Identifying Media Frames and Frame Dynamics Within and Across Policy Issues

Amber E. Boydstun, University of California, Davis
Justin H. Gross, University of North Carolina, Chapel Hill
Philip Resnik, University of Maryland, College Park
Noah A. Smith, Carnegie Mellon University

September 16, 2013

Abstract

Framing is a central concept in political communication and a powerful political tool. Understanding what frames are used to define specific issues and also what general patterns are evidenced by the evolution of frames over time is hugely important. It is also a serious challenge, thanks to the volume of text data, the dynamic nature of language, and the variance in applicable frames across issues (e.g., the 'innocence' frame of the death penalty debate is not used in discussing smoking bans). We describe a project that advances framing research methodology in two ways. First, we are developing a unified coding scheme for content analysis across issues, whereby issue-specific frames (e.g., innocence) are nested within high-level dimensions (or frame *types*) that cross-cut issues (e.g., fairness); we are validating this coding scheme by applying it to news coverage of immigration, same-sex marriage, and smoking/tobacco in the United States over the course of the past twenty-three years. Second, we are developing methods for semi-automated and automated frame discovery aimed at both replicating manual coding and isolating patterns of frame evolution that might not be readily visible to human inspection. Our goal is to employ strategies heavily informed by existing work in natural language processing, but tailored to the specific needs and professional sensibilities of political communications scholars.

1 Introduction

Framing—portraying an issue from one perspective to the necessary exclusion of alternative perspectives—is a central concept in political communication (see Schaffner and Sellers, 2009, Introduction for a nice overview, and the remainder of the edited volume for several illustrations). It is widely accepted that framing can have a significant influence on public attitudes toward important policy issues (e.g., Chong and Druckman, 2007; Nelson et al., 1997) and on the application of policy issues directly (e.g., Baumgartner et al., 2008). Understanding, for a given issue, what frames are used by politicians, the media, and the voting public to communicate about it, is an enormous challenge, due to the dynamic and creative nature of language and the growing volume of data in which frames appear and develop over time. As engagement by citizens in the political discourse broadens via the widespread adoption of blogging, commenting, and other social media, scientific study of the political world requires reliable analysis of how issues are framed, ideally in real time. Yet the process by which a political scientist or communications scholar identifies the catalogue of frames in a political discourse about a particular issue (**frame discovery**) is complex and labor-intensive; so is the secondary process of coding instances of framing in text (**frame analysis**) in order to reveal patterns in frame usage.

Moreover, the very definition of framing has been notoriously slippery. The most widely employed definition among current researchers in political communication is provided by Robert Entman: "Framing essentially involves *selection* and *salience*. To frame is to *select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation* for the item described" (Entman, 1993, p. 52, emphasis by the author). Beyond the challenge of gaining consensus on a conceptual definition, the matter of operationalizing the notion of issue framing presents its own set of difficulties. As Matthes and Kohring point out in a recent article aimed at improving the reliability and validity of content analytic measures of media frames, "a frame is a quite abstract variable that is hard to identify and hard to code in content analysis" (Matthes and Kohring, 2008, p. 258). As a result, measurement is challenging; in the words of another researcher, "it is extremely difficult to neutralize the impact of the researcher in framing research" (Van Gorp, 2005, p. 503, cited in Matthes and Kohring, 2008).

Despite the important nuances of different conceptual definitions of framing, all these definitions treat language as central, making the tools of natural language processing especially important. Whether the aim is to study frames defined as memes, dimensions of debate, or specific arguments for or against something, all of these framing techniques can be identified and analyzed through language signals ranging from simple lexical clues to word clusters to choices of syntactic structure. Natural language processing promises to help us meet some of the challenges inherent in operationalizing and measuring frames, as we will discuss below.

Additionally, beyond the frames associated with specific issues, scholars and citizens alike face the challenge of being able to trace how frames are used across multiple policy debates. We can examine how the death penalty is framed in terms of "an eye for an eye" vs. "cruel and unusual punishment," for example, but understanding the mechanisms and effects of framing as broader phenomena requires better methods that allow us to trace overarching tropes *across* a variety of issue debates; for example, framing various policies in terms of potential losses or potential gains (Kahneman and Tversky, 1979), using divergent metaphors of the nation as a family (Lakoff, 2004), or invoking a value that is widely embraced, yet whose meaning is deeply contested (e.g. "freedom"), across a spectrum of settings (Lakoff, 2006). Here, too, language is key to identifying and analyzing frames that cross-cut issues, and again, therefore, natural language processing is a key research tool.

Whether we are looking at politicians' communications, at traditional news media, or particularly at social media, the availability of online data creates an unprecedented opportunity to track framing in near-real-time, and understand it as an evolutionary process across time, issues, and communication venues (e.g., different media formats). This means that understanding framing as a general phenomenon requires large-scale text data analysis well beyond what has been accomplished by expert manual annotation alone.

Our three-year interdisciplinary project, funded by the U.S. National Science Foundation, takes first steps toward a data-driven, expert-informed, computational model of framing that augments an expert's ability to discover frames and analyze their use in textual discourse. The project's goals are to:

1. Develop algorithms for automated frame **analysis**, leveraging high-level domain knowledge and issue-specific knowledge from expert political scientists, annotated examples, and unannotated text and contextual metadata linked to the text. Concretely, we are developing algorithms to produce similar results to human coders who achieve high inter-coder reliability at identifying frames used in text passages from contemporary political discourse, relating to a range of issues.

2. Integrate insights from political science experts with statistical analysis of political discourse

in order to enable frame **discovery**. Concretely, we are developing tools that speed up the work of human experts and will reveal frames that might be missed by the subjective "naked eye."

3. Apply both of these methods to better understand the **evolutionary** process behind frame development and dynamics for specific issues in current American politics, across multiple years and multiple traditional and social media streams.

This paper describes some of our key research activities during the first year. We focus primarily on the initial development of the Policy Frames Codebook (goal 1; §2), and we briefly summarize the development of two statistical models. One measures proportions of ideological content in political speeches, representing ideology as a discrete variable (§3), and the second induces a two-level latent topic structure inspired by the theory of framing as second-level agenda setting (McCombs, 2002; §4).

2 Policy Frames Codebook

It is perhaps fitting that issue-framing, also known as issue-definition, has been defined in many different ways. These varied definitions have helped produce a burgeoning framing literature, but they leave us without the ability to examine patterns in framing both within and across issues over time. Here, we describe the benefits of framing schemas that cross-cut policy issues, and we introduce just such a schema: the Policy Frames Codebook. Just as the Policy Agendas Codebook[1] provides a system for categorizing topics across policy agendas, our Policy Frames Codebook provides a system for categorizing frames across policy issues. As a key outcome of our project intended for use in the wider community, this codebook will be a carefully validated resource that is useful in both human and automated content analysis. For those who wish to use it for conventional hand-coded content analysis, it will provide a common framework for cross-project comparison and replication, while remaining general enough to allow project-specific code development based on idiosyncrasies of individual issues and research questions about these issues. Should such researchers then wish to scale up to analysis of a larger corpus than can be efficiently handled by a small team of human coders, automated content analysis will then be an option, without having to start from scratch, as the codebook is designed and validation exercises conducted with scalability in mind. In this section we briefly motivate our approach and describe the codebook's present state of development.

Framing research has already benefitted from certain well-established schemas that generalize across issues. Iyengar (1991), for instance, identifies episodic frames (focused on specific incidents or cases) as distinct from thematic frames (focused on larger trends or context). Such general schemas facilitate invaluable insights into high-level patterns of political communication and, most importantly, their influence on public attitudes. For example, people who consume stories about poverty that are framed episodically by focusing on unemployed individuals are more likely to blame poverty on individual failings. People who consume thematic poverty stories, focused on national unemployment rates, are more likely to blame poverty on the government or other forces beyond an individual's control (Iyengar, 1991). However, existing generalized frame schemas do not unpack the topical content of frames as second-level agenda items (in the sense of McCombs, 2002), offering very little information about how the nature of a given policy debate shifts from one substantive dimension of the issue to another.

Other frame schemas are issue specific. For example, Baumgartner and colleagues trace framing in the case of capital punishment using an extensive codebook of frames specific to that issue (the

[1] http://www.policyagendas.org

death penalty does/does not deter crime, the death penalty system is/is not subject to error, etc.; Baumgartner et al., 2008). These issue-specific schemas are wonderfully detailed, but they do not allow us to examine patterns—and test hypotheses—across issues. In the death penalty study, for instance, the authors find suggestive evidence that conceptually-linked frames can "piggyback" on one another. The rise of the "innocence" frame in the mid-1990's, for instance, was accompanied by a rise in frames related to evidence, due process, classism, and racism. These related frames likely gained attention on the coattails of the innocence frame, but then in turn helped fuel that frame's momentum and attention to the death penalty overall. From this single case, it appears that the appearance of one frame may increase the likelihood of substantively linked frames being used (Baumgartner et al., 2008). However, without a coding schema that cross-cuts issues, we have no way of testing this hypothesis or others.

Our Policy Frames Codebook is intended to provide the best of both worlds: a general system for categorizing frames across policy issues designed so that it can also be specialized in issue-specific ways. The codebook contains 14 categories of frame "dimensions" (plus an "other" category) that are intended to be applicable to any policy issue (abortion, immigration, foreign aid, etc.) and in any communication context (news stories, Twitter, party manifestos, legislative debates, etc.). The dimensions are listed below.

1. **Economic frames**: The costs, benefits, or monetary/financial implications of the issue (to an individual, family, community or to the economy as a whole).

2. **Capacity and resources frames**: The lack of or availability of physical, geographical, spatial, human, and financial resources, or the capacity of existing systems and resources to implement or carry out policy goals.

3. **Morality frames**: Any perspective—or policy objective or action (including proposed action)—that is compelled by religious doctrine or interpretation, duty, honor, righteousness or any other sense of ethics or social responsibility.

4. **Fairness and equality frames**: Equality or inequality with which laws, punishment, rewards, and resources are applied or distributed among individuals or groups. Also the balance between the rights or interests of one individual or group compared to another individual or group.

5. **Constitutionality and jurisprudence frames**: The constraints imposed on or freedoms granted to individuals, government, and corporations via the Constitution, Bill of Rights and other amendments, or judicial interpretation. This deals specifically with the authority of government to regulate, and the authority of individuals/corporations to act independently of government.

6. **Policy prescription and evaluation**: Particular policies proposed for addressing an identified problem, and figuring out if certain policies will work, or if existing policies are effective.

7. **Law and order, crime and justice frames**: Specific policies in practice and their enforcement, incentives, and implications. Includes stories about enforcement and interpretation of laws by individuals and law enforcement, breaking laws, loopholes, fines, sentencing and punishment. Increases or reductions in crime.

8. **Security and defense frames**: Security, threats to security, and protection of one's person, family, in-group, nation, etc. Generally an action or a call to action that can be taken to protect the welfare of a person, group, nation sometimes from a not yet manifested threat.

9. **Health and safety frames**: Healthcare access and effectiveness, illness, disease, sanitation, obesity, mental health effects, prevention of or perpetuation of gun violence, infrastructure and building safety.

10. **Quality of life frames**: The effects of a policy on individuals' wealth, mobility, access to resources, happiness, social structures, ease of day-to-day routines, quality of community life, etc.

11. **Cultural identity frames**: The social norms, trends, values and customs constituting culture(s), as they relate to a specific policy issue

12. **Public opinion frames**: References to general social attitudes, polling and demographic information, as well as implied or actual consequences of diverging from or getting ahead of public opinion or polls.

13. **Political frames**: Any political considerations surrounding an issue. Issue actions or efforts or stances that are political, such as partisan filibusters, lobbyist involvement, bipartisan efforts, deal-making and vote trading, appealing to one's base, mentions of political maneuvering. Explicit statements that a policy issue is good or bad for a particular political party.

14. **External regulation and reputation frames**: The United States' external relations with another nation; the external relations of one state with another; or relations between groups. This includes trade agreements and outcomes, comparisons of policy outcomes or desired policy outcomes.

15. **Other frames**: Any frames that do not fit into the above categories.

Researchers may choose to employ only these categories as listed here, or they could also nest issue-specific frames (or arguments) within each category. For example, in the case of capital punishment, the "innocence" frame would be a frame specific to that issue but categorized under the dimension of "fairness and equality." In this way, scholars can apply the Policy Frames Codebook to new content analysis projects or take existing datasets that employed issue-specific frames and categorize those frames into the dimensions provided here.

We developed these categories through a mix of inductive and deductive methods. We began by brainstorming—amongst our team and several colleagues—categories that we imagined would cross-cut most, if not all, policy issues while also examining a random sampling of newspaper stories and blog posts to see which frames appeared and how we might categorize them. Then we tried applying our preliminary list of frame categories to a random sample of front-page newspaper stories covering a wide range of issues, and revised our categorization scheme accordingly. Next, we shopped our list around, sending it to additional colleagues and presenting it at an international conference (the 20th International Conference of Europeanists), again revising our schema based on this feedback. Finally, we did another round of test coding. Throughout this testing process, we developed and revised not only our list of categories but also a codebook that defines and gives examples for each category.

Researchers can apply these categories in whatever way suits their research aims. However, we advocate coding each piece of communication (e.g., newspaper story, blog post, Congressional bill) according to the *primary* frame category used, as well as the presence of any additional frames employed. For example, a news story focused on the economic impacts of immigration but with additional discussion about the challenges of immigrants' quality of life and cultural assimilation would receive three frame dimension codes—economic, quality of life, and cultural identity—but the economic dimension would be marked as primary. Moreover, in upcoming work, we will be hand-coding numerous example documents at a more granular level, letting coders select specific passages (paragraphs, sentences, phrases) that evoke particular frames from within a hierarchically-organized codebook particular to the policy area at hand.

Additionally, we suggest tracking the tone of each text. We differentiate among *positive*, *negative*, and *neutral* tones, where the precise definition varies according to issue being studied and the partition according to these designations will depend on the researcher's operationalization choices. For example, in pilot testing our immigration codebook on newspaper articles, we define these tones from the perspective of immigrants and their advocates; we might have instead defined them from the perspective of supporters of greater restrictionism, possibly with equivalent results (but perhaps not).

- **Positive tone**: Immigration and immigrants' rights are portrayed in a positive light or from a generally sympathetic point of view, so that immigrant advocates and supporters of less restrictive immigration laws would be pleased to see the news article.
- **Negative tone**: Immigration and immigrants' rights are portrayed in a negative light or in a non-sympathetic manner, so that immigrant advocates and supporters of less restrictive immigration laws would be disappointed or upset to see the news article.
- **Neutral tone**: Immigration and immigrants' rights are portrayed using both positive and negative tones that balance each other out, *or* the news article does not appear to discuss the issue either positively or negatively.

One can imagine other partitions we might have drawn on the space of tones. For example, one could define the perspective to be that of undocumented immigrants, so that an article drawing attention to the positive image of authorized immigrants in contrast with the undocumented would be negative in tone; in the coding above, it would be labeled "neutral" due to the ambiguity. Furthermore, some researchers may wish to study only *explicit* framing of an issue by clear advocates for a position—we expect examples to abound within editorials and op-ed columns, blog posts, and opinion commentary on cable news or talk radio, but they may also be found in "straight news" stories via the quotes of activists, politicians, and interest group members. In our definition of the three tone types above, we allow for detection of *implicit* frames as well. The coder is simply asked to put herself in the position of an individual directly affected by the issue at hand and must essentially decide whether the article would be appealing or distressing. In this case, the aspects of the issue receiving attention from a journalist may themselves rub certain readers the wrong way despite not overtly taking sides in a conflict. (Within recent computational linguistics literature, Recasens et al. (2013) draw a related distinction between *framing bias*, which involves explicitly subjective words or phrases linked with a particular point of view, and *epistemological bias*, which involves implicit assumptions and presuppositions in ostensibly neutral writing.)

Within the structure we are proposing, many options are left to the judgment of the researcher, but adopting this structure ensures that such judgment will be made explicit and defended within the context of one's research program.

In the coming months, we will work on automated frame analysis that is grounded in human coding based on the Policy Frames Codebook, focusing on the benefits of semi-automated iterative development—for example, supervised learning based on human-coded texts, and semi-supervised modeling designed to bring framing distinctions to the surface. In order to produce results that will be useful to the codebook development and coding team, however, we face not insignificant modeling challenges. In particular, much of current text modeling technology has been developed either for "shallow tasks" (e.g., Web-scale document retrieval and organization) or extraction of the propositional content (i.e., factual claims) from a piece of text into a structured form. Although there are many recent developments in analyzing subjectivity and sentiment in text, these tend to focus primarily on what computational linguists call polarity (Pang and Lee, 2008), i.e. positive/neutral/negative distinctions (cf. *tone*).

This means that significantly new developments will be required to apply natural language processing in studying framing. We have begun to experiment with promising approaches in two related areas. In the first, discussed in Section 3, we focus on identification of ideological type within candidate speeches; the method is used to test the hypothesis that successful candidates "move to the center," based on Downsian theory (Downs, 1957). In the second, summarized in Section 4, we develop a hierarchical Bayesian approach to frame discovery inspired by the view of framing as second-level agenda setting (McCombs, 2002). These preliminary models are consistent with the view of framing embodied in the Policy Frames Codebook, but not yet integrated with it; integrating methodological advances with our big-picture view of frame categories is a primary focus for our second year of the project.

3 Measuring Ideological Proportions

When seeking to quantify the use of frames in political communication, we believe it will be useful to take into account how frames correlate with the ideology of the speaker or author. Textual evidence for frame use and for ideology share key properties: lexical cues can be helpful but are often ambiguous, and reasoning about them will lead to better conclusions if we exploit contextual information and manage uncertainty with care. For these reasons, we conducted a study on ideology in text, in order to explore the possibilities for new statistical methods that might be appropriate for analysis of frames, ideology, and their interactions. We briefly review this work here; extended presentations may be found in Gross et al. (2013) and Sim et al. (2013).

The "median voter theorem" from Downsian theory (Black, 1948; Downs, 1957; Hotelling, 1929) predicts that presidential candidates should shift toward the general electorate's median voter after securing their parties' nominations.[2] Motivated by this common-sense though largely untested hypothesis, we test the theory using candidates' campaign speeches as data. This is accomplished in two stages. First, we automatically construct lexicons of ideologically-associated terms from a reference corpus of texts whose authors are strongly identified with known ideologies. We represent ideologies as discrete categories that can be organized hierarchically; see Figure 1.

The second stage makes use of these lexicons, representing a speech by a presidential candidate as a series of alternating "cues" (terms from the lexicons) and "lags" (sequences of uninteresting filler). Thus, we make no effort to classify all (or even most) of the tokens in a text; rather, we restrict focus to those phrases that are likely to be most informative about the perspective of the speaker. An example is shown in Figure 2.

Our approach assumes an unknown hidden Markov model over these terms, with each term emitted by a state corresponding to one of the ideologies (or "background") in Figure 1. We then

[2] This oversimplifies a bit. Strictly speaking, extending the logic of spatial modeling to multicandidate elections under the plurality rule does not lead to the same expectations (Cox, 1985; Feddersen et al., 1990). Primary elections are not, in fact, operating under the plurality rule, and, due to the complex multistage process by which candidates tend to pull out at different points in the nomination process, it is not obvious what should be anticipated under standard spatial models. Nonetheless, whatever the most strategic quantile pivot point may be, that of the Republican and Democratic primaries would naturally fall to the right and left, respectively, of the median during the general election.

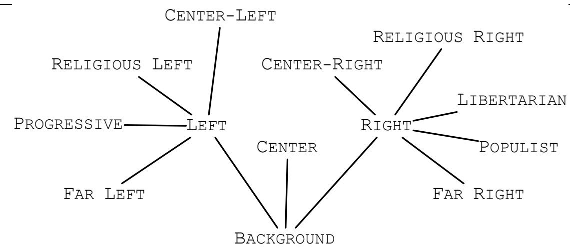

Figure 1: Ideology tree showing the labels for the ideological corpus and used in the HMM.

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	. . . $\xrightarrow{6}$ ronald_reagan $\xrightarrow{2}$ presid_reagan $\xrightarrow{3}$ econom_crisi $\xrightarrow{5}$ unemploy_rate $\xrightarrow{17}$ econom_growth $\xrightarrow{1}$ job_creation $\xrightarrow{9}$ . . .

Figure 2: Example of the cue-lag representation.

perform approximate Bayesian inference on this representation to infer a posterior distribution over each term instance's ideology state and, by extension, an estimate over the proportions of ideologies evoked in a speech. These can be further aggregated across speeches to create profiles for different candidates at different stages of each campaign; see Figure 3.

Two key assumptions are built into our model. The first is that states that are closer in the graph (Figure 1) are more likely to transition to each other. The ideology space, then, is not strictly categorical; there are notions of differing distance. The second assumption is that longer lags increase the probability of "restarting" the Markovian walk and forgetting the previous state. The Bayesian approach allows us to avoid many commitments, including (i) which lexicon terms associate with which ideologies, (ii) relative "distances" between the discrete ideologies, and (iii) the strength of a candidate's tendency to stay in the same state vs. move to another.

After performing validation and robustness checks, we fit the model using presidential candidates' speeches from 2008 and 2012. The results show that Barack Obama, John McCain and Mitt Romney did indeed make substantively significant shifts away from the ideological extremes after securing their parties' presidential nominations. Further, variants of our model that do not exploit the structured ideology space (Figure 1), or that weaken or strengthen the Markovian dependencies (by, respectively, always or never restarting) underperform our approach on a benchmark of checking preregistered hypotheses about candidates' associations with different ideologies.

Moving forward, the Policy Frames Codebook offers a starting point for developing a structured representation of the space of frames as well as textual descriptions of frames. Annotated examples can provide evidence in developing cue lexicons. The modeling methodology explored in this study provides a principled statistical framework for measurement of framing in text data while managing uncertainty.

4 A Hierarchical Bayesian Model of Framing as Second-Level Agenda Setting

In its concern with the subjects or issues under discussion in political discourse, agenda setting maps neatly to Bayesian topic modeling (Blei et al., 2003b) as a means of discovering and characterizing

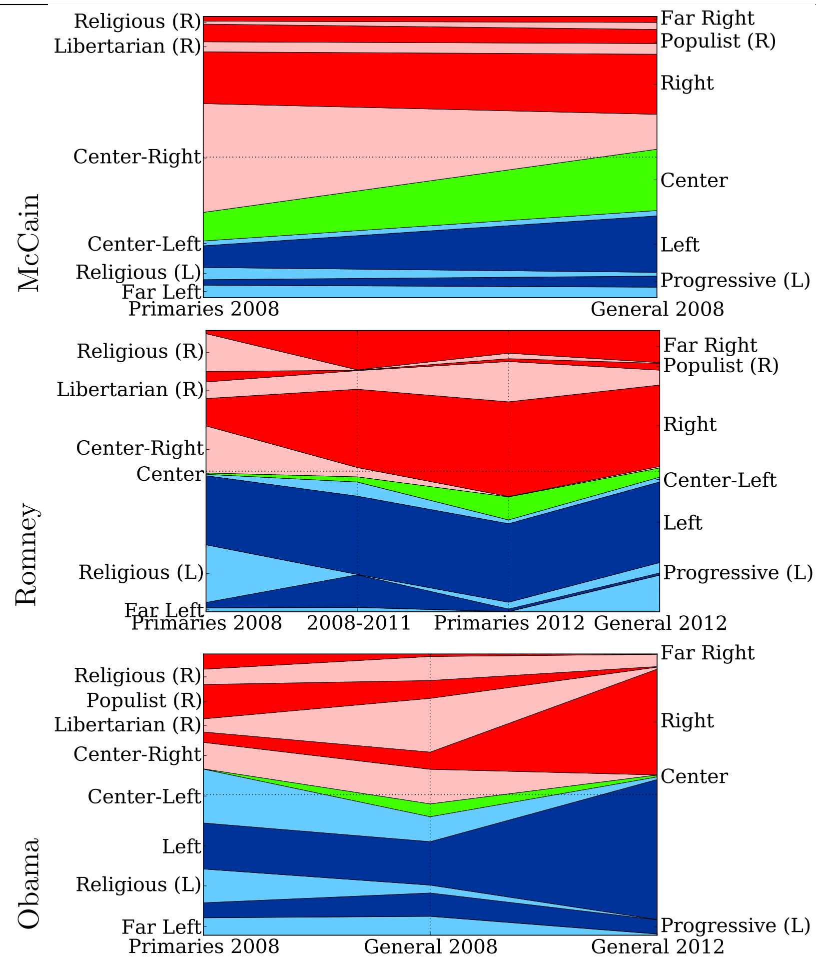

Figure 3: Proportion of time spent in each ideology by McCain, Romney, and Obama during the 2008 and 2012 Presidential election seasons.

those issues (e.g. Grimmer, 2010; Quinn et al., 2010). Interestingly, one line of communication theory seeks to unify agenda setting and framing by viewing frames as a second-level kind of agenda (McCombs, 2002): just as agenda setting is about which objects of discussion are considered salient, framing is about which *attributes* of those objects are considered salient. The key is that what communication theorists consider an attribute in a discussion about a topic can itself be an object (or sub-topic), as well. For example, the question of wrongful convictions is one attribute of the death penalty discussion, but it can also be viewed as an object of discussion in its own right. Similarly, drivers' licenses for non-registered immigrants is an object in its own right, while serving as an attribute of the larger immigration debate.

This two-level view leads naturally to the idea of using a hierarchical topic model to formalize both agendas and frames within a uniform setting, and in this project we have recently introduced a new model to do exactly that (Nguyen et al., 2013). Our approach falls into the general category of supervised topic models: it jointly captures topic structure along with a continuous-valued response variable that, in this setting, is used to represent of competing perspectives or ideology. In particular, our model, which we call *supervised hierarchical LDA* (SHLDA), simultaneously learns lexical and hierarchical regressions: the lexical regression discovers the *context-independent* effect of individual words in the lexicon on the response variable, and the hierarchical topic-based regression, inspired by

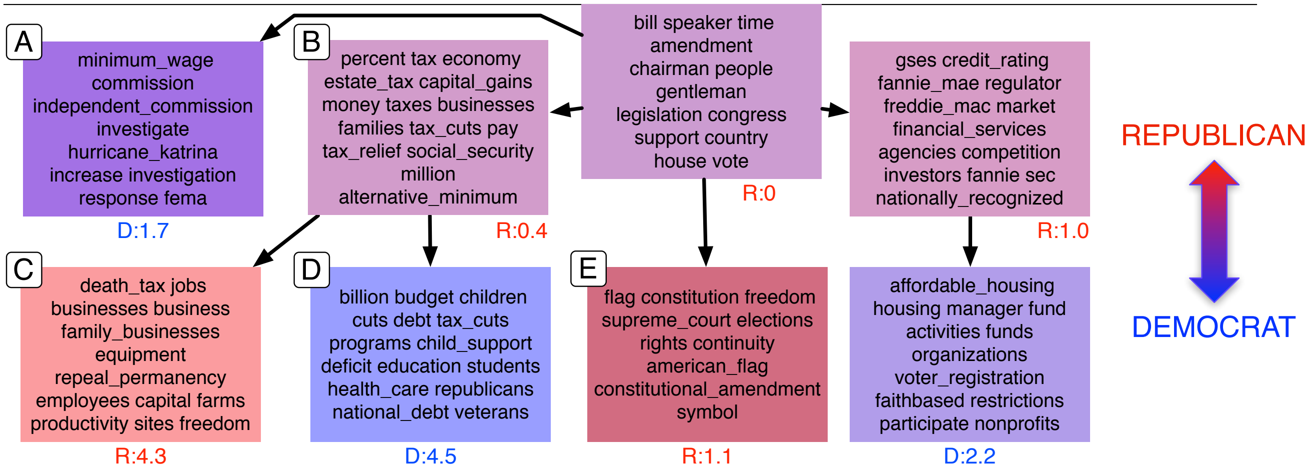

Figure 4: Fragment of an agenda/framing topic hierarchy discovered from Congressional floor debates. Many first-level topics are bipartisan (purple), while lower level topics are associated with specific political positions (Democrats blue, Republicans red). The number below each topic denotes the strength with which the model associates the topic with a position on the political spectrum.

hierarchical topic models Blei et al. (2003a), captures *context-specific* effect of second-level agendas.

To situate the model for readers not already familiar with the family of Bayesian topic models, the most popular such models today are variants of Latent Dirichlet Allocation (LDA, Blei et al., 2003b), which provides a way to automatically discover latent or implicit *topics* in otherwise unstructured collections of text. Blei and McAuliffe (2007) introduced *supervised* LDA, which allows that automatic topic-discovery process to be informed by an observed variable associated with each document, e.g. the political orientation of its author. The result is a flat set of topics that is often more interesting or useful than what LDA alone would produce, because the additional observed-variable information like author partisanship has been taken into account. Blei et al. (2003a) developed *hierarchical* LDA, a model that lacks extra observable information like partisanship, but which can automatically discover a hierarchy of topics, rather than a flat set, where the structure of the hierarchy is entirely determined by the text data. In the model we have introduced, we combine the two — supervised LDA and hierarchical LDA — to produce structured hierarchies of topics in a way that takes extra information, such as document-level ideology, into account.

To give a little bit more detail for those who are familiar with the background literature, at its core SHLDA extends hierarchical LDA (HLDA, Blei et al., 2003a) in the same way that supervised LDA (SLDA, Blei and McAuliffe, 2007) extended LDA. That is, like SLDA, SHLDA assumes that each document d is associated with a response variable y_d of interest, which can be any real-valued number. But, like HLDA, SHLDA assumes that words in documents are generated from topics associated with non-terminal nodes of a latent unbounded-branching tree, by using the nested Chinese Restaurant Process (nCRP) as the prior.[3] Sometimes, of course, words are strongly associated with extremes on the response variable continuum regardless of underlying topic structure. Therefore, in addition to hierarchical regression parameters, we include global *lexical regression parameters*, which model the interaction between specific words and the response variable. We also generalize HLDA by permitting documents to be assigned to multiple paths in the document hierarchy using sentence-level information, rather than just to a single path. Nguyen et al. (2013) describe SHLDA in full detail, along with formal evaluation.

[3] Although HLDA is capable of working with trees with infinite depth, in practice it is often truncated to a fixed depth, as is the case here.

As an example, Figure 4 illustrates part of a hierarchy of topics discovered automatically in an analysis Congressional floor debates, taking into account the political party of the person speaking. Blue is used to denote topics that are more strongly associated with Democrats, and red those more associated with Republicans. The root of the tree is a catch-all. The next level constitutes an automatically derived topical breakdown of the Congressional agenda; as such, many of its topics are purple (bipartisan). The level below that breaks agendas into, literally, second-level agendas, in the form of sub-topics where the data support that distinction. For example, the "tax" topic (B) is bipartisan, but it has two children that we can associate with distinct ideological frames connected with taxation: the Democratic-leaning child (D) focuses on the benefits to society provided by taxation ("child_support", "children", "education", "health_care"), while the Republican-leaning child (C) focuses on business-related impacts of taxation ("death_tax", "jobs", "family_businesses"). Crucially, this structure has been discovered automatically from a collection of documents (individual speaker turns on the floor of Congress), each one associated with an additional observed variable (political party of the person speaking), with no additional manual intervention.[4]

It is worth noting that including both hierarchical and lexical regression parameters in the model is important. For example, in the U.S. the term "liberty" tends to be an effective indicator of a conservative speaker regardless of context. But on the other hand, this is not true for "cost", which is a conservative-leaning indicator in environmental policy contexts but liberal-leaning in debates about foreign policy. SHLDA integrates both hierarchically context-sensitive and lexical, context-insensitive components within a single, unified model. These are closely related to the issue-specific and issue-general frame distinction of §2.

5 Conclusion

We have presented an overview of a project whose goal is to bring together the empirical study of framing with scalable, state of the art computational modeling of text. A key research activity during our first year has been the initial design and development of the Policy Frames Codebook, a theoretically motivated resource that provides both a general, cross-issues inventory of frame categories and the instantiation of those categories for a set of issues highly relevant in U.S. policy studies. In tandem with that resource development, we have begun developing statistical models that pertain to the core of framing: relating the language that politicans use to the underlying ideological or partisan positions that they hold or wish to emphasize. One model captures properties of language in relation to a manually curated structure of ideological categories. The other discovers agenda/frame-like structure automatically. In the coming year, we will be focused on refining these resources and models, integrating them into a human-in-the-loop process for frame discovery and identification, and investigating their utility as tools for political science domain experts.

Acknowledgments

The authors gratefully acknowledge project members and collaborators who contributed to this work: Brice Acree, Jordan Boyd-Graber, Viet-An Nguyen, Yanchuan Sim, and Kristina Victor. This research was supported in part by NSF IIS collaborative research funding under grants 1211201, 1211266, 1211277, and 1211153. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

[4]Examples like this are encouraging, but of course examples can only provide face validity. We have shown that the model's predictive power — e.g., predicting political party based on what the speaker said — significantly improves on prior models (Nguyen et al., 2013); more rigorous analysis of its potential for frame discovery is something we are now beginning to investigate.

References

Frank R. Baumgartner, Suzanna L. De Boef, and Amber E. Boydstun. *The Decline of the Death Penalty and the Discovery of Innocence*. Cambridge University Press, 2008.

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

David Blei and Jon McAuliffe. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. MIT Press, Cambridge, MA, 2007.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *NIPS*. MIT Press, 2003a. ISBN 0-262-20152-6.

David M. Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. *JMLR*, 3, 2003b.

Dennis Chong and James N. Druckman. Framing theory. *American Political Science Review*, 10: 103–126, 2007.

Gary W Cox. Electoral equilibrium under approval voting. *American Journal of Political Science*, pages 112–118, 1985.

Anthony Downs. *An Economic Theory of Democracy*. Harper, New York, 1957.

Robert Entman. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993.

Timothy J Feddersen, Itai Sened, and Stephen G Wright. Rational voting and candidate entry under plurality rule. *American Journal of Political Science*, pages 1005–1016, 1990.

Justin Grimmer. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1):1–35, 2010.

Justin Gross, Brice Acree, Yanchuan Sim, and Noah A. Smith. Testing the etch-a-sketch hypothesis: A computational analysis of Mitt Romney's ideological makeover during the 2012 primary vs. general elections, 2013. APSA 2013 Annual Meeting paper. Available at SSRN: `http://ssrn.com/abstract=2299991`.

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

Shanto Iyengar. *Is Anyone Responsible? How Television Frames Political Issues*. University of Chicago Press, 1991.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–291, 1979.

George Lakoff. *Don't Think of an Elephant! Know Your Values and Frame the Debate—The Essential Guide for Progressives*. Chelsea Green Publishing, White River Junction, NY, 2004.

George Lakoff. *Whose freedom?: the battle over America's most important idea*. Macmillan, 2006.

J. Matthes and M. Kohring. The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, 58(2):258–279, 2008.

Maxwell McCombs. The agenda-setting role of the mass media in the shaping of public opinion. *North*, 2009(05-12):21, 2002.

Thomas E. Nelson, Rosalee A. Clawson, and Zoe M. Oxley. Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, 91(3):567–583, 1997.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. Lexical and hierarchical topic regression. In *Neural Information Processing Systems*, 2013.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. How to analyze political attention with minimal assumptions and costs. *American Journal of*

Political Science, 54(1):209–228, 2010.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of ACL*, 2013.

Brian F Schaffner and Patrick J Sellers. *Winning with words: the origins and impact of political framing*. Routledge, 2009.

Yanchuan Sim, Brice Acree, Justin H. Gross, and Noah A. Smith. Measuring ideological proportions in political speeches. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*, October 2013.

B. Van Gorp. Where is the frame?. victims and intruders in the belgian press coverage of the asylum issue. *European Journal of Communication*, 20(4):484–507, 2005.