# Measuring Ideological Proportions in Political Speeches

**Yanchuan Sim**[*] **Brice D. L. Acree**[†] **Justin H. Gross**[†] **Noah A. Smith**[*]

[*]Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{ysim,nasmith}@cs.cmu.edu

[†]Department of Political Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA
{brice.acree,jhgross}@unc.edu

## Abstract

We seek to measure political candidates' ideological positioning from their speeches. To accomplish this, we infer ideological cues from a corpus of political writings annotated with known ideologies. We then represent the speeches of U.S. Presidential candidates as sequences of cues and lags (filler distinguished only by its length in words). We apply a domain-informed Bayesian HMM to infer the proportions of ideologies each candidate uses in each campaign. The results are validated against a set of preregistered, domain expert-authored hypotheses.

## 1 Introduction

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; Hart, 2009; Deirmeier et al., 2012; Hart et al., 2013) and computational linguistics (Thomas et al., 2006; Fader et al., 2007; Gerrish and Blei, 2011, *inter alia*). In American politics, candidates for office give speeches and write books and manifestos expounding their ideas. Every political season, however, there are accusations of candidates "flip-flopping" on issues, with opinion shows, late-night comedies, and talk radio hosts replaying clips of candidates contradicting earlier statements. 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."

A more general observation, often stated but not yet, to our knowledge, tested empirically, is that successful primary candidates "move to the center" before a general election. The expectation follows directly from long-standing and widely influential theories of political competition that are collectively referred to in their simplest form as the "median voter theorem" (Hotelling, 1929; Black, 1948; Downs, 1957). Thus 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.

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 (Laver et al., 2003; Monroe and Maeda, 2004; Hillard et al., 2008), 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).[1] The corpus then serves as evidence for

[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,"'

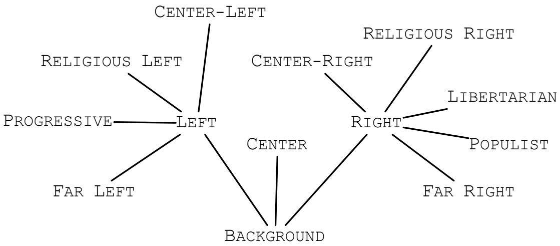

Figure 1: Ideology tree showing the labels for the ideological corpus in §2.1 (excluding BACKGROUND) and corresponding to states in the HMM (§3.3).

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 (§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 1) encoding 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 (§4).

## 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.

| | | |
|---|---|---|
| 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 1: Ideology corpus statistics. Note that some documents are not labeled with finer-grained ideologies.

### 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. 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[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 Fig. 1.[3] Table 1 summarizes key details about the ideological corpus.

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

etc.). Analysis of positions on specific issues is left for future work.

[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.

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; the full list can be found in the supplementary materials, along with a table summarizing key details about the corpus, which contains 32.8 million tokens.

## 2.2 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).

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:

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

Each of the $\eta$ 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 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 $\ell_1$ 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). To reduce the complexity of the hyperparameter space (the possible values of all $\lambda_a$) and to encourage similar levels of sparsity across the different effect vectors, we let, for each ideology attribute $a$,

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

where $\mathcal{V}(a)$ is the set of term types appearing in the data with attribute $a$ (i.e., its vocabulary) , and $\lambda$ 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).

## 2.3 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.[4] 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 cue discovery model.

### 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) (Schneider, 2005; Cover and Thomas, 2012). Selected features were used to classify the ideologies of held-out documents from our corpus.[5] We evaluated these feature selection methods within naïve Bayes classification in a 5-fold cross-validation setup. We vary $\lambda$ 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 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.

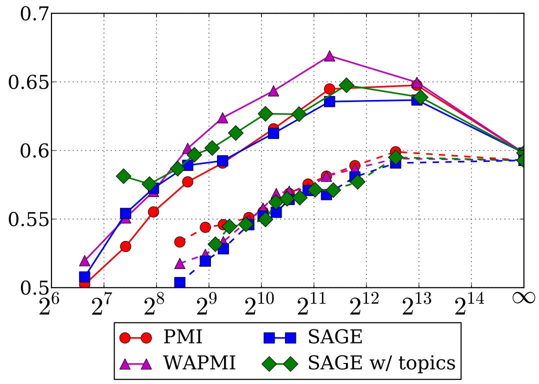

Figure 2: Plot of average classification accuracy for 5-fold cross validation against the number of features. Dashed lines refer to using only unigram features, while solid lines refer to using bigram and trigram features.

### 2.5 Cue Lexicon

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. The supplementary materials include top cue terms associated with various ideologies and a heatmap of similarities among SAGE vectors.

We conducted a small, relatively informal study in which seven subjects (including four scholars of American politics) were asked to match brief descriptions of the classes, including prominent prototypical individuals exemplifying each, to cue terms. About 70% of ideologies were correctly matched by experts, with relatively few confusions between Left and Right. More details are given in supplementary materials.

## 3 Second Stage: Cue-Lag Ideological Proportions

The main contribution of this paper is a technique for measuring ideology proportions in the prose of political candidates. We adopt a Bayesian approach that manages our uncertainty about the cue lexi-

[4] 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).

[5] 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.

con $\mathcal{L}$, 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 (Poole and Rosenthal, 1985; Poole and Rosenthal, 2000; Clinton et al., 2004). 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.

### 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." Table 2 presents a breakdown of the candidates and speeches in our corpus.

### 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 3 shows an example of our cue-lag representation.

| 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 2: 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.

### 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 (Fig. 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 §3.3.3).
2. Let the initial state be the BACKGROUND state.
3. For $t \in \{1, 2, \dots, T\}$:[6]
   (a) Transition to state $S_t$ based on the transition distribution, discussed in §3.3.1. This transition is conditioned on the previous state $S_{t-1}$ and the lag at timestep $t-1$, denoted by $L_{t-1}$.

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

| 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 3: Example of the cue-lag representation.

(b) Emit cue term $W_t$ from the lexicon $\mathcal{L}$ and lag $L_t$ based on the emission distribution, discussed in §3.3.2.

We turn next to the transitions and emissions.

### 3.3.1 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 Fig. 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:

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

$Path(s_i, s_j)$ 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 $\zeta_{s_j}$.

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, $\ell$. The probability of restarting the walk from the BACKGROUND state is a noisy-OR model with parameter $\rho$. This gives the transition distribution:

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

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 $|\mathcal{J}|^2$ transitions with $O(|\mathcal{J}|)$ parameters.[7] 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; Mihalcea, 2005; Toutanova et al., 2004; Collins-Thompson and Callan, 2005).

### 3.3.2 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 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 $\nu$.

[7] More precisely, there are $|\mathcal{J}|$ edges (since there are $|\mathcal{J}|+1$ 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 $2|\mathcal{J}|-(|\mathcal{J}|+1) = |\mathcal{J}|-1$ degrees of freedom for $\boldsymbol{\theta}$. There are $|\mathcal{J}|$ $\boldsymbol{\zeta}$-parameters and a single $\rho$, for a total of $2|\mathcal{J}|$ degrees of freedom.

### 3.3.3 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 $\{\mathcal{L}(i)\}_i$, 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 on the tree walk probabilities $\boldsymbol{\theta}$; its parameter is $\alpha$. For the cue emission distribution associated with ideology $i$, $\boldsymbol{\psi}_{s_i}$, we use an *informed* Dirichlet prior with two different values, $\beta_{cue}$ for cues in $\mathcal{L}(i)$, and a smaller $\beta_{def}$ for those in $\mathcal{L} \setminus \mathcal{L}(i)$.[8]

Learning proceeds by collapsed Gibbs sampling for the hidden states and slice sampling (with vague priors) for the hyperparameters ($\alpha$, $\boldsymbol{\beta}$, $\rho$, and $\boldsymbol{\zeta}$). Details of the sampler are given in the supplementary materials. At each Gibbs step, we resample the ideology state and restart indicator variable for every cue term in every speech.

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 his/her ideological proportions.

In order 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 ideology $i$ during timestep $t$, we say that he spends $\frac{1}{2}(L_{t-1} + L_t)$ amount of time in ideology $i$. 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 the epoch.

## 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 (§2.1), 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. 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 3.

## 5 Evaluation

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

- HMM: rather than §3.3.1, 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 $\boldsymbol{\theta}$.

[8] 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.

| 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 3: 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.

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

**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.

**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 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. Similarly, *barack_obama* and *presid_obama* are Left cues (though *senat_obama* is a Right cue). In future work, we believe filtering candidate names in the first stage will be beneficial.

**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.

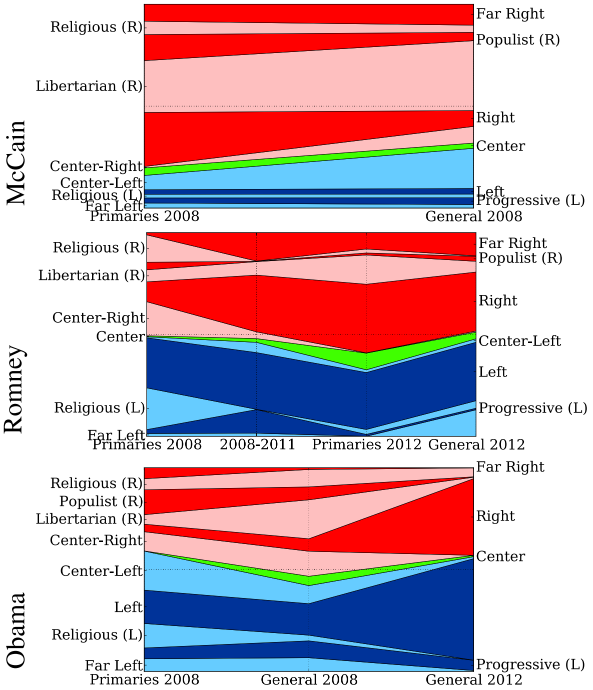

Figure 4: Proportion of time spent in each ideology by McCain, Romney, and Obama during the 2008 and 2012 Presidential election seasons.

**"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 4, 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), 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 plan to use CLIP as a text analysis method to support substantive inquiry in political science, such as following trends in expressed ideology over time.

## 6 Related Work

As early as the 1960s, there has been research on modeling ideological beliefs using automated systems (Abelson and Carroll, 1965; Carbonell, 1978; Sack, 1994). These early works model ideology at a sophisticated level, involving the actors, actions and goals; they require manually constructed knowledge bases. Poole and Rosenthal (1985) used congressional roll call data to demonstrate the ideological divide in Congress, and provided a methodology for measuring ideological positions. Gerrish and Blei (2011; 2012) augmented the methodology with text from congressional bills using probabilistic models to uncover lawmakers' positions on specific political issues, putting them on a left-right spectrum, while Thomas et al. (2006) made use of floor debate speeches to predict votes. Likewise, taking advantage of the proliferation of text today, numerous techniques have been developed to identify topics and perspectives in the media (Gentzkow and Shapiro, 2005; Lin et al., 2008; Fortuna et al., 2009; Gentzkow and Shapiro, 2010); determine the political leanings of a document or author (Laver et al., 2003; Efron, 2004; Mullen and Malouf, 2006; Fader et al., 2007); or recognize stances in debates (Somasundaran and Wiebe, 2009; Anand et al., 2011). Going beyong lexical indicators, Greene and Resnik (2009) investigated syntactic features to identify perspectives or implicit sentiment.

## 7 Conclusions

We introduced CLIP, a domain-informed, Bayesian model of ideological proportions in political language. We showed how ideological cues could be discovered from a lightly labeled corpus of ideological writings, then incorporated into CLIP. The resulting inferences are largely consistent with a set of preregistered hypotheses about candidates in the 2008 and 2012 Presidential elections.

## Acknowledgments

For thoughtful feedback on this research, the authors thank: several anonymous reviewers, Amber Boydstun, Philip Resnik, members of the ARK group at CMU, and participants in Princeton University's Political Methodology Colloquium and PolMeth XXX hosted by The University of Virginia. This work was supported in part by an A*STAR fellowship to Y. Sim, NSF grants IIS-1211201 and IIS-1211277, and Google's support of the Reading is Believing project at CMU.

## References

Robert P Abelson and J Douglas Carroll. 1965. Computer simulation of individual belief systems. *American Behavioral Scientist*, 8(9):24–30.

Pranav Anand, 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*, pages 1–9.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of l1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 33–40, New York, NY, USA. ACM.

Duncan Black. 1948. On the rationale of group decision-making. *The Journal of Political Economy*, 56(1):23–34.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Jaime G. Carbonell. 1978. Politics: Automated ideological reasoning. *Cognitive Science*, 2(1):27–51.

Jonathan Charteris-Black. 2005. *Politicians and Rhetoric: The persuasive power of metaphor*. Palgrave-MacMillan.

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

Kevyn Collins-Thompson and Jamie Callan. 2005. Query expansion using random walk models. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 704–711, New York, NY, USA. ACM.

Thomas M Cover and Joy A Thomas. 2012. *Elements of information theory*. Wiley-interscience.

Daniel Deirmeier, Jean-Francois Godbout, Bei Yu, and Stefan Kaufmann. 2012. Language and ideology in congress. *British Journal of Political Science*, 42(1):31–55.

Anthony Downs. 1957. *An economic theory of democracy*. Harper, New York.

Miles Efron. 2004. Cultural orientation: Classifying subjective documents by cociation analysis. In *AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on Machine learning*, ICML '11.

Anthony Fader, 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)*, pages 658–666, Prague, Czech Republic, June. Association for Computational Linguistics.

Blaz Fortuna, Carolina Galleguillos, and Nello Cristianini. 2009. Detecting the bias in media with statistical learning methods. *Text mining: classification, clustering, and applications 10*, 27.

Matthew Gentzkow and Jesse Shapiro. 2005. Media bias and reputation. Technical report, National Bureau of Economic Research.

Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71.

Sean M. Gerrish and David M. Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of the 28th international conference on Machine learning*, ICML '11.

Sean M. Gerrish and David M. Blei. 2012. How they vote: Issue-adjusted models of legislative behavior. In *Advances in Neural Information Processing Systems 25*, pages 2762–2770.

Stephan Greene 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, pages 503–511, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roderick P Hart, Jay P Childers, and Colene J Lind. 2013. *Political Tone: How Leaders Talk and Why*. University of Chicago Press.

Roderick P Hart. 2009. *Campaign talk: Why elections are good for us*. Princeton University Press.

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

Harold Hotelling. 1929. Stability in competition. *The Economic Journal*, 39(153):41–57.

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

Wei-Hao Lin, 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, pages 17–32. Springer-Verlag, Berlin, Heidelberg.

Rada Mihalcea. 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*, HLT '05, pages 411–418, Stroudsburg, PA, USA. Association for Computational Linguistics.

Burt L Monroe and Ko Maeda. 2004. Talk's cheap: Text-based estimation of rhetorical ideal-points. In *Annual Meeting of the Society for Political Methodology*, pages 29–31.

Tony Mullen and Robert Malouf. 2006. A preliminary investigation into sentiment analysis of informal political discourse. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 159–162.

Keith T. Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):pp. 357–384.

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

Warren Sack. 1994. Actor-role analysis: ideology, point of view, and the news. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, August.

Karl-Michael Schneider. 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, pages 252–263, Berlin, Heidelberg. Springer-Verlag.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 226–234, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matt Thomas, 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, pages 327–335, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristina Toutanova, 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, pages 103–, New York, NY, USA. ACM.