information for Feature Selection in Text Categorization”. In: *Proceedings of the 9th European conference on Principles and Practice of Knowledge Discovery in Databases*. PKDD ’05. Porto, Portugal: Springer-Verlag, pp. 252–263. URL: http://dl.acm.org/citation.cfm?id=2101264 (cit. on p. 16).

Segal, Jeffrey A. and Albert D. Cover (1989). “Ideological Values and the Votes of U.S. Supreme Court Justices”. In: *The American Political Science Review* 83.2, pp. 557–565. URL: http://www.jstor.org/stable/1962405 (cit. on p. 97).

Sim, Yanchuan, Brice D. Acree, Justin H. Gross, and Noah A. Smith (2013). “Measuring Ideological Proportions in Political Speeches”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. EMNLP ’13. Seattle, WA, USA: Association for Computational Linguistics, pp. 91–101. URL: http://www.aclweb.org/anthology/D13-1010 (cit. on p. 9).

Sim, Yanchuan, Bryan Routledge, and Noah A. Smith (2015). “The Utility of Text: The Case of Amicus Briefs and the Supreme Court”. In: *AAAI Conference on Artificial Intelligence*. AAAI ’15. Austin, Texas, USA, pp. 2311–2317. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9361 (cit. on pp. 37, 57, 60, 78, 83, 107).

Slapin, Jonathan B and Sven-Oliver Proksch (2008). “A Scaling Model for Estimating Time-series Party Positions from Texts”. In: *American Journal of Political Science* 52.3, pp. 705–722. URL: http://www.wordfish.org/uploads/1/2/9/8/12985397/slapin_proksch_ajps_2008.pdf (cit. on p. 35).

Somasundaran, Swapna and Janyce Wiebe (2009). "Recognizing Stances in Online Debates". In: *Proceedings of ACL*. ACL '09. Singapore: Association for Computational Linguistics, pp. 226–234. URL: http://dl.acm.org/citation.cfm?id=1687878.1687912 (cit. on p. 35).

Spaeth, Harold J., Sara Benesh, Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh (2015). *2016 Supreme Court Database, Version 2015 Release 03*. URL: http://supremecourtdatabase.org (cit. on pp. 79, 85, 107).

Stamatatos, Efstathios (2009). "A Survey of Modern Authorship Attribution Methods". In: *Journal of the American Society for Information Science and Technology* 60.3, pp. 538–556. URL: http://onlinelibrary.wiley.com/doi/10.1002/asi.21001/abstract (cit. on pp. 2, 56).

Steyvers, Mark and Tom Griffiths (2006). "Probabilistic Topic Models". In: *Latent Semantic Analysis: A Road to Meaning*. Ed. by T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch. Laurence Erlbaum. URL: http://cocosci.berkeley.edu/tom/papers/SteyversGriffiths.pdf (cit. on p. 4).

Swaminathan, Hariharan and Janice A. Gifford (1985). "Bayesian Estimation in the Two-Parameter Logistic Model". In: *Psychometrika* 50.3, pp. 349–364. URL: http://link.springer.com/article/10.1007/BF02294110 (cit. on p. 106).

Tanner, Chris and Eugene Charniak (2015). "A Hybrid Generative/Discriminative Approach To Citation Prediction". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL '15. Denver, CO, USA: Association for Computational Linguistics, pp. 75–83. URL: http://www.aclweb.org/anthology/N15-1008 (cit. on p. 57).

Thomas, Matt, Bo Pang, and Lillian Lee (2006). "Get out the Vote: Determining Support or Opposition from Congressional Floor-debate Transcripts". In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP '06. Sydney, Australia: Association for Computational Linguistics, pp. 327–335. URL: http://dl.acm.org/citation.cfm?id=1610075.1610122 (cit. on pp. 8, 35).

Tierney, Luke (1994). “Markov Chains for Exploring Posterior Distributions”. In: *The Annals of Statistics* 22.4, pp. 1701–1728. URL: http://www.jstor.org/stable/2242477 (cit. on pp. 47, 81, 82).

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer (2003). “Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network”. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. NAACL ’03. Edmonton, AB, Canada: Association for Computational Linguistics, pp. 173–180. URL: http://dl.acm.org/citation.cfm?id=1073478 (cit. on p. 50).

Toutanova, Kristina, Christopher D. Manning, and Andrew Y. Ng (2004). “Learning Random Walk Models for Inducing Word Dependency Distributions”. In: *Proceedings of the 21st International Conference on Machine Learning*. ICML ’04. Banff, AB, Canada: ACM, pp. 103–. URL: http://dl.acm.org/citation.cfm?id=1015442 (cit. on p. 23).

Turian, Joseph, Lev Ratinov, and Yoshua Bengio (2010). “Word Representations: A Simple and General Method for Semi-supervised Learning”. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL ’10. Uppsala, Sweden: Association for Computational Linguistics, pp. 384–394. URL: http://dl.acm.org/citation.cfm?id=1858681.1858721 (cit. on p. 103).

Vogel, Adam, Max Bodoia, Christopher Potts, and Daniel Jurafsky (2013). “Emergence of Gricean Maxims from Multi-Agent Decision Theory”. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACl ’13. Atlanta, GA, USA: Association for Computational Linguistics, pp. 1072–1081. URL: http://www.aclweb.org/anthology/N13-1127 (cit. on p. 57).

Wainwright, Martin J. and Michael I. Jordan (2008). “Graphical Models, Exponential Families, and Variational Inference”. In: *Foundations and Trends in Machine Learning* 1.1-2, pp. 1–305. URL: http://dx.doi.org/10.1561/2200000001 (cit. on p. 4).

Wang, William Yang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar (2012). "Historical Analysis of Legal Opinions with a Sparse Mixed-effects Latent Variable Model". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. ACL '12. Jeju Island, Korea: Association for Computational Linguistics, pp. 740–749. URL: `http://dl.acm.org/citation.cfm?id=2390629` (cit. on p. 98).

Wei, Greg CG and Martin A Tanner (1990). "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms". In: *Journal of the American Statistical Association* 85.411, pp. 699–704. URL: `http://www.jstor.org/stable/2290005` (cit. on p. 47).

Wikipedia (2016). *Procedures of the Supreme Court of the United States*. Online; accessed 10-June-2016. URL: `https://en.wikipedia.org/wiki/Procedures_of_the_Supreme_Court_of_the_United_States` (cit. on p. 71).

Yano, Tae, William W. Cohen, and Noah A. Smith (2009). "Predicting Response to Political Blog Posts with Topic Models". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL '09. Boulder, CO, USA: Association for Computational Linguistics, pp. 477–485. URL: `http://dl.acm.org/citation.cfm?id=1620754.1620824` (cit. on p. 2).

Yano, Tae, Dani Yogatama, and Noah A. Smith (2013). "A Penny for Your Tweets: Campaign Contributions and Capitol Hill Microblogs". In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. Cambridge, MA, USA: The AAAI Press. URL: `http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6084/6337` (cit. on p. 35).

Yogatama, Dani, Manaal Faruqui, Chris Dyer, and Noah A. Smith (2015). "Learning Word Representations with Hierarchical Sparse Coding". In: *Proceedings of The 32nd International Conference on Machine Learning*. Ed. by David Blei and Francis Bach. Vol. 37. ICML '15. Lille, France: JMLR, pp. 87–96. URL: `http://www.jmlr.org/proceedings/papers/v37/yogatama15.pdf`.

Yogatama, Dani, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith (2011). "Predicting a Scientific Community's Response to an Article". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, UK: Association for Computational Linguistics, pp. 594–604. URL: `http://dl.acm.org/citation.cfm?id=2145432.2145501` (cit. on pp. 2, 38, 43, 51, 52, 56).

Zhu, Yaojia, Xiaoran Yan, Lise Getoor, and Cristopher Moore (2013). "Scalable Text and Link Analysis with Mixed-topic Link Models". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. Chicago, IL, USA: ACM, pp. 473–481. URL: `http://doi.acm.org/10.1145/2487575.2487693` (cit. on p. 57).