Thesis Proposal:
Text as Strategic Choice

Yanchuan Sim
Language Technologies Institute
Carnegie Mellon University

ysim@cs.cmu.edu

Last updated: Jun 5, 2015

Thesis committee

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

Abstract

Humans write — lawyers submit briefs, legislators draft bills, scientists publish papers, teenagers tweet, and politicians give speeches which are first written by their speechwriters — for a multitude of purposes. Choosing the right things to say is a complex process and requires significant effort. In this thesis, I propose a framework for text analysis based on the idea that text production is a strategic process dependent on author's social attributes and his beliefs about audience responses. The social attributes of an author deeply influence and bias his language production; while authors are motivated to evoke responses from his audience. Current research has treated the above as disjoint problems. Instead, this thesis proposes to take a decision theoretic approach to jointly model both author's social attributes and responses from text – we treat text as a choice variable and model an individual author's text production as a utility function.

Throughout this proposal, we will present several examples of strategic behavior and how we can model it computationally. We will present a model for identifying strategic behavior amongst candidates during the U.S presidential elections. Following, we use the Supreme Court as a case study to incorporate utility functions into models of text. Lastly, we discuss methods and challenges to characterize authors' strategic behavior in the domains of scientific community and judicial politics.

Contents

1 Introduction

Humans write — lawyers submit briefs, legislators draft bills, scientists publish papers, teenagers tweet, and politicians give speeches which are first written by their speechwriters — for a multitude of purposes. Choosing the right things to say is a complex process and requires significant effort. Lawyers aim to best frame their arguments so as to convince their judges, while teenagers might seek social status among their peers. Therefore, the text we observe is the result of the author's *strategic choice*[1], which reflects his social attributes and motivations.

The *social attributes* of an author deeply influence and bias his language production. Social contexts shape and inform the language that is being used, and inferring these attributes is a research subfield in itself. 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. These attributes influence the authored text; a conservative politician is more likely to write about lowering taxes, whereas a teenager will more likely be seen tweeting about Taylor Swift, than Exxon's recent financial statement.

However, language production is not just influenced by an author's social attributes; authors are motivated to *evoke responses* from his audience. For instance, politicians give speeches to rally his constituents, and journalists report on events that their readers will be interested in. Yano (2011) calls these *actuating texts*. Many models have been built to predict the responses to actuating text, e.g. whether a blog post will be popular (Yano et al., 2009), or a tweet will be reposted (Petrović et al., 2011), predicting roll call votes from bills (Gerrish and Blei, 2011), or if a research paper will be cited (Yogatama et al., 2011), etc.

Current research has treated the above as disjoint problems. Instead, this thesis proposes to take a decision theoretic approach to jointly model both authors' social attributes and responses from actuating text. We treat the text as a choice variable and model text production as a utility function:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \; U(\boldsymbol{\theta}, \text{attributes}, \text{response}) \tag{1.1}$$

where the optimal document, $\boldsymbol{\theta}^*$, is a function of the author's attributes, and the responses he seeks to evoke. Given that we observe $\boldsymbol{\theta}_{\text{obs}}$, we reason that $\boldsymbol{\theta}_{\text{obs}} \approx \boldsymbol{\theta}^*$. In fact, the function U encodes the author's strategy; within the constraints of his attributes (i.e., his language, political orientation, expertise), he produces text that maximizes his desired response (i.e., votes, retweets, 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 opponents), and the lawyer's desired response is having the judge decide the case in his favor. On the other hand, he is constrained by what he can say, for instance, following procedural rules and invoking legal arguments that are relevant to the case, and staying within the allotted time limits. Within these constraints, he has a choice of arguments, which he knows from experience, the different "costs" and "benefits". Hence, the arguments that we observe are the result of his strategic choice, U.

Throughout this thesis, we will present several examples of strategic behavior and how we can model it computationally. In §2, we will present a model for identifying U.S. presidential candidates' strategic behavior from their transcribed speeches. Subsequently in §3, we will consider a model that explicitly treats actuating

[1] In this thesis, we describe strategy as the purposeful decision making act of an author. It does not necessarily entail decisions that depend on that of others (i.e., game theoretic strategies). Our proposed work does not consider game theoretic effects, but it will be an interesting direction for future work.

text in the U.S. Supreme Court as a strategic choice. In §4, we will propose further work on two datasets where we will further explore questions such as: What is the author's strategy? What are his trade-offs, is he trying to maximize his response, at the expense of higher costs? Or is the author just self-interested?

Thesis Statement

In this thesis, we claim that text production is a strategic process dependent on an author's social attributes and his beliefs about audience responses. We develop a set of novel probabilistic models to characterize authors' strategic behaviors from their attributes, text and responses, which allow new inferences that can be drawn about authors' strategic choices from textual evidence. Using these models, we examine two particular domains — politics and the scientific community — and develop the necessary tools to evaluate our hypothesis. We anticipate that our models will be useful in applications such as recommendation systems, and will offer new quantitative perspectives for researchers in other fields (e.g, social sciences) to perform statistical text analysis.

2 Identifying Strategic Behavior

Our prior work, Sim et al. (2013), analyzed the political speeches during U.S. presidential campaigns. It is a general observation, although not empirically tested before us that successful primary candidates ideologically "move to the center" before a general election. During the transition from primary to general elections, more ideologically concentrated voters are being replaced by a set of voters who are more widely dispersed across the ideological spectrum; as a result, candidates will present themselves as more moderate in an effort to capture more votes.

As such, 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.

Our main contribution here is a probabilistic technique for inferring proportions of ideologies expressed by a candidate. The inputs to the model are the cue-lag representation of a speech (example in Fig. 2) and a domain-specific topology relating ideologies to each other. The topology tree (shown in Fig. 1) encodes the closeness of different ideologies and, by extension, the odds of transitioning between them within a speech.

2.1 Building a Cue Lexicon from Contemporary Political Writings

We operationalized ideologies in a novel empirical way, exploiting contemporary political writings published in explicitly ideological books and magazines, whose authors are perceived as representative of one

particular ideology.[2] 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.

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. A political science domain expert who is a co-author of this work manually labeled each element in a collection of 112 books and 10 magazine titles[3] 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 Fig. 1.[4]

In addition to ideology labels, individual chapters within the books were manually tagged with topics that the chapter was about.[5] 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.

Cue Discovery Model. We use the ideological corpus to infer ideological cues: terms that are strongly associated with an ideology. 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).

SAGE parameterizes the language model 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 document 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:

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

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.

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

[3] 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.

[4] 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.

[5] For instance, in Barack Obama's book *The Audacity of Hope*, his chapter titled "Faith" is labeled as RELIGIOUS. There are a total of 61 topics; the full list can be found in the original paper (Sim et al., 2013).

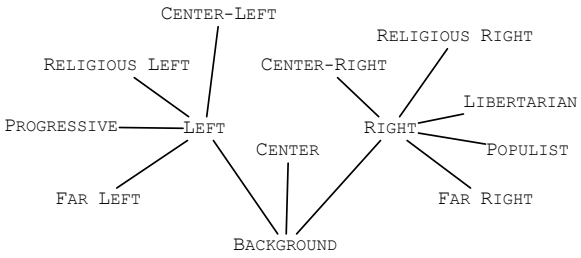


Figure 1: Ideology tree.

- $\boldsymbol{\eta}^{i_c}$, 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 Fig. 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:

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

We use the OWL-QN algorithm to solve the convex objective function (Andrew and Gao, 2007).

2.2 Cue-Lag Ideological Proportions (CLIP) Model

We adopt a Bayesian approach that manages our uncertainty about the cue lexicon $\mathcal{L}$ (§2.1), 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).

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

Party	Pri'08	Gen'08	Pri'12	Gen'12
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 1: 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.

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: Example of the cue-lag representation.

pledged delegates to win the party nomination are "from the primary." Table 1 presents a breakdown of the candidates and speeches in our corpus.

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. Fig. 2 shows an example of our cue-lag representation.

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 (Fig. 1) or BACKGROUND. The ideologies are organized into a tree based on their hierarchical relationships; see Fig. 1. The ideology tree is used in defining the transition distribution in the HMM. 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. In CLIP, the walk corresponds to a *single* transition — the speaker does not emit anything from the states passed through along the path.

Furthermore, 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. For a longer lag, it is more likely for the model to restart the walk from the BACKGROUND state. On the other hand, the emission from a state consists of a cue from the cue lexicon and an integer lag value. To incorporate our prior beliefs based on our ideology-specific cue lexicons, we use an informed prior on the multinomial emission distribution from a given ideological state.

We estimate the posterior distributions of CLIP using MCMC techniques such as Gibbs sampling and slice sampling. From the posterior distribution, we are able to infer the expected amount of time a candidate spends in each ideology. By running the model on subset of a politician's transcribed speeches (i.e., primary elections vs general elections), we can estimate the expected ideological proportions within each stage of the campaign.

2.3 Evaluation Using 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, 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. Before running our model, we identified a set of hypotheses, which we **pre-registered** as expectations. These 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.

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

- HMM: rather than ideology tree transitions, 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 presents a summary of which hypotheses the models' inferences are in accordance with. CLIP is not consistently outperformed by any of the competing baselines.

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

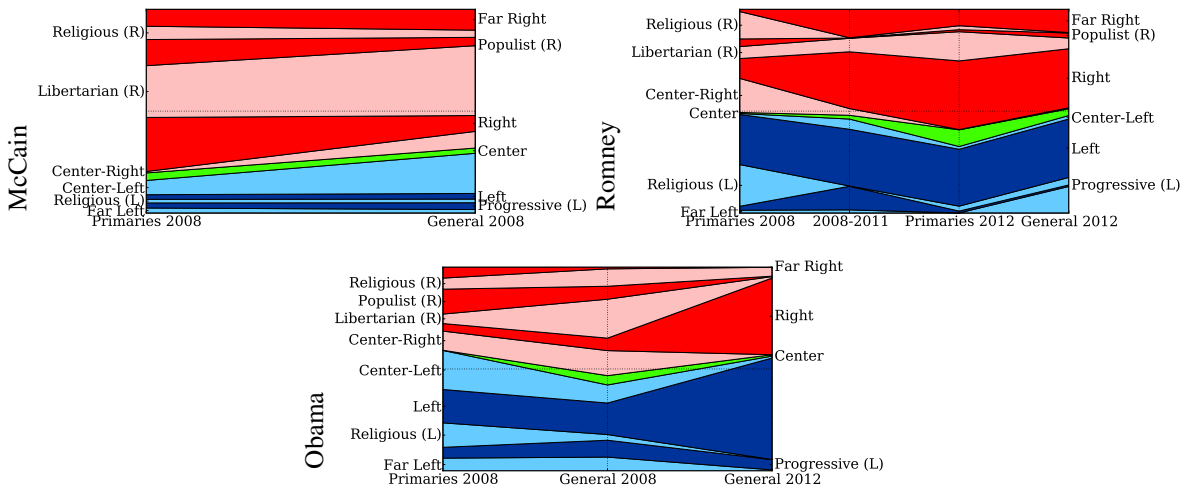


Figure 3: Proportion of time spent in each ideology by McCain, Romney, and Obama during the 2008 and 2012 Presidential election seasons.

Sanity checks (S1–3). 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.

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.

"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 3, 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).

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.

CLIP's inferences on the corpus of political speeches can be browsed at `http://www.ark.cs.cmu.edu/CLIP`. 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 (§2.1), but also of incorrect hypotheses.

2.4 Significance

Although the methodology of pre-registration is not new or uncommon[6], 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). Thus, we anticipate that pre-registration will be used by researchers when dealing with computational models of subjective texts, and also in our future work.

In this section, we built a model to empirically identify evidence of strategic behavior – namely that of presidential candidates ideologically moving towards the center during the elections. However, we did not explicitly take into account the attributes of a candidate or the responses he seek; we will consider models that do so in subsequent sections.

3 Text as Choice

Some pieces of text are written with clear goals in mind. Economists and game theorists use the word *utility* for the concept of satisfying a need or desire, and a huge array of theories and models are available for analyzing how utility-seeking agents behave. Therefore, an author's strategic choice can be modeled explicitly as an utility maximizing action, where his textual output is the choice variable. In our completed work, Sim et al. (2015), we take first steps into explicitly incorporating utility functions into models of text.

[6]Pre-registration is commonly performed in other fields such as psychology and social sciences.

We used the Supreme Court of the United States (SCOTUS) as a case study. 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* (Latin for "friends of the court," hereafter "amici," singular "amicus"), and the textual artifacts they author — known as amicus briefs — reveal explicit attempts to sway justices one way or the other. 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.

SCOTUS Terminology. SCOTUS reviews the decisions of lower courts and (less commonly) resolves disputes between states.[7] In a typical case, 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 (not necessarily allotted for every case) conclude, the justices vote and author one or more opinions. In this work, we relate the votes of justices to merits and amicus briefs.

3.1 Ideal Point Models

We build on a well-established methodology from political science known as *ideal points* for analyzing votes. Specifically, Lauderdale and Clark (2014) combined descriptive text and ideal points in a probabilistic topic model. 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.

Lauderdale and Clark (2014) incorporated text as evidence and infer dimensions of IP that are grounded in "topical" space. They build on latent Dirichlet allocation (Blei et al., 2003), 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

$$p(v_{i,j} = \text{petitioner} \mid \boldsymbol{\psi}_j, \boldsymbol{\theta}_i, a_i, b_i) = \sigma\left(a_i + \boldsymbol{\psi}_j^\top (b_i \boldsymbol{\theta}_i)\right) \tag{3.1}$$

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 Sim et al. (2015), we proposed 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:

$$p(v_{i,j} = \text{petitioner} \mid \psi_j, \boldsymbol{\theta}_i, \boldsymbol{\Delta}_i, a_i, b_i, c_i) = \sigma\left(a_i + \boldsymbol{\psi}_j^\top \left(b_i\boldsymbol{\theta}_i + c_i^{\text{p}}\boldsymbol{\Delta}_i^{\text{p}} + c_i^{\text{r}}\boldsymbol{\Delta}_i^{\text{r}}\right)\right) \tag{3.2}$$

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

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.

3.2 Random Utility Models of Amici Agents

Amici are purposeful decision makers who write briefs hoping to sway votes on a case. 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 $v_1, \ldots, v_9$,

$$u(v_1, v_2, \ldots, v_9) = \textstyle\sum_{j \in \mathcal{J}} \mathbb{I}(v_j = s), \tag{3.3}$$

where $\mathbb{I}$ is the indicator function. This is her **utility**.

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 $\boldsymbol{\theta}$:

$$C(\boldsymbol{\Delta}, \boldsymbol{\theta}) = \tfrac{\xi}{2} \|\boldsymbol{\Delta} - \boldsymbol{\theta}\|_2^2 \tag{3.4}$$

where $\xi > 0$ is a hyperparameter controlling the cost (relative to the vote evaluation). The function captures 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.

Expected Utility. The outcome of the case is uncertain, so the amicus' objective will consider her *expected* utility:

$$\max_{\boldsymbol{\Delta}} \ \mathbb{E}_{\boldsymbol{\Delta}}[u(v_1, \ldots, v_9)] - \tfrac{\xi}{2} \|\boldsymbol{\Delta} - \boldsymbol{\theta}\|_2^2 \tag{3.5}$$

When an amicus writes her brief, we assume that she has knowledge of the justices' IPs, case parameters, and contents of the merits briefs, but ignores other amici.[8] As such, taking linearity of expectations, we can compute the expected utility for an amicus on side s:[9]

$$\max_{\boldsymbol{\Delta}} \ \textstyle\sum_{j \in \mathcal{J}} \sigma\left(a + \boldsymbol{\psi}_j^\top (b\boldsymbol{\theta} + c^s \boldsymbol{\Delta})\right) - \tfrac{\xi}{2} \|\boldsymbol{\Delta} - \boldsymbol{\theta}\|_2^2 \tag{3.6}$$

There are several conceivable ways to incorporate amici's optimization into our estimation of justices' IP. We could maximize the likelihood and impose the constraint on $\boldsymbol{\Delta}$ that solve our expected utility optimization

[8] A model with strategic amici agents (a petitioner amicus choosing brief topics considering a respondent amicus' brief) is a very complicated game theoretical model and, we conjecture, would require a much richer representation of policy and goals.

[9] The first-order conditions for amicus' purposeful maximization wrt $\boldsymbol{\Delta}$ lead to interesting brief writing trade offs, which can be found in supplementary section A of Sim et al. (2015).

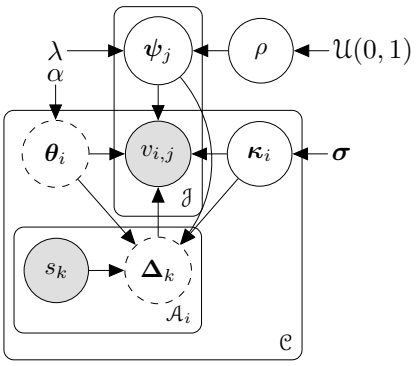


Figure 4: Plate diagram for the amicus random utility model. $\mathcal{J}, \mathcal{C}$ and $\mathcal{A}_i$ are the sets of justices, cases, and amicus briefs (for case i), respectively. $\boldsymbol{\psi}_j$ is the IP for justice j; $\boldsymbol{\kappa}_i$ is the set of case parameters a_i, b_i, c_i^{p} and c_i^{r} for case i; and $\boldsymbol{\alpha}, \boldsymbol{\sigma}, \lambda$, and ρ are hyperparameters.

(either directly or by checking the first order conditions).[10] Or, we can view such (soft) constraints as imposing a prior on $\boldsymbol{\Delta}$:

$$p_{\text{util}}(\boldsymbol{\Delta}) \propto \mathbb{E}_{\boldsymbol{\Delta}}[u(v_1, \dots)] + \xi(1 - \tfrac{1}{2}\|\boldsymbol{\Delta} - \boldsymbol{\theta}\|_2^2) \tag{3.7}$$

where the constant is added so that p_{util} is non-negative. Note, were the utility negative, the amici would have chosen not to write a the brief. 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. Hence, the behavior we observe, (i.e., the amicus' topic mixture proportions), has a likelihood that is proportional to expected utility.

Considering all the amici, we estimate the full likelihood

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{\Delta}, \boldsymbol{\kappa}) \times \left[\textstyle\prod_{k \in \mathcal{A}} p_{\text{util}}(\boldsymbol{\Delta}_k)\right]^{\eta} \tag{3.8}$$

where $\mathcal{L}(\cdot)$ is the likelihood of our amici IP model (Eq. 3.2), and η is a hyperparameter that controls influence of utility on parameter estimation.

Eq. 3.8 resembles *product of experts* model (Hinton, 2002). For the likelihood of votes in a case to be maximized, it is necessary that no individual component—generative story for votes, amicus briefs—assigns a low probability. Accordingly, this results in a principled manner for us to incorporate our assumptions about amici as rational decision makers, each of whom is an "expert" with the goal of nudging latent variables to maximize her own expected utility.

Our random utility model for persuasive text is similar to a classical generative model (plate diagram in Figure 4) and can be estimated using familiar algorithms (i.e., Gibbs sampling (Geman and Geman, 1984), Metropolis-Hastings (Hastings, 1970), etc.). Details of inference and parameter estimation can be found in §4 of Sim et al. (2015).

3.3 Experiments and Analysis

Vote Prediction We evaluate each model's ability to predict how justices would vote on a case out of the training sample. To compute the probability of justices' votes, we first infer the topic mixture proportions

[10] 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.

Model	Accuracy
Logistic regression w/ topics	0.715 ± 0.008
Unanimous	0.714 ± 0.003
Unidimensional IP	0.583 ± 0.037
Issues IP (Lauderdale and Clark, 2014)	0.671 ± 0.008
Amici IP (Eq. 3.2)	0.690 ± 0.021
Random utility IP	$\mathbf{0.742} \pm 0.006$

Table 3: Average pairwise vote partition accuracy (five-fold cross-validation). We have two naïve baselines, (i) where all justices vote unanimously, and (ii) where we trained an ℓ_1-regularized logistic regression classifier for each justice using the concatenated topic proportions of $\boldsymbol{\theta}$ and $\boldsymbol{\Delta}$ as features for each case. Unidimensional IP does not take into account any information about the cases (i.e., $\boldsymbol{\theta}_i$=1).

for the case's merits briefs ($\boldsymbol{\theta}$), and amicus briefs ($\boldsymbol{\Delta}$). Then, given all the justice's IPs $\boldsymbol{\psi}_j$, we find the most likely vote outcome for the case by integrating over the case parameters $\boldsymbol{\kappa}$.

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 would not be able to distinguish the actual side that the justice will favor, but we can identify the most likely partitioning of the justices into two groups.[11] We can then evaluate, for each case, an average pairwise accuracy score,

$$\binom{9}{2}^{-1} \sum_{j,j' \in \mathcal{J}: j \neq j'} \mathbb{I}[\mathbb{I}[\hat{v}_j = \hat{v}_{j'}] = \mathbb{I}[v^*_j = v^*_{j'}]] \tag{3.9}$$

where $\hat{\boldsymbol{v}}$ ($\boldsymbol{v}^*$) are predicted (actual) votes. We performed 5-fold cross validation, and present the vote partition accuracy in Table 3. Our model incorporating both amicus briefs and their utility outperformed baselines on the task, supporting the case that capturing the strategic motives of an amicus brief writer is useful in the context of SCOTUS.

Counterfactual Analysis. Following Pearl (2000), we can query the model and perform counterfactual analyses. As an illustration, we consider *National Federation of Independent Business (NFIB) v. Sebelius (HHS)* (132 S. Ct. 2566), a landmark 2011 case in which the Court upheld Congress's power to enact most provisions of the Affordable Care Act (ACA; "Obamacare").[12]

In the merits briefs, the topics discussed revolve around *interstate commerce* and the *individual mandate*, while there is an interesting disparity in topics between briefs supporting NFIB and HHS.[13] 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 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.

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

[11] Given the partitioning of justices, domain experts should be able to identify the side each group of justices would favor.

[12] 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.

[13] The merits briefs were estimated at 41% and 20% on the *interstate commerce* and the *individual mandate* topics, respectively. NFIB amicus briefs were 15% on *interstate commerce* and 41% on *individual mandate*; these figures switch to 36% and 22% for HHS amicus briefs.

set and attempt to predict the votes under the hypothetical circumstances with the random utility model. Fig. 5 (left) shows the resulting IP of hypothetical situations where no amicus briefs were filed, or when only briefs supporting one side are filed. If no amici filed briefs, 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.

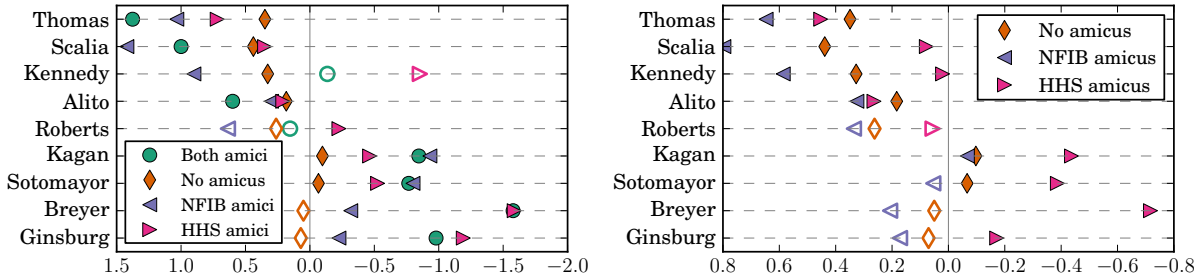


Figure 5: Counterfactual analyses for *NFIB v. Sebelius (HHS)*. *Left*: What if amicus briefs for one side were not filed? *Right*: What if a single amicus files an "optimally" written brief? An IP towards the left (right) indicates higher log-odds of vote favorable to NFIB (HHS). Hollow markers denote that prediction differed from the actual outcome.

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. 3.7). Ideally, one would compute such a topic mixture by maximizing over both $\boldsymbol{\Delta}$ and vote outcome 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 (Fig. 6) faced by a single amicus as we vary the topic proportions of *individual mandate* and *interstate commerce* topics over multiples of 0.1.

Consequently, the 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 *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 Fig. 5 (right) using these "best" proportions. The "best" proportions IPs are different (sometimes worse) from that in Fig. 5 (left) because in the latter,

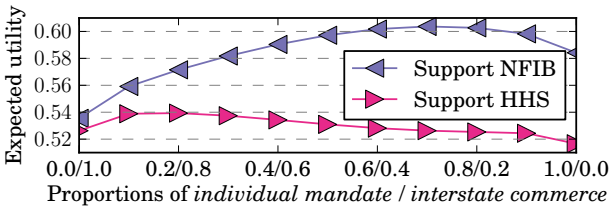


Figure 6: Expected utility when varying between proportions of *individual mandate* and *interstate commerce* topics.

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.

3.4 Significance

In this section, we have introduced a framework for explicit modeling of authors as utility maximizing agents, and incorporated their utility functions into a probabilistic model over their text and responses. Our approach led to improved vote prediction in SCOTUS, and demonstrated that such model captures the structure of amicus briefs better than simpler treatments of the text. Since our model characterizes the amicus brief as a (probabilistic) function of the case parameters, it could also be used to ask how the amici would have altered their briefs given different merits facts or a different panel of justices.

Moreover, we anticipate that our model will be a useful tool for quantitative analysis and hypothesis generation; political scientists can use our tools to support substantive research on the judiciary, while our model can recommend "optimal topics" for amici to pursue their legal agendas. Although we focus on SCOTUS, our model is applicable to any setting where textual evidence for competing goals is available alongside behavioral response. In subsequent sections (and the rest of the thesis), we will build on the tools we have developed here to characterizes authors' strategic behaviors.

4 Characterizing Authors' Strategic Behaviors

In our prior works (described in §2 and §3), we have either assumed that author's attributes were known, or independent of his strategic choice. However, authors have different motivations and thus make different choices. Here, we propose to treat authors as unique entities with latent attributes and thus different utility functions. In the subsequent sections, we will consider datasets in two different domains — the scientific community (§4.1) and judicial politics (§4.2) — and propose extensions to our prior work to characterize these author's utility functions.

4.1 Strategic Behavior in the Scientific Community

(Part of the work in this subsection has been submitted to EMNLP 2015 and is currently under review.)

Authoring a scientific paper is a complex process involving many decisions. As researchers, we write papers to report new scientific findings, but this is not the whole story. Authoring a paper involves a huge amount of decision-making that may be influenced by factors such as institutional incentives, attention-seeking, and pleasure derived from research on topics that excite us.

We propose that text collections and associated metadata can be analyzed to reveal optimizing behavior by authors. Specifically, we consider the ACL Anthology Network Corpus (Radev et al., 2013), along with author and citation metadata. Our main contribution is a method that infers 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.

The method is based on a probabilistic model that 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 preferences), papers (i.e., what they are "about"), and the community as a whole (in responding with citations). This method can be used to make predictions; 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?"

This method can also be used for exploration and to generate hypotheses. We provide an intriguing example relating author tradeoffs to age within the research community.

Author Utility. 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; here, 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 (or downloads) 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.

Our main assumption about author a is that she is optimizing: 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 $\boldsymbol{\theta}_d$:

$$U(\boldsymbol{\theta}_d) = \kappa_a y_d - \frac{1}{2}\|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2 \tag{4.1}$$

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

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 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 the topic.

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

$$y_d = \boldsymbol{\beta}^\top \boldsymbol{\theta}_d + \xi_d \tag{4.2}$$

where $\xi_d \sim \mathcal{N}(0, 1)$. Under this assumption, the author's *expected* utility assuming she is aware of $\boldsymbol{\beta}$ (often called "rational expectations" in discrete choice models), is:

$$\mathbb{E}[U(\boldsymbol{\theta}_d)] = \kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta}_d - \frac{1}{2}\|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2 \tag{4.3}$$

(This is obtained by plugging the expected value of y_d, from Eq. 4.2, into Eq. 4.1.) An author's decision will therefore be

$$\hat{\boldsymbol{\theta}}_d = \arg\max_{\boldsymbol{\theta}} \kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta} - \frac{1}{2}\|\boldsymbol{\theta} - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2 \tag{4.4}$$

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

$$\kappa_a \beta_j - (\hat{\theta}_{d,j} - (\eta_{a,j} + \epsilon_{d,a,j})) = 0, \ \forall 1 \leq j \leq K \tag{4.5}$$

Eq. 4.5 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.

Coauthorship Utility. We model the joint expected utility of a set of authors, $\boldsymbol{a}_d$, in writing $\boldsymbol{\theta}_d$ as the average of the group's utility.

$$\mathbb{E}[U(\boldsymbol{\theta}_d)] = \frac{1}{|\boldsymbol{a}_d|} \sum_{a \in \boldsymbol{a}_d} \left(\kappa_a \boldsymbol{\beta}^\top \boldsymbol{\theta}_d - \frac{1}{2} c_{d,a} \|\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})\|_2^2 \right) \tag{4.6}$$

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

$$\sum_{a \in \boldsymbol{a}_d} \kappa_a \boldsymbol{\beta} - c_{d,a}(\boldsymbol{\theta}_d - (\boldsymbol{\eta}_a + \boldsymbol{\epsilon}_{d,a})) = \mathbf{0} \tag{4.7}$$

Modeling Document Content. There are several possible ways for modeling the document content $p(\boldsymbol{w}_d \mid \boldsymbol{\theta}_d)$. Following considerable past work, we treat $\boldsymbol{\theta}_d$ a mixture of topics, where each topic is a categorical distribution over words (e.g. Blei et al. (2003), Hofmann (1999)). Since $\boldsymbol{\theta}_d \in \mathbb{R}^K$ is unconstrained, we transform it into a categorical distribution over topics using the softmax transformation. Using mixtures of topics instead of a *bag-of-words* representation provides us with a low dimensional interpretable representation that will be useful for characterizing authors' behaviors and preferences. Each dimension, j of an author's preference, is thus 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 assume a different prior for the topic mixtures.

Following the random utility model approach similar to that described in §3, we can make use of MCMC machinery for estimating the parameters in the probabilistic graphical model.

Evaluation. A familiar corpus of scientific publications in the computational linguistics community is the ACL Anthology Network Corpus[14] (Radev et al., 2013), which currently contains 21,212 papers, published

[14] `http://clair.eecs.umich.edu/aan/`

Figure 7: 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.

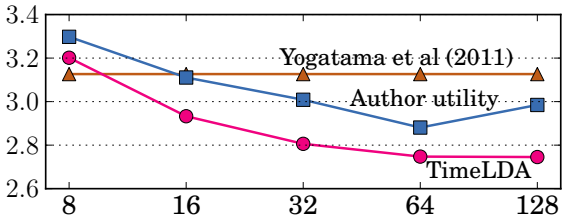
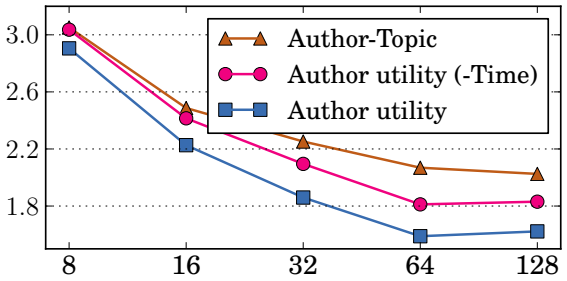


Figure 8: Held-out perplexity ($\times 10^3$, y-axis) with varying number of topics K (x-axis). The differences are significant between all models at $K \geq 64$ (paired t-test, $p < 0.01$). There are 523,381, 529,397, 533,792 phrase tokens in each random split of the test set.



in our field between 1965 and 2013, and written by 17,792 distinct authors. The corpus contains papers that are, and provides metadata such as authors, venue and in-community citation networks.

We quantitatively evaluate our model in two ways: (i) mean absolute error when predicting citation counts (Fig. 7), and (ii) predicting the content of held out documents using perplexity (Fig. 8). For the former, Yogatama et al. (2011) is a strong baseline for predicting responses using n-gram features and metadata features in a generalized linear model with the time series prior. "TimeLDA" is a version of our model without the author utility component; this equates to replacing Yogatama et al.'s features with LDA topic mixtures, and performing joint learning of the topics and citations. With sufficiently many topics ($K \geq 16$), low-dimensional 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. baseline.

For perplexity, we compared to the Author-Topic model of Rosen-Zvi et al. (2004). 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 $\boldsymbol{\eta}_a$. We include a version of our author utility model that ignores temporal information ("–time"), i.e., setting $T = 1$ and collapsing all timesteps. We find that perplexity improves with the addition of the utility model as well as the temporal dynamics.

Qualitatively, we find intriguing patterns in author trade-offs (i.e., κs) that correlate with academic seniority. In Fig. 9, 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. [15] 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

[15] 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.

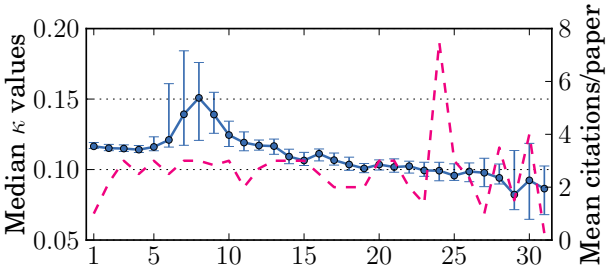


Figure 9: Plot of authors' median κ (blue, solid) and mean citation counts (magenta, dashed) against their academic age in this dataset.

more 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.

4.2 Strategic Behavior in the Supreme Court

In §3, we modeled the behavior of amici through their textual artifacts as a group (through the means of their topic mixtures), but did not account for the strategic motivations of individual amicus. Amici curiae are fundamentally, *interest groups*, and SCOTUS is an attractive forum to pursue their agendas. They participate in the judicial system by filing amicus briefs, which is relatively inexpensive compared to direct litigation.

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 in §3, 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, the strategic behavior that we are interested in is: can we characterize an amicus' by the extent to which they pursue their policy goals?

Amicus Utility Model. Let θ denote the "facts of the case" (i.e., representation of the merits briefs), ψ_j be the preference of justice j. Suppose there are M different amici interested in the case, and each of them with preferences η_k files a brief Δ_k. After deliberation, SCOTUS justices will jointly author an opinion Ω.[16] The utility function of an amicus m is

$$U(\Delta_m; \theta, \{\psi\}_j, \eta_m) = \kappa_m s_{\text{reward}}(\Delta_m, \Omega) - s_{\text{cost}}(\Delta_m, \eta_m) \tag{4.8}$$

[16] In case of disagreements between justices, more than one opinion may be written to explain the rationale behind their decisions. But for our example here, we assume that just one opinion was written.

where $s(x, y)$ denote some similarity metric between x and y. For an amicus who is particularly focused on positioning themselves close to the outcome of a case, they will have a large κ_m, and therefore seek to write briefs that will be similar to the majority opinion (i.e., maximize the first term in Eq. 4.8, $s_{\text{reward}}(\cdot, \cdot)$). On the other hand, a small κ_m would imply that the amicus tend to write close to their preferences. In the judicial domain, we are interested in the legal arguments being presented; hence, we propose to define similarity in terms of citation overlaps between Ω and Δ_m. Precedents play a central role in common law judiciary (Fowler et al., 2007), and we expect that citations will serve as a useful proxy to the legal issues and arguments that are being discussed. Since citations are discrete entities, it will be a computational challenge to compute expectations (of $s_{\text{reward}}(\cdot, \cdot)$) over the citation space — we will need to use a continuous low-dimensional representation (like our past work), or develop approximation algorithms.

The amicus will not know before filing his brief what the eventual opinion is — we will need a model for predicting court opinions. We propose to model opinions as a mixture of the merits, θ, justices' preferences, $\{\psi\}_j$, and amicus brief, Δ_m. One can think of opinions as being co-authored by $J + 2$ different parties — we can build on existing methods based on topic models (Rosen-Zvi et al., 2004), or sparse additive effects models (Eisenstein et al., 2011).

The second term of Eq. 4.8, $s_{\text{cost}}(\cdot, \cdot)$, measures how far the amicus brief has strayed from their intrinsic preferences — like our past work, we can use a squared L_2 norm.

The amicus utility (Eq. 4.8) follows a similar form as the author utility (Eq. 4.6). However, our response variable is no longer a simple linear function, which means that there is no simple closed form for the first-order optimality conditions. Thus, we will need to incorporate it as part of the objective similar to the random utility model in §3.2.

Evaluation. Using our SCOTUS corpus in §3 we can evaluate our amicus utility model on predicting held out justices' opinions (by measuring perplexity). Unlike §3, predicting vote outcomes is no longer trivial in our proposed model. However, we can still predict the justices who are most likely to be involved in writing the majority opinion; this will serve as a proxy for predicting the votes of individual justices. We can also pre-register hypothesis of our model's output regarding how amici behaves and check the validity of our models and assumptions. Qualitatively, we can analyze the behaviors of interest groups filing amicus briefs — Which interest groups are particularly focused on pursuing their policy goals?

Note that special care is needed when interpreting our model's results. It is tempting to say that an amicus, who consistently files briefs with high s_{reward}, is effective at influencing policy — this is not necessarily true. It is equally likely that justices were going to write the same opinions, regardless of the briefs. In this case, the amicus is merely a good forecaster, and probably not an influencer. Discerning between each of the above will require careful experimental design (e.g., identifying similar cases to perform counterfactuals) and additional information beyond will be necessary (e.g., instances of explicit references to amicus briefs in the text).

5 Conclusion

The work presented in this thesis describes several instances of strategic behavior and presents several computational models to infer their behavior from data. In our models, we have made simplifying assumptions

about how authors behave — that what authors care about can be captured in a utility function. In a perfect world where these assumptions are true, our models will be able to recover the true preferences of authors and their tradeoffs between cost and benefits. However, the construction of a complete utility function is intractable and hence, recovering the underlying motives of authors will be impossible. Throughout this work, we have highlighted some of these pitfalls when drawing conclusions from the results.

We view our models as tools for in-depth exploration of text data and hypothesis generation. We used our models to empirically testify widely-held hypothesis about the U.S presidential elections, and to perform in-depth analysis of SCOTUS. In §4, we proposed to extend the idea of *text as a strategic choice* further by characterizing each author differently. Under certain assumptions of rational and strategic behavior, we are able to make use of utility models to improve predictive performance; that is, even though our models cannot recover "true motives" of authors, they do contain useful signals. Likewise, using our models, we found empirical evidence that were consistent with our prior knowledge — that researchers tend to be less concerned about extrinsic payoffs over time.

With further work on the proposed topics, we hope to develop new machinery to make inferences about how authors will behave, and provide us with deeper insight into understanding how language is being used for influence. We envision that these models will be useful to authors of actuating texts (i.e., politicians, journalists, lobbyists, etc) by recommending content that help them better frame their writings to pursue their agendas. From the perspective of social science, this will be an analysis tool that leverages text to enable a more complex analysis of individual actor's behaviors.

6 Timeline

The timeline for this thesis will be organized around paper submission deadlines, and the overall goal is to complete this thesis within a year (defending in June 2016).

- June 2015: Submit paper on strategic behavior in the scientific community (§4.1). Target: EMNLP 2015, deadline May 30, 2015.
- Fall 2015: Work on modeling strategic behavior in the U.S. Supreme Court (§4.2). Target: ACL 2016, deadline Jan/Feb, 2016.
- Spring 2016: Write thesis dissertation.
- June 2016: Thesis defense.

Bibliography

Galen Andrew and Jianfeng Gao. Scalable training of l_1-regularized log-linear models. In *Proceedings of ICML*, ICML '07, pages 33–40. ACM, 2007. URL http://doi.acm.org/10.1145/1273496.1273501.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, May 2014a. ISSN 1467-9841. doi: 10.1111/josl.12080.

David Bamman, Ted Underwood, and Noah A. Smith. 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)*, pages 370–379, Baltimore, Maryland, June 2014b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P14/P14-1035.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. Guiding semi-supervision with constraint-driven learning. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 280–287, June 2007. URL http://cogcomp.cs.illinois.edu/papers/ChangRaRo07.pdf.

Joshua Clinton, Simon Jackman, and Douglas Rivers. The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370, 2004.

Paul M Collins. *Friends of the Supreme Court: Interest Groups and Judicial Decision Making*. Oxford University Press, August 2008. ISBN 019537214X. URL http://www.psci.unt.edu/~pmcollins/FOSC.htm.

Scott Alson Comparato. *Amici Curiae and Strategic Behavior in State Supreme Courts*. Greenwood Publishing Group, 2003.

Michael D. Conover, Bruno Goncalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 192–199, October 2011. doi: 10.1109/PASSAT/SocialCom.2011.34.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287, Cambridge, MA, USA, October 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1870658.1870782.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. Sparse additive generative models of text. In *Proceedings of International Conference on Machine Learning*, ICML '11, Bellevue, WA, USA, 2011. URL http://www.icml-2011.org/papers/534_icmlpaper.pdf.

James H. Fowler, Timothy R. Johnson, James F. Spriggs, Sangick Jeon, and Paul J. Wahlbeck. Network analysis and the law: Measuring the legal importance of precedents at the U.S. Supreme Court. *Political Analysis*, 15(3):324–346, May 2007.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, August 2010. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1756006.1859918.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984. doi: 10.1109/TPAMI.1984.4767596. URL http://dl.acm.org/citation.cfm?id=2286617.

Sean Gerrish and David Blei. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning*, ICML '11, pages 489–496, Bellevue, WA, USA, June 2011. ACM. URL http://www.icml-2011.org/papers/333_icmlpaper.pdf.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57 (1):pp. 97–109, 1970. URL http://www.jstor.org/stable/2334940.

Dustin Hillard, Stephen Purpura, and John Wilkerson. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46, 2008.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, August 2002. ISSN 0899-7667. URL http://dx.doi.org/10.1162/089976602760128018.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57. ACM, 1999. ISBN 1-58113-096-1. URL http://doi.acm.org/10.1145/312624.312649.

Benjamin E. Lauderdale and Tom S. Clark. Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, 58(3):754–771, February 2014. ISSN 1540-5907. doi: 10.1111/ajps.12085. URL http://dx.doi.org/10.1111/ajps.12085.

Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *The American Political Science Review*, 97(2):311–331, 2003. URL http://www.jstor.org/stable/3118211.

Andrew McCallum, Gideon Mann, and Gregory Druck. Generalized expectation criteria. Technical Report UM-CS-2007-60, University of Massachusetts, Amherst, MA 01003, USA, August 2007.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, 1974.

Burt L. Monroe and Ko Maeda. Talk's cheap: Text-based estimation of rhetorical ideal-points, 2004. Presented at the Annual Meeting of the Society for Political Methodology.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Saša Petrović, Miles Osborne, and Victor Lavrenko. RT to win! predicting message propagation in twitter. July 2011. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2754/3209.

Keith T. Poole and Howard Rosenthal. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384, 1985. URL http://www.jstor.org/stable/2111172.

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

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. The ACL anthology network corpus. *Language Resources and Evaluation*, pages 1–26, 2013. URL http://dx.doi.org/10.1007/s10579-012-9211-2.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494. AUAI Press, 2004. URL http://dl.acm.org/citation.cfm?id=1036843.1036902.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 91–101. Association for Computational Linguistics, October 2013. URL http://www.aclweb.org/anthology/D13-1010.

Yanchuan Sim, Bryan Routledge, and Noah A. Smith. The utility of text: The case of amicus briefs and the Supreme Court. In *AAAI Conference on Artificial Intelligence*, AAAI '15, January 2015. URL http://arxiv.org/abs/1409.7985.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, March 2009. ISSN 1532-2882. doi: 10.1002/asi.v60:3.

Tae Yano. *Text as Actuator: Text-Driven Response Modeling and Prediction in Politics*. PhD thesis, Carnegie Mellon University, September 2011. URL http://www.cs.cmu.edu/~taey/pub/tae-yano-actuating-text.pdf.

Tae Yano, William W. Cohen, and Noah A. Smith. 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, pages 477–485, Boulder, CO, USA, June 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1620754.1620824.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 594–604, Edinburgh, UK, July 2011. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2145432.2145501.