

ABSTRACT

Title of dissertation: **GUIDED PROBABILISTIC TOPIC MODELS
FOR AGENDA-SETTING AND FRAMING**

Viet-An Nguyen, Doctor of Philosophy, 2015

Dissertation directed by: Professor Philip Resnik
Department of Linguistics and
Institute for Advanced Computer Studies

Professor Jordan Boyd-Graber
College of Information Studies
Institute for Advanced Computer Studies

Probabilistic topic models are powerful methods to uncover hidden thematic structures in text by projecting each document into a low dimensional space spanned by a set of topics. Given observed text data, topic models infer these hidden structures and use them for data summarization, exploratory analysis, and predictions, which have been applied to a broad range of disciplines.

Politics and political conflicts are often captured in text. Traditional approaches to analyze text in political science and other related fields often require close reading and manual labeling, which is labor-intensive and hinders the use of large-scale collections of text. Recent work, both in computer science and political science, has used automated content analysis methods, especially topic models to substantially reduce the cost of analyzing text at large scale. In this thesis, we follow this approach and develop a series of new probabilistic topic models, *guided*

by additional information associated with the text, to discover and analyze *agenda-setting* (i.e., *what* topics people talk about) and *framing* (i.e., *how* people talk about those topics), a central research problem in political science, communication, public policy and other related fields.

We first focus on study agendas and agenda control behavior in political debates and other conversations. The model we introduce, *Speaker Identity for Topic Segmentation* (SITS), is able to discover what topics that are talked about during the debates, when these topics change, and a speaker-specific measure of agenda control. To make the analysis process more effective, we build *Argviz*, an interactive visualization which leverages SITS’s outputs to allow users to quickly grasp the conversational topic dynamics, discover when the topic changes and by whom, and interactively visualize the conversation’s details on demand. We then analyze policy agendas in a more general setting of political text. We present the *Label to Hierarchy* (L2H) model to learn a hierarchy of topics from multi-labeled data, in which each document is tagged with multiple labels. The model captures the dependencies among labels using an interpretable tree-structured hierarchy, which helps provide insights about the political attentions that policymakers focus on, and how these policy issues relate to each other.

We then go beyond just agenda-setting and expand our focus to framing—the study of how agenda issues are talked about, which can be viewed as *second-level agenda-setting*. To capture this hierarchical views of agendas and frames, we introduce the *Supervised Hierarchical Latent Dirichlet Allocation* (SHLDA) model, which jointly captures a collection of documents, each is associated with a contin-

uous response variable such as the ideological position of the document’s author on a liberal-conservative spectrum. In the topic hierarchy discovered by SHLDA, higher-level nodes map to more general agenda issues while lower-level nodes map to issue-specific frames. Although qualitative analysis shows that the topic hierarchies learned by SHLDA indeed capture the hierarchical view of agenda-setting and framing motivating the work, interpreting the discovered hierarchy still incurs moderately high cost due to the complex and abstract nature of framing. Motivated by improving the hierarchy, we introduce *Hierarchical Ideal Point Topic Model* (HIPTM) which jointly models a collection of votes (e.g., congressional roll call votes) and both the text associated with the voters (e.g., members of Congress) and the items (e.g., congressional bills). Customized specifically for capturing the two-level view of agendas and frames, HIPTM learns a two-level hierarchy of topics, in which first-level nodes map to an interpretable policy issue and second-level nodes map to issue-specific frames. In addition, instead of using pre-computed response variable, HIPTM also jointly estimates the ideological positions of voters on multiple interpretable dimensions.

GUIDED PROBABILISTIC TOPIC MODELS
FOR AGENDA-SETTING AND FRAMING

by

Viet-An Nguyen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:

Professor Philip Resnik, Chair/Co-Advisor
Professor Jordan Boyd-Graber, Co-Advisor
Professor Héctor Corrada Bravo
Professor Hal Daumé III
Professor Wayne McIntosh
Professor Hanna Wallach

© Copyright by
Viet-An Nguyen
2015

Acknowledgments

I am eternally indebted to the many people, whose help and support have made this dissertation possible. I can only hope that the following words can express just how grateful I am to have such amazing people in my life.

First and foremost, I wholeheartedly thank my co-advisors, Philip Resnik and Jordan Boyd-Graber, for their continuous guidance and advice throughout my graduate career. They have been an ideal team, providing as much support as I needed while giving me the freedom to explore plenty of sidetracks. In particular, I would like to thank Jordan for guiding me patiently through many technical challenges and for helping me tirelessly in preparing and presenting our work at various conferences. I am grateful to Philip for pushing me hard to make sure this thesis is coherent and well-written. I also very much appreciate his delicious dinners during his sabbatical.

Many thanks also go to the other members of my committee, Héctor Corrada Bravo, Hal Daumé III, Wayne McIntosh, and Hanna Wallach, for their insightful questions and valuable feedbacks, which provide new perspectives and help establish new connections to improve this thesis. I also like to thank Hal for his helpful comments on my various practice talks and for his excellent *Computational Linguistics I* course which introduced me to and got me interested in NLP. I also thank Hanna for her great dissertation, from which I have learned so much about hierarchical topic models and slice sampling.

I have been very fortunate to have the opportunities to work with wonderful

coauthors and collaborators. I thank Deborah Cai, Jennifer Midberry, and Yuanxin Wang for annotating excellent sets of data; Stephen Altschul for his great lectures and encouragements which allowed me to explore my way in an alien land called biology; and Kristina Miler for tirelessly providing helpful feedbacks and insights, without which Chapter 6 would not have been possible.

Various ideas in this thesis were shaped and influenced through the many conversations with my colleagues in the CLIP lab. Yuening Hu, Ke Zhai and Mohit Iyyer deserve a special mention for being such great friends and collaborators. I cherished my discussions with Amit Goyal, Vlad Eidelman, Ke Wu, Taesun Moon, Alvin Grissom II, Leonardo Claudino, Naho Orita, Thang Nguyen, Andy Garron, and Peter Enns. I enjoyed interacting with Eric Hardisty, Chang Hu, Kristy Hollingshead, Earl Wagner, Greg Sanders, John Morgan, Junhui Li, Jiarong Jiang, He He, Sudha Rao, Snigdha Chaturvedi, Jagadeesh Jagarlamudi, Ferhan Ture, and Ning Gao. I also want to especially thank Joe Webster, Janice Perrone, and other UMI-ACS staffs for their helps and supports.

During my PhD journey, I had the chance to spend two wonderful summers interning at Facebook. I am extremely grateful to all the people who helped make my time at Facebook such a blast including Cameron Marlow, Danny Ferrante, Sofus Macskassy, and other members of the Core Data Science team. My special thanks go to Jonathan Chang and Carlos Diuk for their guidance and mentorship and Nicole Gurries for opening up the opportunity for me in the first place.

Of course, my past few years would have been much more difficult and boring without my dear friends: Bao Nguyen, Chanh Kieu, My Le, Ha Nguyen, Tho Ho,

Toan Ngo, Hien Tran, Hien Nguyen, and Dzung Ta. Thank you all.

Finally, I want to thank my family, although I know there aren't enough words to describe my gratitude to them. I thank my parents for their unconditional love; for always believing in me and teaching me to believe in myself; for having sacrificed greatly to provide for me an amazing education and let me pursue my dreams. I thank my sister for her endless support, advice and encouragement. And most importantly, I thank my beloved wife Yen, who has always been there with me, sharing every moment, through all the highs and the lows in this adventure. I know, together, we could overcome any obstacles and accomplish anything.

Table of Contents

| | |
|---|-----|
| List of Tables | ix |
| List of Figures | xii |
| 1 Introduction | 1 |
| 1.1 The Needs for Automated Methods for Analyzing Political Text in the Big Data Era | 1 |
| 1.2 The Importance of Agendas and Frames in Political Science Research | 4 |
| 1.3 Analyzing Agendas and Frames: Methods and Costs | 5 |
| 1.3.1 Human Coding | 8 |
| 1.3.2 Supervised Learning | 9 |
| 1.3.3 Topic Modeling | 11 |
| 1.4 Automated Content Analysis at Lower Cost | 12 |
| 1.4.1 Analyzing Agendas and Agenda Control in Political Debates | 15 |
| 1.4.2 Learning Agenda Hierarchy from Multi-labeled Legislative Text | 16 |
| 1.4.3 Discovering Agendas and Frames Discovery in Ideologically Polarized Text | 18 |
| 1.4.4 Discovering Agendas and Frames from Roll Call Votes and Congressional Floor Debates | 20 |
| 1.5 Main Technical Contributions | 22 |
| 2 Probabilistic Topic Modeling Foundations | 23 |
| 2.1 Latent Dirichlet Allocation: The Basic Topic Model | 24 |
| 2.2 Beyond LDA: Topic Modeling Extensions | 27 |
| 2.2.1 Using Bayesian Nonparametrics | 28 |
| 2.2.2 Incorporating Metadata | 33 |
| 2.2.3 Adding Hierarchical Structure | 37 |
| 2.2.4 Other extensions | 39 |
| 2.3 MCMC Inference and the Importance of Averaging | 40 |
| 2.3.1 Learning and Predicting with MCMC | 41 |
| 2.3.2 MCMC in Topic Modeling | 42 |
| 2.3.3 Averaging Strategies | 43 |

| | | |
|-------|--|-----|
| 2.3.4 | Unsupervised Topic Models | 44 |
| 2.3.5 | Supervised Topic Models | 48 |
| 2.3.6 | Discussion and Conclusion | 52 |
| 3 | Agenda Control in Political Debates | 53 |
| 3.1 | Introduction | 53 |
| 3.1.1 | Presidential Debates: Unique Setting for Agenda Control | 54 |
| 3.1.2 | Agenda Control to Influence in Multi-party Conversations | 55 |
| 3.1.3 | Topic Segmentation to Capture Conversational Structures | 58 |
| 3.1.4 | Chapter Structure | 60 |
| 3.2 | SITS: Speaker Identity for Topic Segmentation | 60 |
| 3.2.1 | Overview of our Approach | 61 |
| 3.2.2 | Generative Process of SITS | 63 |
| 3.2.3 | Posterior Inference for SITS | 66 |
| 3.3 | Data Collections and Annotations | 73 |
| 3.3.1 | Datasets | 74 |
| 3.4 | Evaluating Agenda Control | 79 |
| 3.4.1 | 2008 Election Debates | 80 |
| 3.4.2 | <i>Crossfire</i> | 82 |
| 3.4.3 | 2012 Republican Primary Debates | 85 |
| 3.5 | Detecting Influencers in Conversations | 87 |
| 3.5.1 | Influencer Annotation | 87 |
| 3.5.2 | Computational Methods for Influencer Detection | 90 |
| 3.5.3 | Influencer Detection Problem | 92 |
| 3.5.4 | Experimental Setup | 95 |
| 3.5.5 | Results and Analysis | 97 |
| 3.6 | Evaluating Topic Segmentation | 98 |
| 3.6.1 | Experiment Setups | 98 |
| 3.6.2 | Results and Analysis | 102 |
| 3.7 | <i>Argviz</i> : Interactive Visualization of Topic Dynamics in Conversations | 103 |
| 3.8 | Conclusions and Future Work | 106 |
| 4 | Learning Agenda Hierarchy from Multi-labeled Political Text | 108 |
| 4.1 | Introduction | 108 |
| 4.1.1 | Analyzing Agendas in Legislative Text | 109 |
| 4.1.2 | Topic Models for Multi-labeled Documents | 114 |
| 4.1.3 | Chapter Structure | 116 |
| 4.2 | L2H: Capturing Label Dependencies using a Tree-structured Hierarchy | 116 |
| 4.2.1 | Creating the Label Graph | 118 |
| 4.2.2 | Generating Tree-structured Hierarchy | 119 |
| 4.2.3 | Generating Documents | 120 |
| 4.3 | Posterior Inference | 123 |
| 4.3.1 | Initialization | 123 |
| 4.3.2 | Sampling Assignments $x_{d,n}$ and $z_{d,n}$ | 123 |
| 4.3.3 | Sampling Topics ϕ | 124 |

| | | |
|-------|--|-----|
| 4.3.4 | Updating tree structure \mathcal{T} | 125 |
| 4.4 | Analyzing Political Agendas in U.S. Congresses | 127 |
| 4.5 | Document Modeling and Classification | 131 |
| 4.5.1 | Document modeling | 131 |
| 4.5.2 | Multi-label Classification | 134 |
| 4.6 | Summary | 136 |
| 5 | Discovering Agendas and Frames in Ideologically Polarized Text | 139 |
| 5.1 | Introduction | 139 |
| 5.1.1 | Framing: Going beyond Agenda-setting to Understand How Things are Talked About | 140 |
| 5.1.2 | Framing as Second-level Agenda-setting | 142 |
| 5.1.3 | Framing Research: Traditional vs. Data-driven Approach | 144 |
| 5.1.4 | Chapter Structure | 145 |
| 5.2 | SHLDA: Capturing Text and Continuous Response using Hierarchical Topic Structure | 146 |
| 5.2.1 | Generating Text | 147 |
| 5.2.2 | Generating Responses | 150 |
| 5.3 | Posterior Inference and Optimization | 151 |
| 5.4 | Data: Congress, Products, Films | 156 |
| 5.4.1 | U.S. congressional floor debates: | 156 |
| 5.4.2 | Amazon product reviews | 158 |
| 5.4.3 | Movie reviews | 159 |
| 5.5 | Qualitative Analysis of Topic Hierarchies | 159 |
| 5.6 | Quantitative Prediction of Document Responses | 161 |
| 5.7 | Conclusion | 165 |
| 5.7.1 | Summary | 165 |
| 5.7.2 | Discussions and Future Directions | 166 |
| 6 | Discovering Agendas and Frames from Roll Call Votes and Text | 168 |
| 6.1 | Introduction | 168 |
| 6.1.1 | A Brief Overview of Ideal Point Models | 170 |
| 6.1.2 | On the Dimensionality of Ideal Points | 172 |
| 6.1.3 | Scaling Multi-dimensional Ideal Points using Votes and Text | 174 |
| 6.1.4 | Tea Party in the House | 176 |
| 6.1.5 | Main Contributions | 177 |
| 6.2 | Hierarchical Ideal Point Topic Model | 179 |
| 6.2.1 | Defining the Topic Hierarchy | 182 |
| 6.2.2 | Generating Congressional Speeches | 184 |
| 6.2.3 | Generating Bill Text | 185 |
| 6.2.4 | Generating Roll Call Votes | 185 |
| 6.3 | Posterior Inference | 187 |
| 6.3.1 | Sampling Issue Assignments for Bill Tokens | 187 |
| 6.3.2 | Sampling Frame Assignments for Speech Tokens | 188 |
| 6.3.3 | Sampling Issue Topics | 189 |

| | | |
|-------|--|-----|
| 6.3.4 | Sampling Frame Proportions | 189 |
| 6.3.5 | Optimizing Frame Regression Parameters | 190 |
| 6.3.6 | Updating Ideal Points, Polarity and Popularity | 190 |
| 6.4 | Analyzing Tea Party Ideal Points | 190 |
| 6.4.1 | Data Collection | 190 |
| 6.4.2 | One-dimensional Ideal Points | 191 |
| 6.4.3 | Multi-dimensional Ideal Points | 195 |
| 6.5 | Agendas and Frames: Analyzing Topic Hierarchy | 202 |
| 6.5.1 | Analyzing Agenda Issues | 202 |
| 6.5.2 | Analyzing Issue-specific Frames | 204 |
| 6.6 | Predicting Tea Party Caucus Membership | 209 |
| 6.6.1 | Membership Prediction given Votes and Text | 210 |
| 6.6.2 | Membership Prediction given Text Only | 211 |
| 6.7 | Conclusion and Future Directions | 213 |
| 6.7.1 | Summary | 213 |
| 6.7.2 | Discussion and Future Directions | 214 |
| 7 | Conclusion and Future Work | 216 |
| 7.1 | Summary of Contributions | 216 |
| 7.2 | Directions for Future Work | 219 |

List of Tables

| | | |
|-----|--|----|
| 1.1 | Summary of four models introduced in this thesis with their estimated costs. | 13 |
| 2.1 | Example of ten topics discovered by LDA from a collection of floor debates in U.S. Congress. | 28 |
| 3.1 | Summary of datasets detailing how many distinct speakers are present, how many distinct conversations are in the corpus, the annotations available, and the general content of the dataset. | 73 |
| 3.2 | Example turns from the 2008 election debates annotated by Boydston et al. (2013a) . Each clause in a turn is manually coded with a <i>Question Topic Code</i> (T_Q) and a <i>Response Topic Code</i> (T_R). The topic codes (T_Q and T_R) are from the Policy Agendas Topics Codebook. In this example, the following topic codes are used: Macroeconomics (1), Housing & Community Development (14), Government Operations (20). | 75 |
| 3.3 | List of the 9 Republican Party presidential debates used. | 77 |
| 3.4 | Example of a Wikipedia discussion in our dataset. | 78 |
| 3.5 | Example of turns designated as a topic shift by SITS. We chose turns to highlight speakers with high topic shift tendency π . Some keywords are manually italicized to highlight the topics discussed. | 83 |
| 3.6 | Top speakers by topic shift tendencies from our <i>Crossfire</i> dataset. We mark hosts (\dagger) and “speakers” who often (but not always) appeared in video clips (\ddagger). ANNOUNCER makes announcements at the beginning and at the end of each show; NARRATOR narrates video clips; MALE and FEMALE refer to unidentified male and female respectively; QUESTION collectively refers to questions from the audience across different shows. Apart from those groups, speakers with the highest tendency were political moderates. | 84 |
| 3.7 | Statistics of the two datasets <i>Crossfire</i> and Wikipedia discussions that we annotated influencers. We use these two datasets to evaluate SITS on influencer detection. | 96 |

| | | |
|-----|--|-----|
| 3.8 | Influencer detection results on <i>Crossfire</i> and Wikipedia discussion pages. For both datasets, topic-change-based methods (\star) outperform structure-based methods (\diamond) by large margins. For all evaluation measurements, higher is better. | 97 |
| 3.9 | Results on the topic segmentation task. Lower is better. The parameter k is the window size of the metrics P_k and WindowDiff chosen to replicate previous results. | 100 |
| 4.1 | Major topics with their corresponding codes in the Policy Agendas Topics codebook. [†] The major topic “Immigration” was newly added to the codebook in 2014. | 111 |
| 4.2 | Examples of bills from the 100 th Congress, coded by the Congressional Bills project. The mapping of the Policy Agenda (PA) major topic codes are provided in Table 4.1. | 111 |
| 4.3 | Examples of multiple labels provided by the Congressional Research Service for the three bills shown in Table 4.2 | 112 |
| 5.1 | Notation used in this chapter | 152 |
| 5.2 | Top words based on the global lexical regression coefficient, τ . For the floor debates, positive τ ’s are Republican-leaning while negative τ ’s are Democrat-leaning. | 162 |
| 5.3 | Regression results for Pearson’s correlation coefficient (PCC, higher is better (\uparrow)) and mean squared error (MSE, lower is better (\downarrow)). Results on Amazon product reviews and movie reviews are averaged over 5 folds. Subscripts denote the number of topics for parametric models. For SVM-LDA-VOC and MLR-LDA-VOC, only best results across $K \in \{10, 30, 50\}$ are reported. Best results are in bold | 163 |
| 6.1 | Example voting records of legislators in the 112 th House of Representatives. A legislator might not vote on a bill, which is denoted by ‘-’ in this table. | 171 |
| 6.2 | Words with highest weights in the priors ϕ_k^* for 19 Policy Agendas Topics, estimated by using labeled data from the Congressional Bills Project. | 183 |
| 6.3 | Key votes having “Government operations” as the most probable issue, estimated by our model. The last column shows the estimated probability $\vartheta_{b,k}$. Each key vote is shown with a short description, the preferred voting position of Freedom Works (Y for Yea, N for Nay), the number of Republicans whose votes agree and disagree with Freedom Works (‘All’ denotes all voting Republican legislators, ‘TP’ denotes Tea Party Caucus members, and ‘NTP’ denotes non-Tea Party Caucus members). Bolded key votes are the ones on which the majority of the two groups vote differently. | 199 |

| | | |
|-----|---|-----|
| 6.4 | Key votes having “Macroeconomics” as the most probable issue, estimated by our model. The last column shows the estimated probability $\vartheta_{b,k}$. Each key vote is shown with a short description, the preferred voting position of Freedom Works (Y for Yea, N for Nay), the number of Republicans whose votes agree and disagree with Freedom Works (‘All’ denotes all voting Republican legislators, ‘TP’ denotes Tea Party Caucus members, and ‘NTP’ denotes non-Tea Party Caucus members). Bolded key votes are the ones on which the majority of the two groups vote differently. | 200 |
| 6.5 | Key votes having “Transportation” as the most probable issue, estimated by our model. The last column shows the estimated probability $\vartheta_{b,k}$. Each key vote is shown with a short description, the preferred voting position of Freedom Works (Y for Yea, N for Nay), the number of Republicans whose votes agree and disagree with Freedom Works (‘All’ denotes all voting Republican legislators, ‘TP’ denotes Tea Party Caucus members, and ‘NTP’ denotes non-Tea Party Caucus members). Bolded key votes are the ones on which the majority of the two groups vote differently. Both of these votes focus on the federal spending | 201 |
| 6.6 | Words with highest probabilities for each first-level issue nodes learned by HIPTM. | 203 |

List of Figures

| | | |
|-----|---|----|
| 1.1 | Three main classes of text categorization methods and their corresponding costs (Quinn et al., 2010). | 7 |
| 2.1 | Density plots for four Dirichlet distributions. The densities are over the triangular simplex that represents multinomial distributions over three dimensions and demonstrate how different Dirichlet distributions can accommodate variable concentrations. Darker coloring denotes higher probability density. (a) Dirichlet parameters that are all 1.0 yield a uniform density over multinomial distributions. (b) Dirichlet parameters that are all greater than 1.0 yield a density concentrated near the mean distribution \mathbf{p} , in this case (0.6250, 0.0625, 0.3125). (c) and (d) Dirichlet parameters that are all less than 1.0 yield a density concentrated near the edges and corners of the simplex. Such a density favors sparse multinomial distributions. | 25 |
| 2.2 | Generative process and the plate diagram representation of LDA. In the diagram, nodes represent random variables (shaded ones are observed, clear ones are latent), directed edges are probabilistic dependencies, and plates represents repetition. | 27 |
| 2.3 | Illustration of the stick breaking process $\boldsymbol{\pi} \sim \text{GEM}(\alpha_0)$, in which $\pi_k = \pi'_k \prod_{i=1}^{k-1} (1 - \pi'_i)$ is defined based on the fraction $\pi'_k \sim \text{Beta}(1, \alpha_0)$ that is taken from the remainder of the stick after each break. | 31 |
| 2.4 | Illustration of the Chinese restaurant process metaphor in which there are seven customers currently occupying three tables. A new customer coming in will sit at (1) an existing table with a probability proportional to the number of customers currently sitting at the table or (2) a new table with a probability proportional to α . The exact probabilities are shown inside each table. | 32 |
| 2.5 | Generative process and the plate diagram representation of HDP. | 33 |
| 2.6 | Generative process and the plate diagram of Labeled LDA. | 34 |
| 2.7 | Generative process and the plate diagram representation of sLDA. | 36 |

| | | |
|------|---|----|
| 2.8 | Illustration of training and test chains in MCMC, showing samples used in four prediction strategies studied in this section: Single Final (SF), Single Average (SA), Multiple Final (MF), and Multiple Average (MA). | 42 |
| 2.9 | Perplexity of LDA using different averaging strategies with different number of training iterations T_{TR} . Perplexity generally decreases with additional training iterations, but the drop is more pronounced with multiple test chains. | 47 |
| 2.10 | Performance of sLDA using different averaging strategies computed at each training iteration. | 48 |
| 2.11 | Performance of sLDA using different averaging strategies computed at the final training iteration T_{TR} , compared with two baselines MLR and SVR. Methods using multiple test chains (MF and MA) perform as well as or better than the two baselines, whereas methods using a single test chain (SF and SA) perform significantly worse. | 49 |
| 3.1 | Plate diagrams of our proposed models: (a) nonparametric SITS; (b) parametric SITS. Nodes represent random variables (shaded nodes are observed); lines are probabilistic dependencies. Plates represent repetition. The innermost plates are turns, grouped in conversations. | 63 |
| 3.2 | Diagram of notation for topic shift indicators and conversation segments: Each turn is associated with a latent binary variable <i>topic shift indicator</i> l specifying whether the topic of the turn is shifted. In this example, topic shifts occur in turns τ and $\tau' + 1$. As a result, the topic shift indicators of turn τ and $\tau' + 1$ are equal to 1 (i.e. $l_{c,\tau} = l_{c,\tau'+1} = 1$) and the topic shift indicators of all turns in between are 0 (i.e. $l_{c,t} = 0, \forall t \in [\tau + 1, \tau']$). Turns $[\tau, \tau']$ form a segment s in which all topic distributions $G_{c,\tau}, G_{c,\tau+1}, \dots, G_{c,\tau'}$ are the same and are denoted collectively as $G_{c,s}$. | 65 |
| 3.3 | Illustration of topic assignments in our inference algorithm. Each solid rectangle represents a restaurant (i.e., a topic distribution) and each circle represents a table (i.e., a topic). To assign token n of turn t in conversation c to a table $z_{c,t,n}$ in the corpus-level restaurant, we need to sample a path assigning the token to a segment-level table, the segment-level table to a conversation-level table and the conversation-level table to a globally shared corpus-level table. | 67 |

| | | |
|-----|---|-----|
| 3.4 | Illustration of minimal path assumption. This figure shows an example of the seating assignments in a hierarchy of Chinese restaurants of a higher-level restaurant and a lower-level restaurant. Each table in the lower restaurant is assigned to a table in the higher restaurant and tables on the same path serve the same dish k . When sampling the assignment for table ψ_2^L in the lower restaurant, given that dish $k = 2$ is assigned to this table, there are two options for how the table in the higher restaurant could be selected. It could be an existing table ψ_2^H or a new table ψ_{new}^H , both serving dish $k = 2$. Under the minimal path assumption, it is always assigned to an existing table (if possible) and only assigned to a new table if there is no table with the given dish. In this case, the minimal path assumption will assign ψ_2^L to ψ_2^H | 70 |
| 3.5 | Topic shift tendency π of speakers in the 2008 Presidential Election Debates (larger means greater tendency). IFILL was the moderator in the vice presidential debate between BIDEN and PALIN; BROKAW, LEHRER and SCHIEFFER were the moderators in the three presidential debates between OBAMA and MCCAIN; QUESTION collectively refers to questions from the audiences. Colors denote Republicans , Democrats, Moderators, and Audiences | 80 |
| 3.6 | Topic shift tendency π of speakers in the 2012 Republican Primary Debates (larger means greater tendency). KING, BLITZER and COOPER are moderators in these debates; the rest are candidates. | 86 |
| 3.7 | The <i>Argviz</i> user interface consists of <i>speaker panel</i> (A), <i>transcript panel</i> (B), <i>heatmap</i> (C), <i>topic shift column</i> (D), <i>topic cloud panel</i> (E), <i>selected topic panel</i> (F). | 103 |
| 4.1 | Generative process and the plate diagram notation of L2H. | 118 |
| 4.2 | Example of the weighted directed graph \mathcal{G} and a spanning tree \mathcal{T} generated from \mathcal{G} , created from a set of multi-labeled data having three unique labels: HEALTH CARE, HEALTH CARE COVERAGE AND ACCESS, and MEDICARE. The thickness of an directed edge represents its weight. | 120 |
| 4.3 | Illustration of the <i>candidate set</i> and the <i>complementary set</i> of a document tagged with two labels: HIGHER EDUCATION and MEDICARE | 121 |
| 4.4 | Example of different ways to define the candidate set \mathcal{L}_1 (shaded nodes) and the complementary set \mathcal{L}_0 (white nodes) for a document tagged with two labels (double-circled nodes). | 122 |
| 4.5 | Number of bills and unique labels in our dataset after pre-processing for each Congress. | 128 |
| 4.6 | A subtree rooted at INTERNATIONAL AFFAIRS in the hierarchy learned by L2H using data from the 112 th Congress. | 129 |

| | | |
|------|---|-----|
| 4.7 | A subtree rooted at ENVIRONMENTAL ASSESSMENT, MONITORING, RESEARCH in the hierarchy learned by L2H using data from the 112 th Congress. | 130 |
| 4.8 | A subtree rooted at HEALTH in the hierarchy learned by L2H using data from the 112 th Congress. | 130 |
| 4.9 | Perplexity on held-out documents, averaged over 5 folds (lower is better). | 133 |
| 4.10 | Multi-label classification results. The results are averaged over 5 folds. | 136 |
| 5.1 | Example hierarchy with ideologically polarized topics that SHLDA learns. First-level nodes map to agenda issues, while second-level nodes map to issue-specific frames. Each node is associated with a topic (i.e., a multinomial distribution over words) and an ideological score. | 143 |
| 5.2 | Plate notation diagram of our SHLDA model. | 146 |
| 5.3 | Illustration of SHLDA’s restaurant franchise metaphor. | 149 |
| 5.4 | Distributions of the response variables in the three datasets. | 157 |
| 5.5 | Topics discovered from Congressional floor debates. Many first-level topics are bipartisan (purple), while lower level topics are associated with specific ideologies (Democrats blue, Republicans red). For example, the “tax” topic (B) is bipartisan, but its Democratic-leaning child (D) focuses on social goals supported by taxes (“children”, “education”, “health care”), while its Republican-leaning child (C) focuses on business implications (“death tax”, “jobs”, “businesses”). The number below each topic denotes the magnitude of the learned regression parameter associated with that topic. Colors and the numbers beneath each topic show the regression parameter η associated with the topic. | 159 |
| 5.6 | Topics discovered from Amazon reviews. Higher topics are general, while lower topics are more specific. The polarity of the review is encoded in the color: red (negative) to blue (positive). Many of the first-level topics have no specific polarity and are associated with a broad class of products such as “routers” (Node D). However, the lowest topics in the hierarchy are often polarized; one child topic of “router” focuses on upgradable firmware such as “tomato” and “ddwrt” (Node E, positive) while another focuses on poor “tech_support” and “customer_service” (Node F, negative). The number below each topic is the regression parameter learned with that topic. | 161 |
| 6.1 | Overview of HIPTM’s outputs: (1) first-level nodes map to policy issues, each of which corresponds to a major topic in the Policy Agendas Topics codebook, (2) second-level nodes map to issue-specific frames, and (3) each frame node and each lawmaker are associated with an issue-specific ideological position. | 179 |
| 6.2 | Plate notation diagram of our HIPTM model. | 181 |

| | | |
|------|---|-----|
| 6.3 | Box plots of the estimated one-dimensional Tea Party ideal points for members and non-members of the Tea Party Caucus among Republican Representatives in the 112 th U.S. House. The median of members' ideal points is significantly higher than that of non-members' ideal points, though there are a lot of overlaps between the two distributions. | 192 |
| 6.4 | Republican legislators having the (a) lowest and (b) highest estimated one-dimensional ideal points. | 193 |
| 6.5 | Boxplots of ideal points on 19 dimensions, each of which corresponds to a major topic in the Policy Agendas Codebook estimated by our model. On most issues the ideal point distributions over the two Republican groups (member vs. non-member of the Tea Party Caucus) overlap significantly. The most polarized issues are 'Government Operations' and 'Macroeconomics', which align well with the agenda of the Tea Party movement supporting small government and lower taxes. | 196 |
| 6.6 | Subtree on "Macroeconomics" learned by our model. | 205 |
| 6.7 | Subtree on "Health" issue in the topic hierarchy learned by our model. | 206 |
| 6.8 | Subtree on the "Labor, Employment and Immigration" issue in the topic hierarchy learned by our model. | 207 |
| 6.9 | Tea Party Caucus membership prediction results over five folds using AUC-ROC (higher is better, random baseline achieves 0.5). The features extracted from our model are estimated using both the votes and the text. | 210 |
| 6.10 | Tea Party Caucus membership prediction results over five folds using AUC-ROC (higher is better, random baseline achieves 0.5). The features extracted from our model for unseen legislators are estimated using their text only. | 212 |

Chapter 1: Introduction

1.1 The Needs for Automated Methods for Analyzing Political Text in the Big Data Era

To open up the report in response to President Barack Obama’s request for a 90-day review on how technologies affect our lives in May 2014, Counselor to the President John Podesta and other senior government officials wrote:

“We are living in the midst of a social, economic, and technological revolution. How we communicate, socialize, spend leisure time, and conduct business has moved onto the Internet. The Internet has in turn moved into our phones, into devices spreading around our homes and cities, and into the factories that power the industrial economy. The resulting explosion of data and discovery is changing our world.” (Podesta et al., 2014).

We are indeed living in an era of *big data*, in which the proliferation of data from both digital and analog sources, together with ever-faster computing machines and ever-larger data storage, have brought researchers unique opportunities to study problems in *computational social science* at an unprecedented scale and granularity (Watts, 2007; Lazer et al., 2009; Giles, 2012; Wallach, 2014). For example, the

availability of large social networks over time allows us to study in detail their structural properties and patterns (Leskovec, 2008). Online social networks, in which individuals are able to share information with their peers, enable work to study how information diffuses, propagates and influences through the networks (Bakshy et al., 2012; Bond et al., 2012; Weng, 2014). Fine-grained human behavioral data provide invaluable opportunities to analyze and predict human traits (Vespignani, 2009; Kosinski et al., 2013), infer relational dynamics of individuals (Eagle et al., 2009) and uncover opinions and sentiments (Pang and Lee, 2008; Liu, 2012).

Among those available data, text is arguably one of the most pervasive and persistent sources of information for social science research. *Content analysis of text* has been a major approach for social scientists to study human behaviors for centuries: from at least as early as the late 1600s when the Church examined printed text to detect non-religious materials which were considered a threat to its authority, until today when, for example, close reading of textual contents such as open-ended survey responses provides invaluable insights into respondent’s own thinking (Krippendorff, 2012). However, with the availability of voluminous amount of text today, traditional content analysis methods such as close reading and manual coding become infeasible. This problem raises the need for *automated content analysis of text*, which draws on techniques from natural language processing, machine learning, data mining and information retrieval to analyze text data at large scale (Grimmer and Stewart, 2013; O’Connor, 2014). The outputs of automated methods can potentially (1) provide insights that might not be possible to achieve by close reading, (2) improve the coding schemes created manually by domain experts, and (3) support

rigorous methods for prediction and forecasting.

Within political science, there have recently been significant efforts to bring together researchers from political science, computer science, linguistics and other related fields to develop computer-assisted approaches to analyze political text. For example, the *Special Issue on Text Annotation for Political Science Research* of the *Journal of Information Technology & Politics* “solicits and publishes papers that provide a clear view of the state-of-the-art in text annotation and evaluation, especially for political science” (Cardie and Wilkerson, 2008). The *Political Analysis* journal’s *Special Issue on The Statistical Analysis of Political Text* publishes work on automated methods that are “directed toward specific applications in the study of politics, such as determining ideological position from texts, coding political interactions, and identifying the content of political conflict” (Monroe and Schrodtt, 2008). The annual conference on *New Directions in Analyzing Text as Data*, jointly sponsored by the Ford Center for Global Citizenship at Northwestern University, the Department of Methodology at the London School of Economics, and the Institute for Quantitative Social Science at Harvard University, has provided an unique venue for researchers in multi-disciplinary fields to present and exchange latest work on applying automated content analysis methods to a diverse set of applications and problems in political science.¹

¹Work done in this dissertation has contributed to talks in this conference including: *Interactive Modeling of Large Datasets and Discovering Topic Influencers* (2011), *“I Want to Talk About, Again, My Record On Energy...”: Modeling Control of the Topic in Political Debates and Other Conversations* (2012) and *Identifying Media Frames and Frame Dynamics within and across Policy Issues* (2013).

1.2 The Importance of Agendas and Frames in Political Science Research

In this thesis, we focus on developing novel automated content analysis techniques for studying *agendas* and *frames* in political text—a central research area in political science for decades (Schattschneider, 1960; McCombs and Shaw, 1972; Baumgartner and Jones, 1993b; Wolfe et al., 2013). Political agenda, as defined by Baumgartner (2001), “is the set of issues that are the subject of decision making and debate within a given political system at any one time”. Examples of political agenda issues include economy, education, health, defense, foreign affairs and transportation. Political agenda-setting, or the ability to influence the salient topics or issues, has been the focus of much research in both political communication and policy studies (Wolfe et al., 2013). Studying political agendas helps shed light on key questions about political systems such as: What are the issues that get more attention by the policymakers, the media and the public? How do these attentions change over time, and why?

Scholars in political communication have mainly focused on how the media influence public agenda (McCombs and Shaw, 1993; McCombs, 2005). In his book *The Press and Foreign Policy*, Cohen (1963) argued that “the press may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to think about”. Another seminal work on this subject is the 1968 Chapel Hill study, in which McCombs and Shaw (1972) showed

that there was a very strong correlation between what 100 residents of Chapel Hill, North Carolina thought was the most important election issue and what the local and national news reported was the most important issue. Researchers in policy studies, on the other hand, focus on *policy agenda-setting*, which emphasizes the political attention of government elites and policymakers (Rogers and Dearing, 1988; Baumgartner and Jones, 1993b; Rogers et al., 1993; Jones and Baumgartner, 2005).

If agenda-setting emphasizes on *what* issues are talked about, the question of *how* things are talked about concerns *framing*. “*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). By highlighting a particular perspective or interpretation and deemphasizing others, it is widely accepted that framing can have significant influence on public opinions towards important policy issues (Chong and Druckman, 2007; Nelson et al., 1997; Boydston et al., 2013c). For example, the rise of the “innocence frame” in the death penalty debate, emphasizing the irreversible consequence of mistaken convictions, has led to a sharp decline in the use of capital punishment in the U.S. (Baumgartner et al., 2008).

1.3 Analyzing Agendas and Frames: Methods and Costs

One central step in *analyzing agendas* in political text is to understand the political attention mentioned in the text, or in other words, to uncover *what* topics

are talked about. A popular approach to tackle this question is *text categorization*—classifying the text of interest into one or more discrete topical categories, each of which maps to an agenda issue. Various methods for text categorization have been introduced in the literature, each one has its own costs and benefits. Following [Quinn et al. \(2010\)](#), we discuss three main classes of text categorization methods (*human coding*, *supervised learning*, and *topic modeling*), differentiated by three types of costs that can incur at three stages of the analysis:²

- *Pre-analysis cost* is the cost incurred before the actual categorization happens where “conceptualization and operationalization are dealt with”. Methods with high pre-analysis cost are the ones that require human with substantive knowledge to prepare the data for the text categorization to happen.
- *Analysis cost* is the cost incurred during the categorization of the text of interest happens. Methods with high analysis cost are the ones that require humans to spend more time per text unit to categorize.
- *Post-analysis cost* is the cost incurred after the categorization where the results are assessed and interpreted. Methods with high post-analysis cost are the ones that require humans to spend more time analyzing the results. Results that are incoherent and uninterpretable also increase this cost.

²[Quinn et al. \(2010\)](#) further split each cost type into (1) domain knowledge required and (2) human time taken. We simplify our analysis by considering only an overall cost for each cost type. They also discuss five text categorization methods with *reading-based method* and *dictionary-based method* being the additional two. We do not discuss these two methods since reading-based method is inarguably impractical for large-scale corpus and dictionary-based method has very similar costs with supervised learning method. An additional dimension considered in ([Quinn et al., 2010](#)) is the set of assumptions each method makes, which is not the focus of our analysis here.

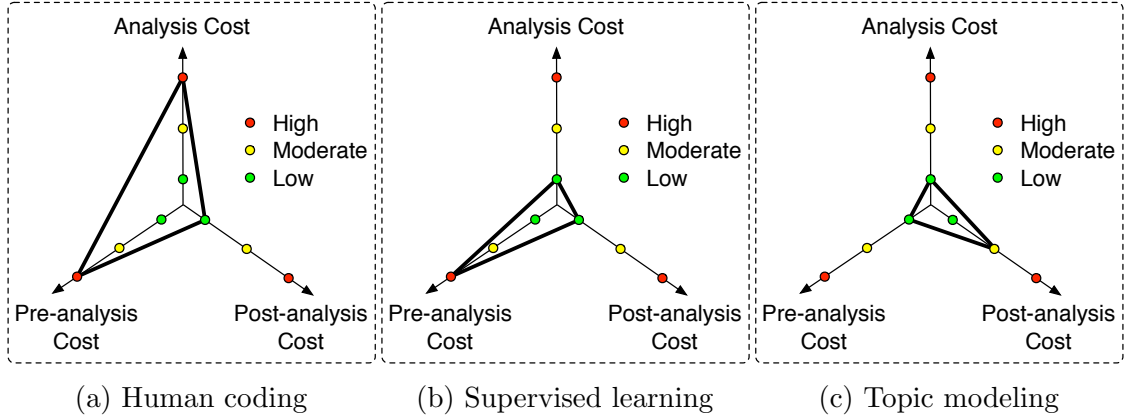


Figure 1.1: Three main classes of text categorization methods and their corresponding costs (Quinn et al., 2010).

Figure 1.1 illustrates the three methods with their corresponding costs, which are evaluated using a three-level rating scheme: *High*, *Moderate*, and *Low*.

If analyzing agendas concerns what topics are talked about, *analyzing frames* concerns *how* these topics are talked about. Identifying and categorizing frames is, however, a much more challenging task. The major reason is because framing is abstract. Boydston et al. (2013c) assert that “the very definition of framing has been notoriously slippery”, for which Entman (1993) called framing a “fractured paradigm”. Interestingly, one line of political communication theory seeks to unify agenda-setting and framing by viewing frames as *second-level agendas* (McCombs et al., 1997; Ghanem, 1997): just as agenda-setting is about which issues (or topics) of discussion are salient, framing is about the salience of aspects (or subtopics) of those issues. This two-level view leads naturally to the idea of using a *hierarchical categorization of topics*, which we use in this thesis to discover and analyze frames and framing.

In the remainder of this section, we review the three existing text categorization methods for political text and their corresponding costs.

1.3.1 Human Coding

Human coding is a standard *manual content analysis* method applied to the problem of identifying issues or topics in political text. It usually consists of (1) defining a coding system by domain experts (e.g., a set of diverse, well-defined political issues), (2) training human coders, and (3) coding the documents of interest manually. In practice, defining the codebook often involves an iterative process where coding issues are conceptualized and repeatedly refined through several pilot studies until a final coding system is achieved. Since the codebook needs to be defined and human coders need to be trained before the actual human coding can happen, the costs in both pre-analysis and analysis phases of this approach are high. However, this approach often provides highly interpretable coding systems, together with a very high quality set of coded documents, which makes its post-analysis cost relatively low (Figure 1.1a).

One of the most successful work following this approach is arguably the *Policy Agendas Project* led by Baumgartner and Jones (1993a), which defines a codebook of 19 major topics and 225 subtopics.³ The codebook has been used extensively to code and study policy agendas in documents from Congress, Supreme Court, news media like the New York Times and various Public Opinion and Interest Groups (John,

³<http://www.policyagendas.org/>

2006).⁴ Specifically built for the U.S. Congress, the *Congressional Bill Project*, led by Adler and Wilkerson (2006) provides an extensive set of more than 400,000 bills coded using the Policy Agendas Topics codebook. In addition, following its success in studying the U.S. political system, the Policy Agendas Project has also been extended and built upon for the European Union (*EU Policy Agendas Project*)⁵ and individual European countries (*Comparative Agendas Project*).⁶

Identifying and coding frames is, as argued above, much more challenging due to the abstract nature of framing. Despite the challenges, Boydston et al. (2013c) have worked on an ambitious project to define a *Policy Frames Codebook*, which consists of 14 categories of frames and an “Other” category. Examples of the frame categories in the codebook include “Economic”, “Capacity & Resources”, “Morality & Ethics” and “Fairness & Equality”. These frame categories are intended to be applicable across multiple policy issues (e.g., abortion, immigration, tobacco, marriage equality etc), just as the Policy Agendas Codebook provides a consistent system for categorizing topics across policy agendas.

1.3.2 Supervised Learning

With the increasing availability of political text, however, the cost of manually coding documents has become impractical. Many recent efforts focus on using *automated content analysis* approach to reduce the analysis cost and trade off between

⁴See <http://www.policyagendas.org/page/datasets-codebook> for examples of datasets coded using the Policy Agendas Topic codebook. The number of books and papers from research programs using this codebook are “too numerous to cite here” (Quinn et al., 2010, p. 210).

⁵<http://www.policyagendas.eu/>

⁶<http://www.comparativeagendas.info/>

the pre-analysis and post-analysis costs.

One relatively straightforward approach to automation for this type of problem is *supervised learning*—leveraging existing supervised learning techniques from machine learning, data mining and related fields. For example, Purpura and Hillard (2006) and Hillard et al. (2008) describe the automated classification system used in the Congressional Bills Project, in which Support Vector Machines (SVMs) and other machine learning techniques are used to classify legislative text into one of the 226 subtopics in the Policy Agendas Topics codebook. Kwon et al. (2007) also use standard supervised learning methods to classify political claims into one of the predefined classes of opinion. A similar approach has also been used to classify German online political news (Scharkow, 2013) and Dutch election manifestos (Verberne et al., 2014).

The supervised learning techniques still require labeled data for training. Thus, similar to the human coding, the supervised learning method still has high pre-analysis cost. However, after the training phase, the learned classifier can be used to automatically label new data, which reduces the analysis cost significantly (Figure 1.1b). A promising approach to reduce the pre-analysis cost of labeling training data is *active learning*, which requires human to label only a subset of the data while still achieving comparable classification accuracy (Hillard et al., 2007; Purpura et al., 2008).

1.3.3 Topic Modeling

Topic modeling is also an *automated content analysis* method which has gained exponential popularity in analyzing political text in particular, and in discovering thematic structure of large-scale text corpus in general (Blei, 2012). Introduced by Blei et al. (2003b), Latent Dirichlet allocation (LDA)—the original unsupervised topic model—extends previous latent variable models including LSA (Deerwester et al., 1990), LSI (Papadimitriou et al., 1998), and PLSA (Hofmann, 1999) and assumes that words in each documents are generated from a mixture of *topics*, each of which is a multinomial distribution over a fixed vocabulary of words. Inferred from the word co-occurrence in documents, each learned topic typically represents a coherent theme which often maps to an *issue* when studying political agendas. Recent work following this approach include applying topic models to examine the agenda in the U.S. Senate from 1997 to 2004 (Quinn et al., 2010), estimating category proportions in opinions about the U.S. presidency (Hopkins and King, 2010), and measuring the expressed agendas in the Senate press releases (Grimmer, 2010).

LDA and other unsupervised extensions require no training data and thus have a low pre-analysis cost. However, each topic is just a multinomial distribution over the vocabulary, which is often represented by a list of words with highest probabilities. If this word list is incoherent, which is not uncommon (Chang et al., 2009b; Mimno et al., 2011; Lau et al., 2014a), the resulting categorized text might be hard to interpret. This difficulty, therefore, raises the cost of the post-analysis phase (Figure 1.1c).

1.4 Automated Content Analysis at Lower Cost

In this thesis, following the automated content analysis approach, we introduce *novel topic models*, which are *guided* by additional information associated with the text and designed to discover and analyze agendas and frames in political discourses at *a lower cost*. We extend existing topic models, an active research area that we will review in Chapter 2, to improve the models’ interpretability, which as a result reduces the post-analysis cost, by

- modeling jointly the text and some *metadata* of interest which are readily available at no additional cost and can guide the model to learn more interpretable topics. Examples of these metadata include *speaker identity* in Chapter 3, *authors’ ideological score* in Chapter 5, and *roll call votes* in Chapter 6.
- using labeled data that incur a lower pre-analysis cost than the traditional approach of creating well-defined coding systems by domain expert, training human coders and coding document manually as described in Section 1.3.1. Examples include using *multi-labeled data* in Chapter 4 and using *existing coded data* as prior in Chapter 6.

Table 1.1 summarizes the four models introduced in this thesis together with their estimated costs.

We first focus on discovering and analyzing agendas in two settings: political debates in Chapter 3 and Congressional bill text in Chapter 4. For political debates, we present an unsupervised nonparametric topic model which incorporates

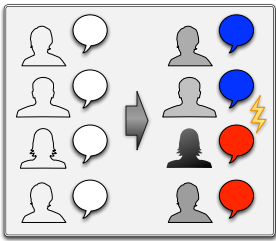
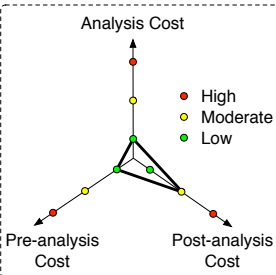
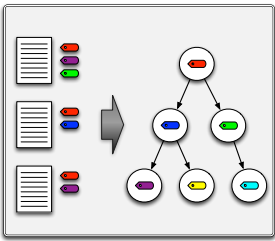
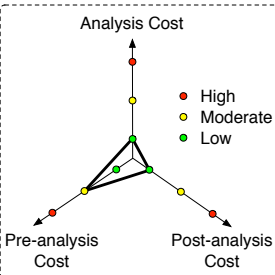
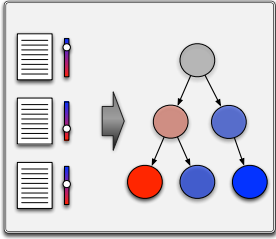
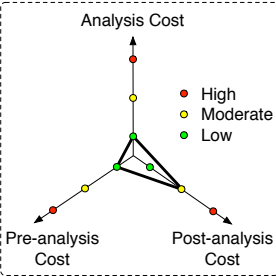
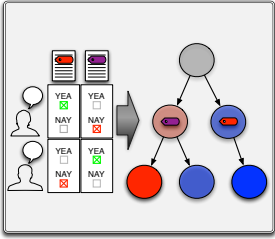
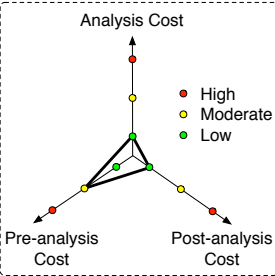
| | Without agenda labels | With agenda labels |
|------------------|--|--|
| Agendas only | <p>SITS (Chapter 3)</p>   | <p>L2H (Chapter 4)</p>   |
| Agendas & Frames | <p>SHLDA (Chapter 5)</p>   | <p>HIPTM (Chapter 6)</p>   |

Table 1.1: Summary of four models introduced in this thesis with their estimated costs.

the speaker identity to analyze agendas and agenda control behaviors of participating individuals. For Congressional bill text, we leverage multi-labeled data, which arguably can be obtained at a lower cost than traditional human coded data, to learn a label hierarchy capturing the dependencies among agenda issues. The learned hierarchy is highly interpretable which helps reduce the post-analysis cost.

Next, we introduce two nonparametric Bayesian hierarchical topic models to study agendas and frames in Congressional floor debates in Chapters 5 and 6. As discussed in Section 1.3, we follow the line of political communication theory that views framing as second-level agenda-setting and discover a hierarchy of topics to analyze agendas and frames. In these learned topic hierarchies, first-level topics map to agenda issues while second-level topics map to issue-specific frames. More specifically, in Chapter 5 we learn the topic hierarchy by jointly modeling the text with their authors' ideological scores. The model uses the metadata ideological scores to guide the discovery of issue-specific frames that are ideologically polarized. However, since no agenda or frame labels are used, the model is still subject to the limitation of unsupervised topic models: the interpretability of the learned topics. Since the topics are organized hierarchically, we estimate that incoherent and uninterpretable topics incur even more post-analysis cost compared with flat-structured unsupervised topics. In Chapter 6, we learn the topic hierarchy from text and roll call votes. By leveraging existing agenda-labeled data as priors, we learn a more interpretable topic hierarchy which, by design, reduces the post-analysis cost.

In the remainder of this chapter, we briefly introduce the four models we will present in subsequent chapters. For each model, we will give an overview of the

problems and applications that the model tries to tackle, how we evaluate it and what its analysis costs are qualitatively. **Although using the the cost-based analysis as a guiding framework across the thesis, our focus in this thesis is not explicitly quantifying the cost but actually developing and applying the models on analyzing agenda-setting and framing in various settings.** Measuring the actual analysis cost of the models developed in this thesis and other automated content analysis methods is an interesting direction for future work which we will discuss in Chapter 7.

1.4.1 Analyzing Agendas and Agenda Control in Political Debates

We first focus on studying agendas in *political debates*. Political debates, especially presidential debates, play a central role in U.S. politics. A key question that has attracted much research in political science on presidential debates is: *How do candidates control the agenda of the debates?* In Chapter 3, we introduce SITS—*Speaker Identity for Topic Segmentation*—a nonparametric Bayesian model that takes into account the speaker identity to discover (1) what topics that are talked about during the debate, (2) when these topics change, and (3) a speaker-specific measure of “agenda control”. We apply our model SITS to qualitatively analyze the agendas and agenda control behaviors of candidates in 2008 election debates and 2012 Republican primary debates.

Being an unsupervised topic model, SITS enjoys *low pre-analysis cost*, but is subject to a *moderately high post-analysis cost* due to the interpretability of the

learned topics. To mitigate this problem, we build an interactive visualization, *Argviz*, which help analysts better access and interpret the outputs of SITS. Using *Argviz* to visualize SITS’s output, domain experts can quickly examine the topical dynamics of the debates, discern when the topic changes and by whom, and interactively visualize the debate’s details on demand.

Although motivated by studying political debates, SITS is also applicable to a much broader setting, *turn-taking multi-party conversations*. In addition to political debates, we also use SITS to analyze business meetings (ICSI meeting transcripts), online discussions (Wikipedia discussion pages) and TV political talk show (CNN’s *Crossfire*). To quantitatively evaluate our model, we conduct extensive experiments on two tasks: topic segmentation and influencer detection. For topic segmentation, the task to divide conversations into smaller and topically coherent segments, SITS outperforms previous models which does not explicitly capture the speaker identity. For the second task, we manually annotate influencers in two datasets: *Crossfire* and Wikipedia. Empirical results show that features extracted from SITS outperforms traditional methods to detect influencer significantly.

1.4.2 Learning Agenda Hierarchy from Multi-labeled Legislative Text

We then transition from studying agendas in the conversational setting to legislative text in the U.S. Congress, the focus of much political science research including the Congressional Bills project and other related research programs.⁷ In

⁷See <http://www.congressionalbills.org/research.html> for some research related to the Congressional Bills project.

the Congressional Bills project, each bill, based on its title of the introduced version, is manually coded with a major topic and a subtopic in the Policy Agendas Topics codebook. Although coding in this way has several advantages such as achieving high inter-annotator reliability and enabling comparisons with other forms of policymaking activity (hearings, laws, executive orders etc) coded using the same codebook, it also incurs a high pre-analysis cost. In Chapter 4, we study agendas in legislative text using the set of labels provided by the Congressional Research Service, in which each bill is coded with *multiple agenda issues*.⁸ The motivations for using this type of *multi-labeled data* include

- First, each bill can be about more than one issue. In the descriptions about its coded data, the Congressional Bills project notes that “researchers should not assume that every bill relating to ‘air pollution’ (for example) will be found among the ‘705’ bills. A bill could address air pollution but be primarily among something else”.^{9,10}
- Second, it is relatively cheaper to code the data using multiple labels since it allows a more flexible coding system: the list of labels needs not to be very well defined beforehand by domain experts and can be accumulatively extended as new labels might be used by human coders.

One drawback of this labeling approach is that the label space can be relatively large, which makes learning the model, predicting labels for new documents and

⁸<http://thomas.loc.gov/help/terms-subjects.html>

⁹<http://www.congressionalbills.org/codebooks.html>

¹⁰In the Policy Agendas Topics codebook, ‘705’ is the code for the subtopic ‘Air pollution, Global Warming, and Noise Pollution’.

interpreting results more difficult. To tackle these problems, we propose L2H—*Label to Hierarchy*—a hierarchical topic model that captures the dependencies among labels by using a tree-based topic hierarchy. By associating each label with a topic (i.e., a multinomial distribution over the vocabulary), L2H learns an interpretable topic hierarchy which provides a natural mechanism for integrating user knowledge and data-driven summaries in a single, consistent structure. We apply L2H to analyze policy agendas in legislative texts in four U.S. Congresses (109th–112th). By using multi-labeled data, we reduce the pre-analysis cost of traditional supervised learning method described in Section 1.3.2, while enjoying its low post-analysis cost, especially with the aid of the learned hierarchy. Moreover, our empirical experiments also show that, using the topic hierarchy can improve the prediction performance in two quantitative tasks: predicting words in held-out documents and predicting multiple labels for unseen text.

1.4.3 Discovering Agendas and Frames Discovery in Ideologically Polarized Text

Going beyond agenda-setting (i.e., *what* topics people talk about), we expand our focus to framing (i.e., *how* they talk about different issues). In its concern with the subjects or issues under discussion in political discourse, agenda-setting maps neatly to topic modeling as a means of discovering and characterizing those issues (Grimmer, 2010; Quinn et al., 2010). Interestingly, one line of political communication theory seeks to unify agenda setting and framing by viewing frames as a

second-level agenda-setting (McCombs et al., 1997; Ghanem, 1997): just as agenda setting is about which objects of discussion are salient, framing is about the salience of attributes of those objects. The key is that what communications theorists consider an attribute in a discussion can itself be an object, as well. For example, mistaken convictions is one attribute of the death penalty discussion, but it can also be viewed as an object of discussion in its own right.

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. In Chapter 5, we present SHLDA—*Supervised Hierarchical latent Dirichlet allocation*—to do exactly that. SHLDA discovers a hierarchy of topics from text, in which the first-level topics map to agenda issues while second-level topics map to issue-specific frames. Using no agenda or frame labels, SHLDA requires low pre-analysis cost, but is also subject to the limitation of unsupervised topic models: the interpretability of the learned topics (Section 1.3.3). Since the topics are organized hierarchically, incoherent and uninterpretable topics might incur even more post-analysis cost compared to flat-structured unsupervised topics. To mitigate this problem, SHLDA jointly models the text—transcribed from Congressional floor debates and the *authors’ ideological scores*—estimated using legislators’ roll call votes.¹¹ By incorporating these metadata, we can discover ideologically polarized frames, which helps improve the interpretability of the learned topic hierarchy, and thus reduce the post-analysis cost.

¹¹The ideological score is the DW-NOMINATE scores obtained from http://voteview.com/dwnomin_joint_house_and_senate.htm.

In addition, the model is predictive: it represents the idea of alternative or competing perspectives via a *continuous-valued response variable*. Although inspired by the study of political discourse, associating texts with “perspectives” is more general and has been studied in various settings such as sentiment analysis (Paul and Girju, 2010; Jo and Oh, 2011) and discovery of regional variation (Ahmed and Xing, 2010a; Eisenstein et al., 2011). We show that the learned hierarchical structure improves prediction of perspective in both a political domain and on sentiment analysis tasks, and we argue that the topic hierarchies exposed by the model are indeed capturing structure in line with the theory that motivated the work.

1.4.4 Discovering Agendas and Frames from Roll Call Votes and Congressional Floor Debates

Similar to SITS, SHLDA learns the set of topics (but organized in a tree-based hierarchy instead) without any topic labels. This provides an exploratory tool to discover agendas and frames jointly without the high pre-analysis cost of manual coding, but still suffers from the high post-analysis cost of interpreting the results. To mitigate the problem, we design SHLDA to capture jointly the text and some continuous response of interest associated with each document. Particularly, in the setting of analyzing agenda-setting and framing from political text, we discover agendas and frames which are polarized on the liberal-conservative spectrum by using DW-NOMINATE—a commonly used score estimated from voting records of lawmakers to approximate their positions, or often called *ideal points*, on a single

dimension of ideology.

In Chapter 6, we introduce HIPTM—a *Hierarchical Ideal Point Topic Model* to discover agendas and frames by jointly modeling a set of votes in the U.S. Congress and the text associated with both the legislators (e.g., congressional speeches) and the bills (e.g., the bill text). HIPTM is different from SHLDA in the following ways. First, we specifically design HIPTM with a two-level hierarchical structure in which first-level nodes map to agenda issues and second-level nodes map to issue-specific frames. Second, we leverage existing labeled data from the Congressional Bills Project to build topic priors (i.e., multinomial distributions over words) for the issue nodes in the hierarchy, each of which maps to one of the 19 major topics in the Policy Agendas Topics Codebook. Third, instead of using pre-computed ideal point like DW-NOMINATE, HIPTM jointly estimate multi-dimensional ideal points of legislators, in which each dimension maps to one of the 19 interpretable topics mentioned above.

We apply HIPTM to analyze how legislators vote and talk similarly or differently in the U.S. Congress with respect to the *Tea Party movement*, a recent American political movement which has attracted much attentions from both public media as well as academic scholars. Using HIPTM, we analyze the difference in language uses (via discovered issue-specific frames) and voting behaviors (via estimated multi-dimensional ideal points) on various policy agenda issues of members of the Tea Party Caucus—the first institutional organization to the Tea Party movement, in comparison with other Republican legislators with no membership with the caucus.

1.5 Main Technical Contributions

Even though all models introduced in this thesis are motivated by specific problems in political science, the models themselves are typically applicable to other more general settings. Besides providing new computational tools for identifying and analyzing policy agendas and issue frames in large-scale political text, this thesis makes the following technical contributions to machine learning and natural language processing:

- Study and empirically compare different sample combination strategies when using MCMC inference for topic models for predictions (Chapter 2)
- Introduce a new nonparametric Bayesian topic model which incorporates speaker identity to capture topics, topic changes, and agenda control behavior in multi-party conversations (Chapter 3)
- Extend current state-of-the-art topic models for multi-labeled documents to learn an interpretable topic hierarchy when the label space is large (Chapter 4)
- Add novel extensions to state-of-the-art nonparametric hierarchical topic models to learn topic hierarchies jointly from text and other metadata of interest including continuous response variables such as the ideal points capturing the position of lawmakers on the liberal-conservative spectrum (Chapter 5) and binary matrices such as the voting records in the U.S. Congress (Chapter 6)

We also summarize in more detail the contributions of this thesis including both the technical contributions of each model introduced and their applications in Chapter 7.

Chapter 2: Probabilistic Topic Modeling Foundations

Probabilistic topic models are powerful methods to uncover hidden thematic structures in text by projecting each document into a low dimensional space spanned by a set of *topics*, each of which is a distribution over words. Given the observed data, topic models discover these hidden structures through posterior inference, and use them for a wide range of applications including data summarization, exploratory analysis, and predictions (Blei, 2012, 2014). In this thesis, as motivated in Chapter 1, we follow this line of research and introduce novel topic models to study agendas and frames in political discourses as well as other related applications. This chapter provides relevant background on this active research area.

In Section 2.1, we first review *Latent Dirichlet Allocation* (Blei et al., 2003b, LDA), the basic topic model which provides the foundation for topic modeling research. Since its introduction in 2003, LDA has become the building block of numerous extensions, many of which motivate the models we introduce in this thesis. We survey these extensions in Section 2.2. In Section 2.3, we review *Markov chain Monte Carlo* (MCMC)—a popular posterior inference technique for topic models, and discuss the importance of averaging over multiple samples when using MCMC for predictions, which is theoretically motivated but often glossed over in practice.

The discussion in Section 2.3 goes beyond reviewing background to include experimentation that makes a new empirical contribution.

Parts of this chapter are based on the materials originally published in [Nguyen et al. \(2013a, 2014a\)](#).

2.1 Latent Dirichlet Allocation: The Basic Topic Model

Introduced by [Blei et al. \(2003b\)](#), Latent Dirichlet Allocation (LDA) is a probabilistic model whose goal is to find a lower dimensional representation of a collections of documents that is useful for various tasks including classification, summarization and exploratory analysis. Following and extending previous latent variable models such as LSA ([Deerwester et al., 1990](#)), LSI ([Papadimitriou et al., 1998](#)), and PLSA ([Hofmann, 1999](#)), the intuition behind LDA is that each document exhibits multiple topics. LDA captures this intuition by assuming that each document is a mixture over a finite number of latent *topics*, each of which is a probabilistic distribution over a vocabulary.

Before describing LDA in detail, we first review the *Dirichlet distribution*, a central concept used in LDA and many of its extensions.

Dirichlet distribution is a distribution over finite discrete probability distributions. A Dirichlet distribution, denoted by $\text{Dirichlet}(\boldsymbol{\alpha})$, is parameterized by a vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ of non-negative real numbers. The density of $\text{Dirichlet}(\boldsymbol{\alpha})$ over a

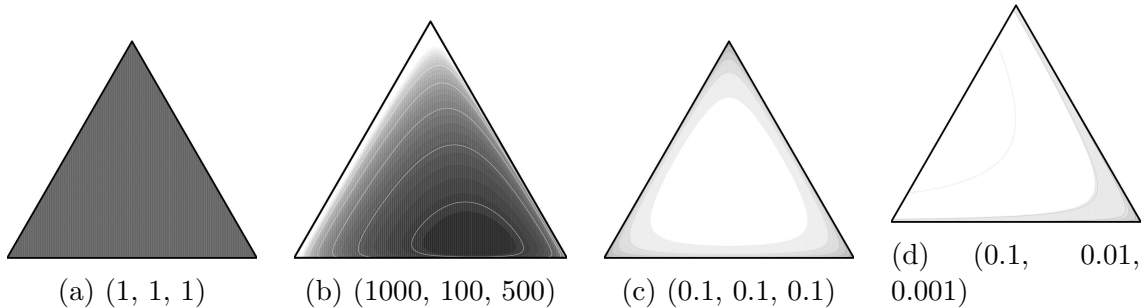


Figure 2.1: Density plots for four Dirichlet distributions. The densities are over the triangular simplex that represents multinomial distributions over three dimensions and demonstrate how different Dirichlet distributions can accommodate variable concentrations. Darker coloring denotes higher probability density. (a) Dirichlet parameters that are all 1.0 yield a uniform density over multinomial distributions. (b) Dirichlet parameters that are all greater than 1.0 yield a density concentrated near the mean distribution \mathbf{p} , in this case $(0.6250, 0.0625, 0.3125)$. (c) and (d) Dirichlet parameters that are all less than 1.0 yield a density concentrated near the edges and corners of the simplex. Such a density favors sparse multinomial distributions.

probability distribution $\mathbf{x} = (x_1, \dots, x_K)$ is

$$p(\mathbf{x} | \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K x_k^{\alpha_k - 1} \quad (2.1)$$

A Dirichlet distribution can also be parameterized by a concentration parameter $\alpha > 0$ and a mean probability distribution \mathbf{p} . This two-parameter Dirichlet distribution, denoted by $\text{Dirichlet}(\alpha, \mathbf{p})$, is equivalent to $\text{Dirichlet}(\boldsymbol{\alpha})$ if we define $\alpha = \sum_{k=1}^K \alpha_k$ and $\mathbf{p} = \boldsymbol{\alpha}/\alpha$. When the mean distribution \mathbf{p} is a uniform distribution, the Dirichlet distribution is called *symmetric* and often denoted by $\text{Dirichlet}(\alpha)$. Figure 2.1 illustrates examples of the probability densities defined by different Dirichlet distributions.

Latent Dirichlet allocation (LDA) takes as input a set of D documents, in which word tokens $\{\mathbf{w}_d\}_{d=1}^D$ are from a vocabulary of V unique word types. LDA posits that there are K shared topics, each of which is a multinomial distribution over the vocabulary drawn from a Dirichlet distribution prior. Figure 2.2 shows the generative process of LDA and its plate notation diagram.

More specifically, the multinomial ϕ_k of topic k is a distribution over words

$$p(\phi_k | \beta) = \frac{\Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v)} \prod_{v=1}^V \phi_{k,v}^{\beta_v - 1} \quad (2.2)$$

In addition, each document d is modeled as a multinomial distribution θ_d over the K topics, also drawn from a Dirichlet prior

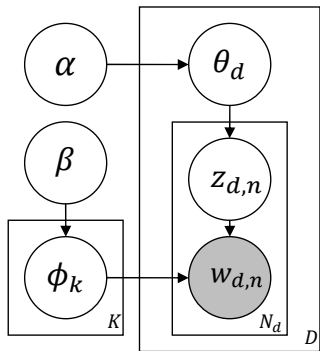
$$p(\theta_d | \alpha) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_{d,k}^{\alpha_k - 1} \quad (2.3)$$

To generate a token in document d , first we draw a topic assignment $z_{d,n}$ from the document-specific multinomial $\theta_{z_{d,n}}$. Given the chosen topic, we draw the word token $w_{d,n}$ from the corresponding multinomial $\phi_{z_{d,n}}$.

The joint probability distribution of a set of documents \mathbf{w} and their topic assignments \mathbf{z} is

$$p(\mathbf{w}, \mathbf{z} | \theta, \phi; \alpha, \beta) = \prod_{k=1}^K p(\phi_k | \beta) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^{N_d} p(z_{d,n} | \theta_d) p(w_{d,n} | \phi_{z_{d,n}}) \quad (2.4)$$

Given the observable documents, through posterior inference, LDA estimates the global topics $\{\hat{\phi}_k\}$ capturing what are talked about *globally* in the whole corpus,



1. For each topic $k \in [1, K]$
 - Draw word distribution $\phi_k \sim \text{Dirichlet}(\beta)$
2. For each document $d \in [1, D]$
 - Draw topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$
 - For each token $n \in [1, N_d]$
 - Draw topic $z_{d,n} \sim \text{Multinomial}(\theta_d)$
 - Draw word $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$

Figure 2.2: Generative process and the plate diagram representation of LDA. In the diagram, nodes represent random variables (shaded ones are observed, clear ones are latent), directed edges are probabilistic dependencies, and plates represents repetition.

and the document-specific topic proportions $\{\hat{\theta}_d\}$ capture what salient topics are talked about *locally* in each document. Table 2.1 shows, as an example, ten topics learned by LDA from a set of floor debates in the 109th U.S. Congress. Here, each topic is represented by a list of words which have the highest probabilities in that topic. As we can see, each word list provides a relatively coherent theme, which can be mapped to a policy agenda issue, e.g., “Immigration” (Topic 1), “Economic” (Topic 2), “Foreign Trade” (Topic 3) etc.

2.2 Beyond LDA: Topic Modeling Extensions

Since its introduction, LDA has become the building block for numerous topic modeling extensions (Blei et al., 2010a; Blei, 2012, 2014). Surveying all probabilistic models that extend and adapt LDA goes beyond the scope of this chapter, but in this section, we provide a brief survey on several directions for extending LDA that motivates various parts of the models presented in this thesis.

| Topics | Words with highest probabilities |
|----------|--|
| Topic 1 | immigration; illegal_immigration; border_patrol; border_security; agent; alien; illegal_alien; deport; southern_border; visa; citizenship |
| Topic 2 | tax_relief; revenue; tax_cut; economic_growth; trillion; raising_tax; tax_increase; tax_policy; cut_tax; american_family; fiscal; tax_revenue |
| Topic 3 | china; trade; free_trade; export; wto; cafta; trade_agreement; chinese; manufacture; world_trade; counterfeit; central_america; tariff |
| Topic 4 | drug; traffick; local_law; murder; penalty; prosecute; police; sentence; mandatory_minimum; task_force; gang; deal; sheriff; attorney |
| Topic 5 | cell; embryo; patient; stem_cell; disease; embryonic_stem; doctor; physician; medicine; cure; nih; adult_stem; stage; drug; ethic |
| Topic 6 | agriculture; animal; farmer; usda; horse; label; manufacture; food_safety; meat; rancher; farm; eat; plant; livestock; slaughter |
| Topic 7 | oil; coal; drill; gasoline; ethanol; electric; gallon; car; peak; pump; plant; burn; crude_oil; shelf; gulf; refinery |
| Topic 8 | teacher; head_start; charter_school; catholic_school; workforce; math; teach; academic; classroom; technical_education; enroll; public_school; community_college |
| Topic 9 | army; air_force; veteran; guard; enemy; nation_guard; navy; wound; dod; active_duty; mobile; deploy; marine_corp; hero |
| Topic 10 | port; amtrak; port_security; highway; route; cargo; airport; rail; custom; aircraft; ship; traffic; plane; pilot |

Table 2.1: Example of ten topics discovered by LDA from a collection of floor debates in U.S. Congress.

2.2.1 Using Bayesian Nonparametrics

One major challenge for practitioners when applying LDA on a text corpus is choosing the number of topics, which is required to be fixed in advance. This is usually done by running LDA with different numbers of topics and choosing the one that gives the best performance on some predefined objectives (e.g., likelihood of held-out documents). *Bayesian nonparametrics* provide an elegant solution to this problem. Essentially, Bayesian nonparametric methods provide priors over an infinite-dimensional space of probability distributions and let the observed data decide the actual dimensionality of the posterior distribution.

In this section, we first review *Dirichlet process* (DP)—a nonparametric prior over infinite discrete distributions. We then discuss how LDA can be extended to learn an unbounded number of topics using the *Hierarchical Dirichlet processes* (HDP) (Teh et al., 2006). More details on DP and HDP can be found in Sudderth (2006, Chapter 2) and Teh and Jordan (2010).

Dirichlet process (DP) like the Dirichlet distribution, is a distribution over distributions. A Dirichlet process, denoted by $\text{DP}(\alpha, G_0)$, is parameterized by (1) a *concentration parameter* $\alpha > 0$ and (2) a *base distribution* G_0 over a space Ω . A draw from a Dirichlet process, $G \sim \text{DP}(\alpha, G_0)$ is a distribution over the same space Ω , such that for any finite measurable partition (A_1, A_2, \dots, A_K) of Ω , the following holds

$$(G(A_1), G(A_2), \dots, G(A_K)) \sim \text{Dir}(\alpha G_0(A_1), \alpha G_0(A_2), \dots, \alpha G_0(A_K)) \quad (2.5)$$

This means that if we draw a random distribution from $\text{DP}(\alpha, G_0)$, there will be on average $G_0(A_k)$ probability mass at $A_k \in \Omega$, and the concentration parameter α controls how tightly $G(A_k)$ concentrates around $G_0(A_k)$ defined by the base distribution.

The existence of the Dirichlet process was established by Ferguson (1973). There are two concepts related to the Dirichlet process: the *stick-breaking process* (SBP) and the *Chinese restaurant process* (CRP).

- The stick-breaking process provides an explicit way to construct a Dirichlet

process. [Sethuraman \(1994\)](#) shows that a random distribution G drawn from $\text{DP}(\alpha, G_0)$ can be defined as

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \quad (2.6)$$

where δ_{ϕ_k} is a probability distribution concentrated at ϕ_k which are iid draws from the based distribution G_0

$$\phi_k |_{k=1}^{\infty} \sim G_0 \quad (2.7)$$

and π_k is defined based on iid draws from $\text{Beta}(1, \alpha_0)$

$$\pi_k = \pi'_k \prod_{i=1}^{k-1} (1 - \pi'_i) \quad \pi'_k |_{k=1}^{\infty} \sim \text{Beta}(1, \alpha_0) \quad (2.8)$$

Equation 2.6 shows that G is *discrete* with probability 1. Intuitively, a draw G from the $\text{DP}(\alpha, G_0)$ can be seen as a distribution over infinite discrete “atoms”, each has a weight π_k and a value ϕ_k drawn from the base distribution. The sequence of weights $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ satisfies $\sum_{k=1}^{\infty} \pi_k = 1$, which makes $\boldsymbol{\pi}$ a probability distribution and usually denoted by $\boldsymbol{\pi} \sim \text{GEM}(\alpha_0)$.

- The Chinese restaurant process shows the *clustering property* of random draws from the distribution G drawn from $\text{DP}(\alpha, G_0)$, in which G is marginalized out ([Blackwell and MacQueen, 1973](#)). Let $\theta_1, \theta_2, \dots, \theta_n$ be a sequence of n random draws from G , the next draw θ_{n+1} is distributed according to the

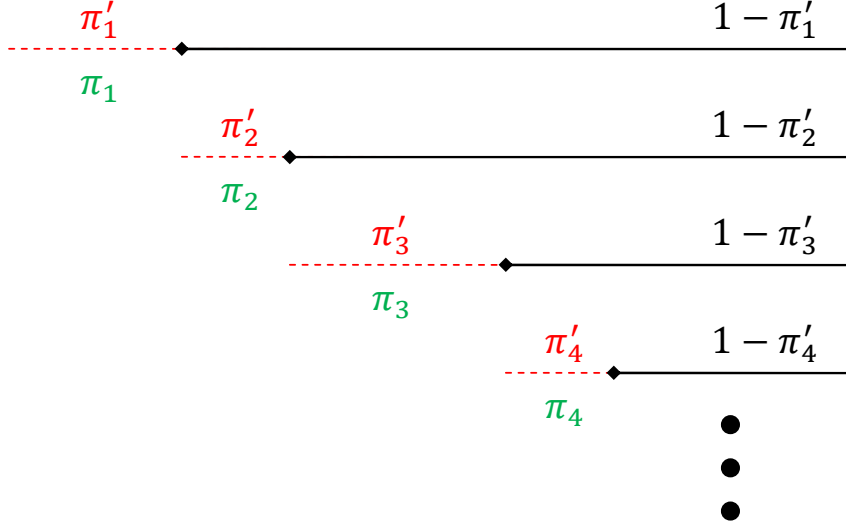


Figure 2.3: Illustration of the stick breaking process $\boldsymbol{\pi} \sim \text{GEM}(\alpha_0)$, in which $\pi_k = \pi'_k \prod_{i=1}^{k-1} (1 - \pi'_i)$ is defined based on the fraction $\pi'_k \sim \text{Beta}(1, \alpha_0)$ that is taken from the remainder of the stick after each break.

following conditional distribution

$$\theta_{n+1} \mid \theta_1, \dots, \theta_n; \alpha, G_0 \sim \sum_{k=1}^K \frac{N_k}{n + \alpha} \delta_{\phi_k} + \frac{\alpha}{n + \alpha} G_0 \quad (2.9)$$

where ϕ_1, \dots, ϕ_K are K distinct atom values that the first n draws $\theta_1, \dots, \theta_n$ take on, and N_k is the number of draws θ_i that take on ϕ_k .

Equation 2.9 illustrates the Chinese restaurant metaphor: There is a Chinese restaurant with infinite number of tables (i.e., atoms). Each customer (i.e., θ_i) comes in and chooses a table to sit. The customer sits at table k (i.e., atom k) with probability proportional to the number of customers already seated there (i.e., N_k) and enjoys the table's existing dish (i.e., ϕ_k). With probability α , the customer sits at a new table and order a new dish (i.e., new draw from the base distribution G).

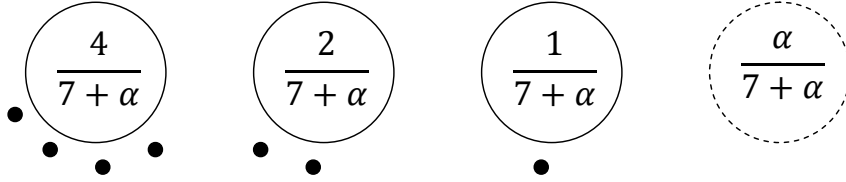


Figure 2.4: Illustration of the Chinese restaurant process metaphor in which there are seven customers currently occupying three tables. A new customer coming in will sit at (1) an existing table with a probability proportional to the number of customers currently sitting at the table or (2) a new table with a probability proportional to α . The exact probabilities are shown inside each table.

Hierarchical Dirichlet processes (HDP) uses the Dirichlet process to extend LDA to capture an infinite number of topics (Teh et al., 2006). In LDA, each document d has a topic proportion θ_d drawn from a finite Dirichlet(α) prior. To handle infinite number of topics, HDP instead draws the topic proportion G_d for each document d from a Dirichlet process $\text{DP}(\alpha, G_0)$. Given the per-document distribution over topics G_d , the generative process for each token is then similar to that of LDA.

All that is left now to fully define the generative process of HDP is specifying the base distribution G_0 . One straightforward way is to use a Dirichlet distribution to define G_0 . However, there is a major disadvantage in doing so: since a Dirichlet distribution is a continuous distribution over topics, words within documents will share the same set of topics (i.e., draws from G_d) but words across different documents will not. To address this problem, the HDP draws G_0 from another Dirichlet process $\text{DP}(\gamma, H)$ with the base distribution H being a Dirichlet distribution Dirichlet(β). Figure 2.5 shows the generative process and plate notation diagram of HDP.

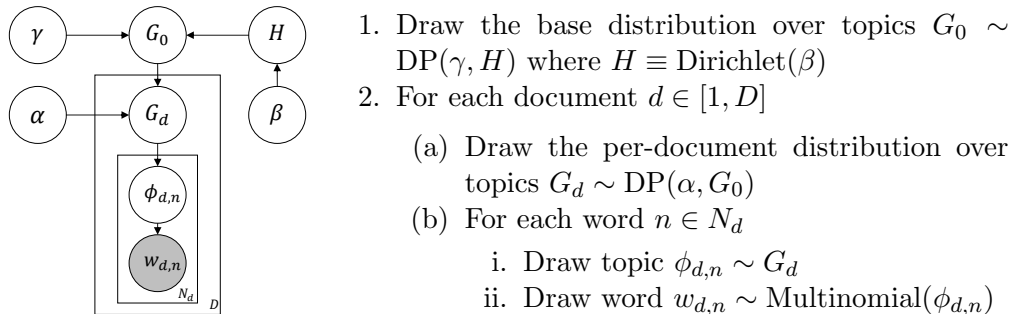


Figure 2.5: Generative process and the plate diagram representation of HDP.

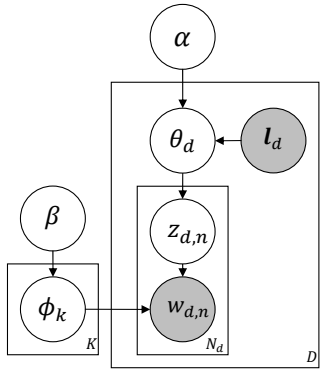
2.2.2 Incorporating Metadata

In many settings, each document in the corpus contains additional information called *metadata* such as author, geographic location, published venue, timestamp etc. The availability of these metadata has motivated various models to jointly capture both the text and the metadata. Accounting for such additional information can not only result in better topics discovered, but also improve the prediction results of documents’ unseen metadata given their text. Following [Mimno and McCallum \(2008\)](#), we consider two broad categories of such models: *upstream models* and *downstream models*.

Upstream models are topic models in which metadata directly or indirectly generate the latent topic variables. We will first review Labeled LDA—a popular upstream topic model for multi-labeled data, which is also the basis of L2H, the model we introduce in Chapter 4.

Proposed by [Ramage et al. \(2009\)](#), Labeled LDA (L-LDA) models multi-labeled documents in which each document is tagged with multiple labels. More specifically, L-LDA takes as input a set of D documents \mathbf{w}_d , where each is tagged

with \mathbf{l}_d labels. Labeled LDA associates each of its L unique labels with a topic (i.e., a multinomial over the vocabulary as in standard LDA). Like LDA, L-LDA models each document d as mixture over topics. However, given the set of observable labels \mathbf{l}_d , L-LDA only allows tokens in d to be generated from topics that are associated with \mathbf{l}_d . Figure 2.6 shows the generative process of L-LDA. With similar goal to model multi-labeled data, Partially Labeled LDA (Ramage et al., 2011, PLDA), Prior-LDA and Dependency-LDA (Rubin et al., 2012) are proposed to capture the dependencies among the labels by projecting them onto a lower-dimensional latent space.



1. For each topic $k \in [1, L]$
 - Draw word distribution $\phi_k \sim \text{Dirichlet}(\beta)$
2. For each document $d \in [1, D]$ with labels \mathbf{l}_d
 - Generate α_d by masking out α using \mathbf{l}_d
 - Draw topic distribution $\theta_d \sim \text{Dirichlet}(\alpha_d)$
 - For each token $n \in [1, N_d]$
 - Draw topic $z_{d,n} \sim \text{Multinomial}(\theta_d)$
 - Draw word $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$

Figure 2.6: Generative process and the plate diagram of Labeled LDA.

Besides multiple labels, various upstream models have been proposed to capture other types of metadata. For example, to include authorship information, Rosen-Zvi et al. (2004) introduce the Author-Topic model and Mimno and McCallum (2007) present the Author-Persona-Topic (APT) model. Lacoste-Julien et al. (2008) introduce Discriminative LDA (DiscLDA) to incorporate single discrete class and use it to modify the document-specific topic proportion by applying a class-dependent linear transformation. Boyd-Graber and Blei (2009) present the Multilin-

gual Topic Model (MuTo) to incorporate language information to analyze unaligned multilingual text. [Mimno and McCallum \(2008\)](#) propose the Dirichlet-multinomial Regression (DMR) model, which can take into account different types of metadata.

Downstream models are topic models in which both the words and the metadata are generated from latent topic variables. In LDA, there are two types of latent topic variables: (1) document-specific topic proportion θ_d and (2) token-specific topic assignment $z_{d,n}$, which provides two general ways to generate metadata in downstream models.

The most straightforward way is arguably to generate both the words and the metadata simultaneously given the latent topic proportions. In this type of models, words and metadata can be considered exchangeable, in which each topic, besides a multinomial over words as in standard LDA, also has additional distributions over metadata values. Examples include models that jointly capture text and metadata such as references ([Erosheva et al., 2004](#)), timestamps ([Wang and McCallum, 2006](#), TOT), named entities (e.g., persons, organizations, locations) ([Newman et al., 2006](#)), and citations ([Nallapati et al., 2008](#)). Polylingual topic model ([Mimno et al., 2009](#), PLTM) also falls under this family of models if we consider aligned text in other languages the metadata.

In the second type of downstream models, the metadata are generated from the empirical topic assignments. Falling under this type of models is Supervised LDA (sLDA)—a flexible downstream topic model that jointly captures text and *continuous responses*. sLDA has become the foundation for various downstream topic

models, including SHLDA—the model we present in Chapter 5. Proposed by [Blei and McAuliffe \(2007\)](#), sLDA is designed to jointly model a set of D documents \mathbf{w}_d , each of which is associated with a continuous response y_d . Examples of this type of data include product reviews associated with their ratings, online status updates associated with their geographical locations, and legislative text accompanied by their author’s ideological score. sLDA captures the relationship between latent topics and metadata by modeling each document’s continuous response variable using a normal linear model, whose covariates are the document’s empirical distribution of topics: $y_d \sim \mathcal{N}(\boldsymbol{\eta}^T \bar{\mathbf{z}}_d, \rho)$ where $\boldsymbol{\eta}$ is the regression parameter vector and $\bar{\mathbf{z}}_d$ is the empirical distribution over topics of document d . The generative process of sLDA is shown in Figure 2.7.

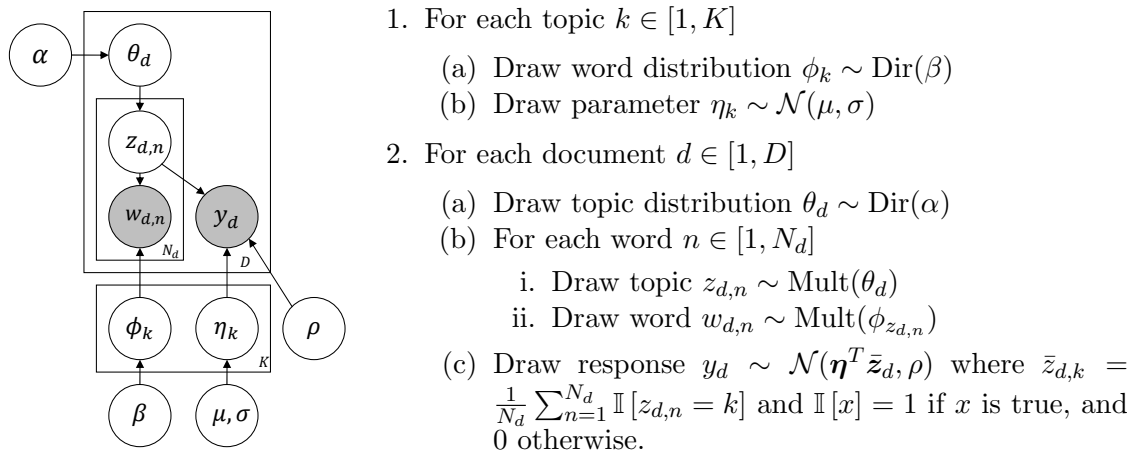


Figure 2.7: Generative process and the plate diagram representation of sLDA.

Since its introduction, sLDA has been extended in many ways. [Wang et al. \(2009\)](#) extend sLDA for multi-class responses. [Chang et al. \(2010\)](#) propose the Relational Topic Model (RTM) to jointly capture text and the links between documents. [Boyd-Graber and Resnik \(2010\)](#) introduce Multilingual Supervised LDA

(or MLSDA) which simultaneously models multilingual text and their accompanied continuous responses. [Zhu et al. \(2012, MedLDA\)](#) improves sLDA by using the discriminative max-margin principle.

2.2.3 Adding Hierarchical Structure

Another active research direction in extending LDA is to use a *hierarchical structure* to capture the correlations among different topics. One early work on this topic is by [Blei et al. \(2003a, 2010b\)](#) who introduce the *nested Chinese restaurant process* (nCRP) as a flexible prior over an infinitely deep, infinitely branching tree-structured hierarchy. A tree can be seen as a nested sequence of partitions. nCRP defines a distribution on trees by putting a probability on each of such sequences. In a nCRP, each node in the tree is a CRP defining a distribution over its infinitely many children. A tree is generated from an nCRP by traversing from the root downward by repeatedly drawing a child node from its parent’s CRP.

Due to its capability to define flexible tree-structured hierarchies, nCRP has been used in many hierarchical topic models to capture the relationships among the latent topics. [Blei et al. \(2003a, 2010b\)](#) use nCRP to define Hierarchical LDA (hLDA) in which each node in the tree is associated with a topic. A document in hLDA is generated by (1) drawing a path in the tree, (2) drawing a node on the chosen path for each document’s token, and (3) using the topic associated with the chosen node to generate the token. Since higher-level nodes are accessible by more documents, this generative process makes the posterior place more general topics

near the root of the tree and more specialized topics further down in the tree.

One major drawback of hLDA, however, is that each document is restricted to only a single path in the tree. Since each path is designed to capture a consistent theme, from more general (i.e., at higher-level nodes) to more specific (i.e., at lower-level nodes), restricting a document to be about a theme is a relatively strong assumption, especially when modeling long documents. Recent work relaxes this restriction by using different priors: tree-structured stick breaking (Adams et al., 2010, TSSB), recursive Chinese restaurant processes (Kim et al., 2012, rCRP), nested Chinese restaurant franchises (Ahmed et al., 2013a,b, nCRF), and nested hierarchical Dirichlet processes (Paisley et al., 2014, nHDP).

Besides trees, other hierarchical structures have also been used to model the topic space. Li and McCallum (2006) use a directed acyclic graph (DAG) to propose the Pachinko allocation model (PAM), which captures arbitrary, nested correlations between topics. Li et al. (2007) use a nonparametric prior for PAM to learn both the number of topics and how the topics are correlated. To add the nested nature of topic hierarchies of hLDA to PAM’s ability to mix topics, Mimno et al. (2007) propose HPAM. Chambers et al. (2010) use a general graph to model topics in GraphLDA.

Combining hierarchical structure with metadata has also attracted much topic modeling research. For example, Slutsky et al. (2013a) introduce Tree Labeled LDA, which extends Labeled LDA for the case where labels are organized in a known tree. Semi-supervised Hierarchical LDA, introduced by (Mao et al., 2012a, SSSLDA), generalizes hLLDA by allowing the document hierarchy labels to be

partially observed, with unobserved labels and topic tree structure then inferred from the data. In addition to these upstream models, [Perotte et al. \(2011\)](#) propose a downstream model called Hierarchically Supervised LDA (HSLDA) which treats documents' hierarchical labels as the response. [Ho et al. \(2012\)](#) propose TopicBlock to learn the topic hierarchy from text with relational links.

2.2.4 Other extensions

So far, we have provided an overview on three directions of extending LDA to use nonparametric priors, incorporate metadata, and capture topic correlation using hierarchical structure. These are a part of a much larger body of topic modeling research which includes other important directions such as

- Relaxation of LDA's assumptions including the bag-of-word assumption ([Griffiths et al., 2004](#); [Wallach, 2006](#); [Boyd-Graber and Blei, 2008a](#)) and the document exchangeability assumption ([Teh et al., 2006](#); [Blei and Lafferty, 2006](#); [Wang et al., 2008](#); [Ren et al., 2008](#); [Fox et al., 2008](#); [Ahmed and Xing, 2008, 2010b](#); [Du et al., 2010](#); [Blei and Frazier, 2011](#))
- Visualization and user interfaces ([Gardner et al., 2010](#); [Gretarsson et al., 2012](#); [Chaney and Blei, 2012](#); [Eisenstein et al., 2012](#); [Chuang et al., 2012](#); [Dou et al., 2013](#))
- Evaluation of topic quality ([Chang et al., 2009b](#); [Newman et al., 2010](#); [Mimno et al., 2011](#); [Stevens et al., 2012](#); [Aletras and Stevenson, 2013a](#); [Lau et al., 2014b](#); [Röder et al., 2015](#))

- Automatic topic labeling (Mei et al., 2007; Magatti et al., 2009; Lau et al., 2010, 2011; Mao et al., 2012b; Aletras and Stevenson, 2013b, 2014)
- Large-scale topic models (Smola and Narayanamurthy, 2010; Zhai et al., 2012; Hoffman et al., 2013; Li et al., 2014)

2.3 MCMC Inference and the Importance of Averaging

In the previous two sections, we have provided an overview of probabilistic topic models and their uses in a wide range of applications. One important aspect of topic models that we have yet to cover is *posterior inference*—estimating the posterior distribution over the latent variables given the observed variables. Exact computation of the posterior is often intractable, which motivates approximate inference techniques (Asuncion et al., 2009). One popular approach is Markov chain Monte Carlo (MCMC), a class of inference algorithms to approximate the target posterior distribution by drawing a set of samples using a Markov chain (Andrieu et al., 2003). In general, given a density $f(x)$ which is hard to compute exactly, MCMC algorithms draw a set of T samples and average over these samples to estimate $f(x)$. The theory behind MCMC—what ensures that we have the correct estimated density—relies on taking the limit as T goes to infinity, which we approximate by only using a large but finite T .

To make predictions, MCMC algorithms generate samples on training data to estimate corpus-level latent variables, and use them to generate samples to estimate document-level latent variables for test data. The underlying theory requires aver-

aging on both training and test samples, but in practice it is often convenient to cut corners: either skip averaging entirely by using just the values of the last sample or use a single training sample and average over test samples.

In this section, we systematically study non-averaging and averaging strategies when performing predictions using MCMC in topic modeling. We review some key concepts about MCMC in general in Section 2.3.1 and specifically for topic models in Section 2.3.2. In Section 2.3.3, we describe different strategies to obtain the final prediction values in topic model using MCMC. Using popular unsupervised (LDA in Section 2.3.4) and supervised (sLDA in Section 2.3.5) topic models via thorough experimentation, we show empirically that cutting corners on averaging leads to consistently poorer prediction.

2.3.1 Learning and Predicting with MCMC

While reviewing all of MCMC is beyond the scope of this section, we need to briefly review key concepts.¹ To estimate a target density $p(x)$ in a high-dimensional space \mathcal{X} , MCMC generates samples $\{x_t\}_{t=1}^T$ while exploring \mathcal{X} using the Markov assumption. Under this assumption, sample x_{t+1} depends on sample x_t only, forming a *Markov chain*, which allows the sampler to spend more time in the most important regions of the density. Two concepts control sample collection:

Burn-in B : Depending on the initial value of the Markov chain, MCMC algorithms might take time to reach the stationary state where samples are drawn from the true

¹For more details please refer to (Neal, 2003; Andrieu et al., 2003; Resnik and Hardisty, 2010).

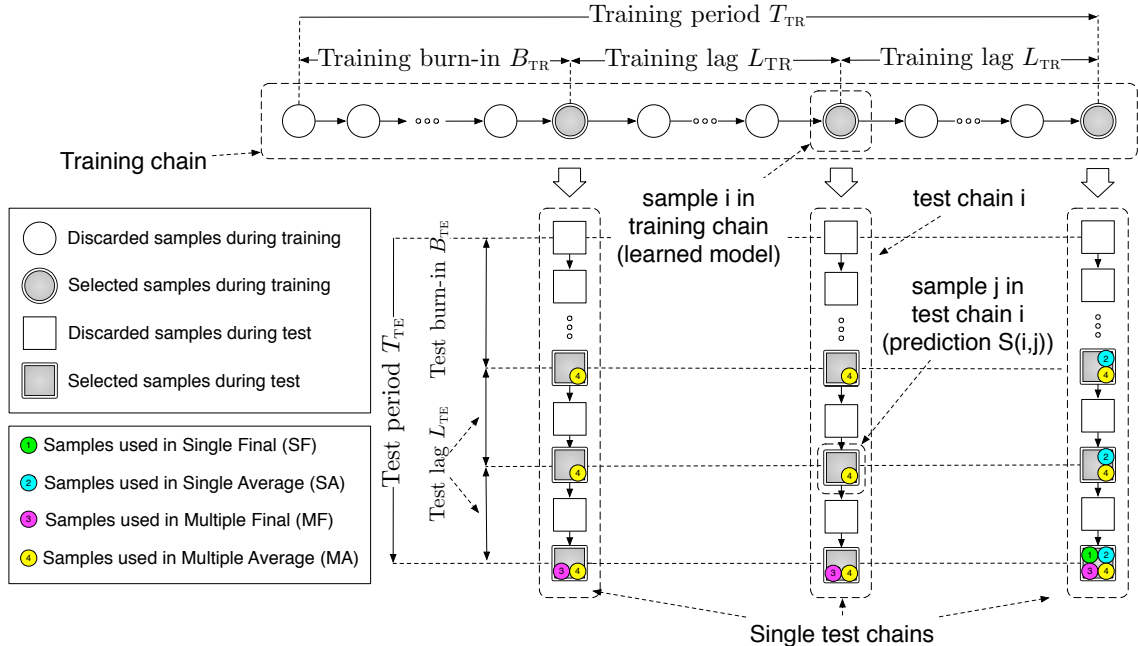


Figure 2.8: Illustration of training and test chains in MCMC, showing samples used in four prediction strategies studied in this section: Single Final (SF), Single Average (SA), Multiple Final (MF), and Multiple Average (MA).

target distribution. Thus, in practice, samples before a burn-in period B are often discarded.

Sample-lag L : Averaging over samples to estimate the target distribution requires i.i.d. samples. However, future samples depend on the current samples (i.e., the Markov assumption). To avoid autocorrelation, we discard all but every L samples.

2.3.2 MCMC in Topic Modeling

As generative probabilistic models, topic models define a joint distribution over latent variables and observable evidence. In our setting, the latent variables consist of corpus-level *global* variables \mathbf{g} and document-level *local* variables \mathbf{l} ; while the evidence consists of words \mathbf{w} and additional metadata \mathbf{y} —the latter omitted in

unsupervised models.

During training, MCMC estimates the posterior $p(\mathbf{g}, \mathbf{l}^{\text{TR}} \mid \mathbf{w}^{\text{TR}}, \mathbf{y}^{\text{TR}})$ by generating a *training Markov chain* of T_{TR} samples.^{2,3} Each training sample i provides a set of fully realized global latent variables $\hat{\mathbf{g}}(i)$, which can generate test data. During test time, given a learned model from training sample i , we generate a *test Markov chain* of T_{TE} samples to estimate the local latent variables $p(\mathbf{l}^{\text{TE}} \mid \mathbf{w}^{\text{TE}}, \hat{\mathbf{g}}(i))$ of test data. Each sample j of test chain i provides a fully estimated local latent variables $\hat{\mathbf{l}}^{\text{TE}}(i, j)$ to make a prediction.

Figure 2.8 shows an overview. To reduce the effects of unconverged and autocorrelated samples as discussed in Section 2.3.1, during training we use a burn-in period of B_{TR} and a sample-lag of L_{TR} iterations. We use $\mathcal{T}_{\text{TR}} = \{i \mid i \in (B_{\text{TR}}, T_{\text{TR}}] \wedge (i - B_{\text{TR}}) \bmod L_{\text{TR}} = 0\}$ to denote the set of indices of the selected models. Similarly, B_{TE} and L_{TE} are the test burn-in and sample-lag. The set of indices of selected samples in test chains is $\mathcal{T}_{\text{TE}} = \{j \mid j \in (B_{\text{TE}}, T_{\text{TE}}] \wedge (j - B_{\text{TE}}) \bmod L_{\text{TE}} = 0\}$.

2.3.3 Averaging Strategies

We use $S(i, j)$ to denote the prediction obtained from sample j of the test chain i . We now discuss different strategies to obtain the final prediction:

- **Single Final (sf)** uses the last sample of last test chain to obtain the predicted value,

$$S_{\text{SF}} = S(T_{\text{TR}}, T_{\text{TE}}). \tag{2.10}$$

²We omit hyperparameters in conditional probabilities for clarity.
³We split data into training (TR) and testing (TE) folds, and denote the training iteration i and the testing iteration j within the corresponding Markov chains.

- **Single Average (sa)** averages over multiple samples in the last test chain

$$S_{\text{SA}} = \frac{1}{|\mathcal{T}_{\text{TE}}|} \sum_{j \in \mathcal{T}_{\text{TE}}} S(T_{\text{TR}}, j). \quad (2.11)$$

This is a common averaging strategy in which we obtain a point estimate of the global latent variables at the end of the training chain. Then, a single test chain is generated on the test data and multiple samples of this test chain are averaged to obtain the final prediction (Chang, 2012; Singh et al., 2012; Jiang et al., 2012; Zhu et al., 2014a).

- **Multiple Final (mf)** averages over the last samples of multiple test chains from multiple models

$$S_{\text{MF}} = \frac{1}{|\mathcal{T}_{\text{TR}}|} \sum_{i \in \mathcal{T}_{\text{TR}}} S(i, T_{\text{TE}}). \quad (2.12)$$

- **Multiple Average (ma)** averages over all samples of multiple test chains for distinct models,

$$S_{\text{MA}} = \frac{1}{|\mathcal{T}_{\text{TR}}|} \frac{1}{|\mathcal{T}_{\text{TE}}|} \sum_{i \in \mathcal{T}_{\text{TR}}} \sum_{j \in \mathcal{T}_{\text{TE}}} S(i, j), \quad (2.13)$$

2.3.4 Unsupervised Topic Models

We evaluate the predictive performance of the unsupervised topic model LDA using different averaging strategies in Section 2.3.3. In LDA, the global latent variables are topics $\{\phi_k\}_{k=1}^K$ and the local latent variables for each document d are topic proportions θ_d .

Train: During training, we use collapsed Gibbs sampling to assign each token in the training data with a topic (Steinberger and Griffiths, 2006). The probability of

assigning token n of training document d to topic k is

$$p(z_{d,n}^{\text{TR}} = k \mid \mathbf{z}_{-d,n}^{\text{TR}}, \mathbf{w}_{-d,n}^{\text{TR}}, w_{d,n}^{\text{TR}} = v) \propto \frac{N_{\text{TR},d,k}^{-d,n} + \alpha}{N_{\text{TR},d,\cdot}^{-d,n} + K\alpha} \cdot \frac{N_{\text{TR},k,v}^{-d,n} + \beta}{N_{\text{TR},k,\cdot}^{-d,n} + V\beta}, \quad (2.14)$$

where $N_{\text{TR},d,k}$ is the number of tokens in the training document d assigned to topic k , and $N_{\text{TR},k,v}$ is the number of times word type v assigned to topic k . Marginal counts are denoted by \cdot , and $^{-d,n}$ denotes the count excluding the assignment of token n in document d .

At each training iteration i , we estimate the distribution over words $\hat{\phi}_k(i)$ of topic k as

$$\hat{\phi}_{k,v}(i) = \frac{N_{\text{TR},k,v}(i) + \beta}{N_{\text{TR},k,\cdot}(i) + V\beta} \quad (2.15)$$

where the counts $N_{\text{TR},k,v}(i)$ and $N_{\text{TR},k,\cdot}(i)$ are taken at training iteration i .

Test: Because we lack explicit topic annotations, we use *perplexity*—a widely-used metric to measure the predictive power of topic models on held-out documents. To compute perplexity, we follow the *estimating θ* method (Wallach et al., 2009, Section 5.1) and evenly split each test document d into $\mathbf{w}_d^{\text{TE1}}$ and $\mathbf{w}_d^{\text{TE2}}$. We first run Gibbs sampling on $\mathbf{w}_d^{\text{TE1}}$ to estimate the topic proportion $\hat{\theta}_d^{\text{TE}}$ of test document d . The probability of assigning topic k to token n in $\mathbf{w}_d^{\text{TE1}}$ is

$$p(z_{d,n}^{\text{TE1}} = k \mid \mathbf{z}_{-d,n}^{\text{TE1}}, \mathbf{w}^{\text{TE1}}, \hat{\phi}(i)) \propto \frac{N_{\text{TE1},d,k}^{-d,n} + \alpha}{N_{\text{TE1},d,\cdot}^{-d,n} + K\alpha} \cdot \hat{\phi}_{k,w_{d,n}^{\text{TE1}}}(i) \quad (2.16)$$

where $N_{\text{TE}_1,d,k}$ is the number of tokens in $\mathbf{w}_d^{\text{TE}_1}$ assigned to topic k . At each iteration j in test chain i , we can estimate the topic proportion vector $\hat{\theta}_d^{\text{TE}}(i, j)$ for test document d as

$$\hat{\theta}_{d,k}^{\text{TE}}(i, j) = \frac{N_{\text{TE}_1,d,k}(i, j) + \alpha}{N_{\text{TE}_1,d,\cdot}(i, j) + K\alpha} \quad (2.17)$$

where both the counts $N_{\text{TE}_1,d,k}(i, j)$ and $N_{\text{TE}_1,d,\cdot}(i, j)$ are taken using sample j of test chain i .

Prediction: Given $\hat{\theta}_d^{\text{TE}}(i, j)$ and $\hat{\phi}(i)$ at sample j of test chain i , we compute the predicted likelihood for each unseen token $w_{d,n}^{\text{TE}_2}$ as

$$S(i, j) \equiv p(w_{d,n}^{\text{TE}_2} | \hat{\theta}_d^{\text{TE}}(i, j), \hat{\phi}(i)) = \sum_{k=1}^K \hat{\theta}_{d,k}^{\text{TE}}(i, j) \cdot \hat{\phi}_{k,w_{d,n}^{\text{TE}_2}}(i) \quad (2.18)$$

Using different strategies described in Section 2.3.3, we obtain the final predicted likelihood for each unseen token $p(w_{d,n}^{\text{TE}_2} | \hat{\theta}_d^{\text{TE}}, \hat{\phi})$ and compute the perplexity as

$$\exp\left(-\frac{\sum_d \sum_n \log(p(w_{d,n}^{\text{TE}_2} | \hat{\theta}_d^{\text{TE}}, \hat{\phi}))}{N^{\text{TE}_2}}\right) \quad (2.19)$$

where N^{TE_2} is the number of tokens in \mathbf{w}^{TE_2} .

Setup: We use three Internet review datasets in our experiment. For all datasets, we preprocess by tokenizing, removing stopwords, stemming, adding bigrams to the vocabulary, and we filter using TF-IDF to obtain a vocabulary of 10,000 words.⁴

⁴To find bigrams, we begin with bigram candidates that occur at least 10 times in the corpus and use a χ^2 test to filter out those having a χ^2 value less than 5. We then treat selected bigrams as single word types and add them to the vocabulary.

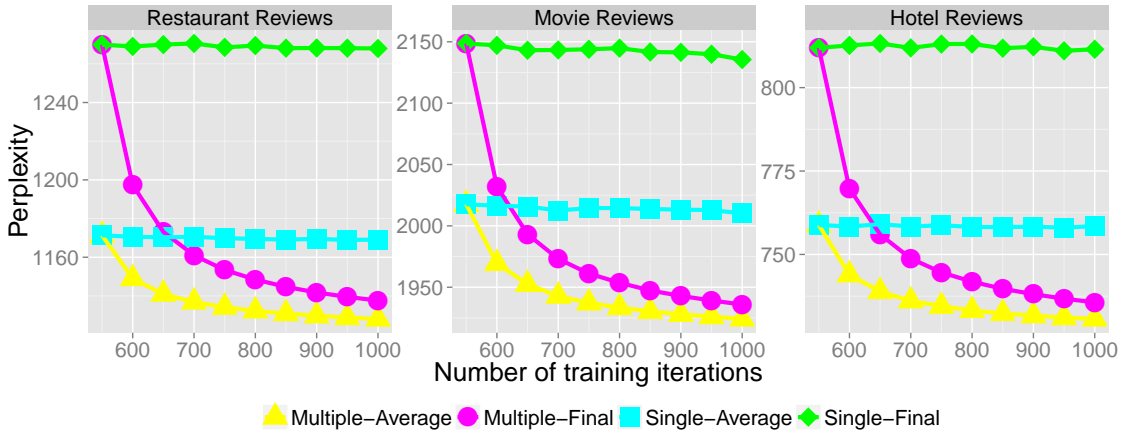


Figure 2.9: Perplexity of LDA using different averaging strategies with different number of training iterations T_{TR} . Perplexity generally decreases with additional training iterations, but the drop is more pronounced with multiple test chains.

The three datasets are:

- HOTEL: 240,060 reviews of hotels from TripAdvisor (Wang et al., 2010).
- RESTAURANT: 25,459 reviews of restaurants from Yelp (Jo and Oh, 2011).
- MOVIE: 5,006 reviews of movies from Rotten Tomatoes (Pang and Lee, 2005)

We report cross-validated average performance over five folds, and use $K = 50$ topics for all datasets. To update the hyperparameters, we use slice sampling (Walach, 2008, p. 62).⁵

Results: Figure 2.9 shows the perplexity of the four averaging methods, computed with different number of training iterations T_{TR} . SA outperforms SF, showing the benefits of averaging over multiple test samples from a single test chain. However, both multiple chain methods (MF and MA) significantly outperform these two methods.

⁵MCMC setup: $T_{\text{TR}} = 1,000$, $B_{\text{TR}} = 500$, $L_{\text{TR}} = 50$, $T_{\text{TE}} = 100$, $B_{\text{TE}} = 50$ and $L_{\text{TE}} = 5$.

This result is consistent with [Asuncion et al. \(2009\)](#), who run multiple training chains but a single test chain for each training chain and average over them. This is more costly since training chains are usually significantly longer than test chains. In addition, multiple training chains are sensitive to their initialization.

2.3.5 Supervised Topic Models

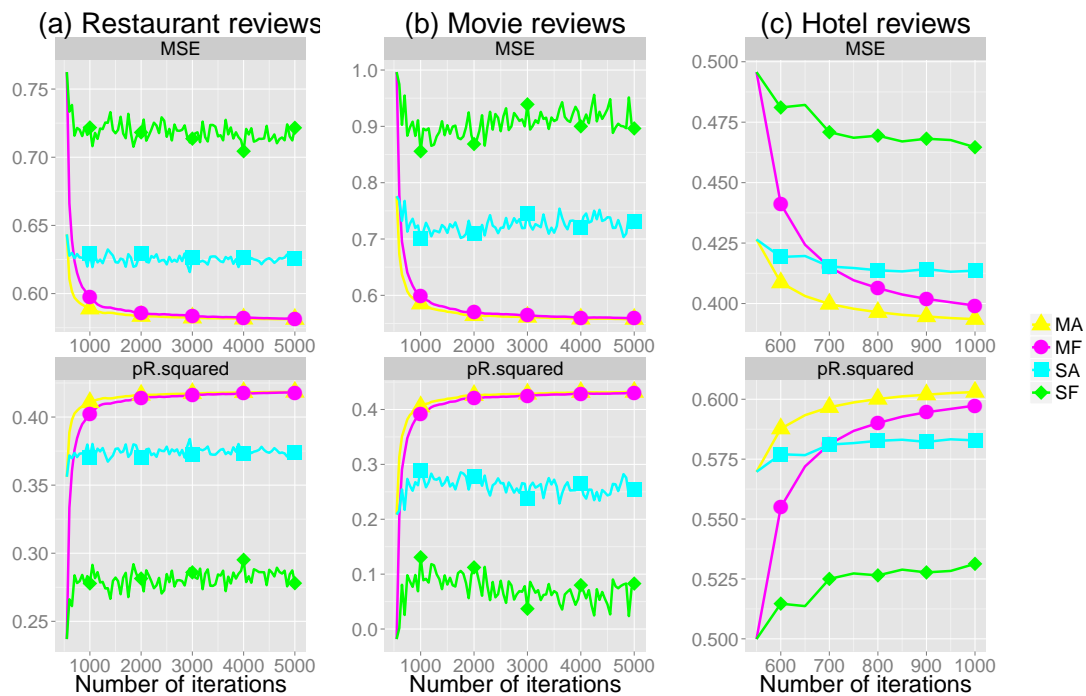


Figure 2.10: Performance of sLDA using different averaging strategies computed at each training iteration.

We evaluate the performance of different prediction methods using supervised latent Dirichlet allocation (sLDA) ([Blei and McAuliffe, 2007](#)) for sentiment analysis: predicting review ratings given review text. Each review text is the document \mathbf{w}_d and the metadata \mathbf{y}_d is the associated rating. In sLDA, in addition to the K multinomials $\{\phi_k\}_{k=1}^K$, the global latent variables also contain the regression parameter

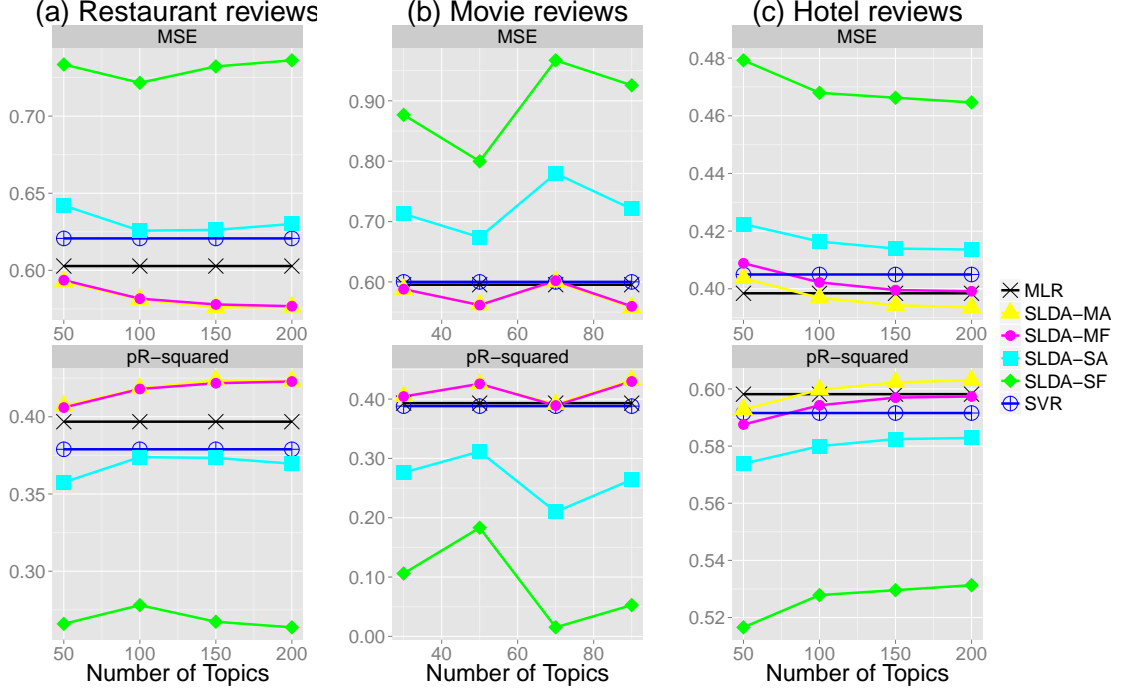


Figure 2.11: Performance of sLDA using different averaging strategies computed at the final training iteration T_{TR} , compared with two baselines MLR and SVR. Methods using multiple test chains (MF and MA) perform as well as or better than the two baselines, whereas methods using a single test chain (SF and SA) perform significantly worse.

η_k for each topic k . The local latent variables of sLDA resembles LDA's: the topic proportion vector θ_d for each document d .

Train: For posterior inference during training, following (Boyd-Graber and Resnik, 2010), we use stochastic EM, which alternates between (1) a Gibbs sampling step to assign a topic to each token, and (2) optimizing the regression parameters. The probability of assigning topic k to token n in the training document d is

$$p(z_{d,n}^{\text{TR}} = k \mid \mathbf{z}_{-d,n}^{\text{TR}}, \mathbf{w}_{-d,n}^{\text{TR}}, w_{d,n}^{\text{TR}} = v) \propto \mathcal{N}(y_d; \mu_{d,n}, \rho) \cdot \frac{N_{\text{TR},d,k}^{-d,n} + \alpha}{N_{\text{TR},d,\cdot}^{-d,n} + K\alpha} \cdot \frac{N_{\text{TR},k,v}^{-d,n} + \beta}{N_{\text{TR},k,\cdot}^{-d,n} + V\beta} \quad (2.20)$$

where $\mu_{d,n} = (\sum_{k'=1}^K \eta_{k'} N_{\text{TR},d,k'}^{-d,n} + \eta_k) / N_{\text{TR},d}$ is the mean of the Gaussian generating y_d

if $z_{d,n}^{\text{TR}} = k$. Here, $N_{\text{TR},d,k}$ is the number of times topic k is assigned to tokens in the training document d ; $N_{\text{TR},k,v}$ is the number of times word type v is assigned to topic k ; \cdot represents marginal counts and $^{-d,n}$ indicates counts excluding the assignment of token n in document d .

We optimize the regression parameters $\boldsymbol{\eta}$ using L-BFGS (Liu and Nocedal, 1989) via the likelihood

$$\mathcal{L}(\boldsymbol{\eta}) = -\frac{1}{2\rho} \sum_{d=1}^D (y_d^{\text{TR}} - \boldsymbol{\eta}^T \bar{\mathbf{z}}_d^{\text{TR}})^2 - \frac{1}{2\sigma} \sum_{k=1}^K (\eta_k - \mu)^2 \quad (2.21)$$

At each iteration i in the training chain, the estimated global latent variables include the a multinomial $\hat{\phi}_k(i)$ and a regression parameter $\hat{\eta}_k(i)$ for each topic k .

Test: Like LDA, at test time we sample the topic assignments for all tokens in the test data

$$p(z_{d,n}^{\text{TE}} = k | \mathbf{z}_{-d,n}^{\text{TE}}, \mathbf{w}^{\text{TE}}) \propto \frac{N_{\text{TE},d,k}^{-d,n} + \alpha}{N_{\text{TE},d,\cdot}^{-d,n} + K\alpha} \cdot \hat{\phi}_{k,w_{d,n}^{\text{TE}}} \quad (2.22)$$

Prediction: The predicted value $S(i, j)$ in this case is the estimated value of the metadata review rating

$$S(i, j) \equiv \hat{y}_d^{\text{TE}}(i, j) = \hat{\boldsymbol{\eta}}(i)^T \bar{\mathbf{z}}_d^{\text{TE}}(i, j), \quad (2.23)$$

where the empirical topic distribution of test document d is defined as $\bar{z}_{d,k}^{\text{TE}}(i, j) \equiv \frac{1}{N_{\text{TE},d}} \sum_{n=1}^{N_{\text{TE},d}} \mathbb{I}[z_{d,n}^{\text{TE}}(i, j) = k]$.

Experimental setup: We use the same data as in Section 2.3.4. For all datasets, the metadata are the review rating, ranging from 1 to 5 stars, which is standardized

using z -normalization. We use two evaluation metrics: mean squared error (MSE) and predictive R-squared (Blei and McAuliffe, 2007).

For comparison, we consider two baselines: (1) multiple linear regression (MLR), which models the metadata as a linear function of the features, and (2) support vector regression (Joachims, 1999, SVR). Both baselines use the normalized frequencies of unigrams and bigrams as features. As in the unsupervised case, we report average performance over five cross-validated folds. For all models, we use a development set to tune their parameter(s) and use the set of parameters that gives best results on the development data at test.⁶

Results: Figure 2.10 shows sLDA prediction results with different averaging strategies, computed at different training iterations.⁷ Consistent with the unsupervised results in Section 2.3.4, SA outperforms SF, but both are outperformed significantly by the two methods using multiple test chains (MF and MA).

We also compare the performance of the four prediction methods obtained at the final iteration T_{TR} of the training chain with the two baselines. The results in Figure 2.11 show that the two baselines (MLR and SVR) outperform significantly the sLDA using only a single test chains (SF and SA). Methods using multiple test chains (MF and MA), on the other hand, match the baseline⁸ (HOTEL) or do better (RESTAURANT and MOVIE).

⁶For MLR we use a Gaussian prior $\mathcal{N}(0, 1/\lambda)$ with $\lambda = a \cdot 10^b$ where $a \in [1, 9]$ and $b \in [1, 4]$; for SVR, we use SVM^{light} (Joachims, 1999) and vary $C \in [1, 50]$, which trades off between training error and margin; for sLDA, we fix $\sigma = 10$ and vary $\rho \in \{0.1, 0.5, 1.0, 1.5, 2.0\}$, which trades off between the likelihood of words and response variable.

⁷MCMC setup: $T_{\text{TR}} = 5,000$ for RESTAURANT and MOVIE and 1,000 for HOTEL; for all datasets $B_{\text{TR}} = 500$, $L_{\text{TR}} = 50$, $T_{\text{TE}} = 100$, $B_{\text{TE}} = 20$ and $L_{\text{TE}} = 5$.

⁸This gap is because sLDA has not converged after 1,000 training iterations (Figure 2.10).

2.3.6 Discussion and Conclusion

MCMC relies on averaging multiple samples to approximate target densities. When used for prediction, MCMC needs to generate and average over both training samples to learn from training data and test samples to make prediction. We have shown that simple averaging—not more aggressive, *ad hoc* approximations like taking the final sample (either training or test)—is not just a question of theoretical aesthetics, but an important factor in obtaining good prediction performance.

Compared with SVR and MLR baselines, sLDA using multiple test chains (MF and MA) performs as well as or better, while sLDA using a single test chain (SF and SA) falters. This simple experimental setup choice can determine whether a model improves over reasonable baselines. In addition, better prediction with shorter training is possible with multiple test chains. Thus, we conclude that averaging using multiple chains produces above-average results.

Chapter 3: Agenda Control in Political Debates

3.1 Introduction

In this chapter, we are interested in discovering *agendas* and *agenda control behaviors* of individuals in political debates and other multi-party conversations. We introduce *Speaker Identity for Topic Segmentation* (SITS), a Bayesian nonparametric topic model which jointly captures topics, topic shifts and individuals' tendency to control the topic of the conversation. The model is capable of discovering (1) the topics used in a set of conversations, (2) how these topics are shared across conversations, (3) when these topics change during conversations, and (4) a speaker-specific measure of agenda control. Using SITS, we analyze the agenda control behaviors of candidates in the 2008 U.S. election debates and the 2012 Republican primary debates, as well as those of participants in a large-scale set of political debate transcripts from CNN's TV show *Crossfire*. To make manual content analysis of conversational transcripts more effective, we build *Argviz*—an interactive visualization which leverages SITS's outputs to help users (e.g., domain experts) quickly grasp the topical dynamics of the conversation, discover when the topic changes and by whom, and interactively visualize the conversation's details on demand. In addition to providing insights on agendas and agenda control in multi-party conversation,

through extensive empirical experiments, we also show that SITS can effectively improve the performance of two quantitative tasks: topic segmentation and influencer detection.

This chapter synthesizes and revises the work originally published in (Nguyen et al., 2012, 2013d, 2014b).

3.1.1 Presidential Debates: Unique Setting for Agenda Control

Presidential debates play a central role in U.S. politics. For example, debuting in 1960 between John Kennedy and Richard Nixon, and having been conducted in every presidential campaign since 1976, televised presidential debates have become a *de facto* election process, in which leading candidates are presented side by side to respond to questions on various important, yet often controversial, issues (Schroeder, 2008). With their unique setting, debates provide candidates distinct opportunities to inform and educate a large and diverse set of audiences about their policy positions, and thus potentially influence votes and elections (Geer, 1988; Holbrook, 1999; Benoit et al., 2002; Blais and Perrella, 2008).¹ Although how much these debates really affect the election outcomes is the subject for a whole different debate (McKinney and Warner, 2013), much previous research agree with Racine Group (2002)’s conclusion that “while journalists and scholars display varying degrees of cynicism about the debates, few deny that viewers find them useful and almost no one doubts that they play an important role in national campaigns” (Benoit et al., 2003; McK-

¹Nielsen (2012) estimated that on average, the 2008 and 2012 presidential debates attracted approximately 57 and 64 million viewers respectively.

inney and Carlin, 2004).

A key question that has attracted much recent research in political science on presidential debates is: *How do candidates control the agenda of the debates?* Previous research suggests that, in general, candidates should and do focus on topics that are most advantageous to them and avoid topics that favor their opponent (Vavreck, 2009; Boydston et al., 2013b). However, *what* are these topics? *When* do the candidates change the topic from one to another? *How often* do they change the topic, especially in the unique setting of debates where candidates do not have complete control over the agenda and are expected to respond to questions from the moderator and the audience? Recent work tackle these questions by performing manual content analysis on debates’ transcripts. Boydston et al. (2013a) manually code the transcripts from each of the three 2008 presidential debates between John McCain and Barack Obama using the Policy Agendas Topics Codebook, to offer empirical support for different types of agenda control behaviors in debates. Boydston et al. (2013b) use similar approach to explore agenda-setting strategies in all presidential debates in 1992, 2004, and 2008. Motivated by this line of work, in this chapter, we introduce SITS, an automated content analysis method using nonparametric Bayesian approach, to study agendas and agenda control in debates.

3.1.2 Agenda Control to Influence in Multi-party Conversations

Although motivated by analyzing political debates, our proposed method SITS is applicable for studying agendas and agenda control in a more general setting:

multi-party conversations. This is a broad category which includes political debates, business meetings, online chats, discussions, conference panels, and many TV or radio talk shows. We also apply SITS to study how participants use agenda control to influence the conversation—an important research problem in communication, sociology and psychology.

Conversation, interactive discussion between two or more people, is one of the most essential and common forms of communication in our daily lives. One of the many functions of conversations is *influence*: having an effect on the belief, opinions or intentions of other conversational participants. Using multi-party conversations to study and identify *influencers*, the people who influence others, has been the focus of researchers in communication, sociology, and psychology (Katz and Lazarsfeld, 1955; Brooke and Ng, 1986; Weimann, 1994), who have long acknowledged that there is a correlation between the conversational behaviors of a participant and how influential he or she is perceived to be by others (Reid and Ng, 2000).

In an early study on this topic, Bales (1970) argues that “to take up time speaking in a small group is to exercise power over the other members for at least the duration of the time taken, regardless of the content.” This statement asserts that *structural patterns* such as speaking time and activeness of participation are good indicators of power and influence in a conversation. Participants who talk most during a conversation are often perceived as having more influence (Sorrentino and Boutillier, 1975; Regula and Julian, 1973; Daley et al., 1977; Ng et al., 1993), more leadership ability (Stang, 1973; Sorrentino and Boutillier, 1975), more dominance (Palmer, 1989; Mast, 2002) and more control of the conversation (Palmer,

1989). Recent work using computational methods also confirms that structural features such as number of turns and turn length are among the most discriminative features to classify whether a participant is influential or not (Rienks et al., 2006; Biran et al., 2012).

However, it is wrong to take Bales's claim too far; the person who speaks loudest and longest is not always the most powerful. In addition to structural patterns, the characteristics of language used also play an important role in establishing influence and controlling the conversation (Ng and Bradac, 1993). For example, particular linguistic choices such as message clarity, powerful and powerless language (Burrell and Koper, 1998), and language intensity (Hamilton and Hunter, 1998) in a message can increase influence. More recently, Huffaker (2010) showed that linguistic diversity expressed by lexical complexity and vocabulary richness has a strong relationship with leadership in online communities. To build a classifier to detect influencers in written online conversations, Biran et al. (2012) also propose to use a set of content-based features to capture various participants' conversational behaviors, including persuasion and agreement/disagreement.

Among many studied behaviors, *agenda control and management* is considered one of the most effective ways to control the conversation (Planalp and Tracy, 1980). Palmer (1989) shows that the less related a participants' utterances are to the immediate topic, the more dominant they are, and then argues, "the ability to change topical focus, especially given strong cultural and social pressure to be relevant, means having enough interpersonal power to take charge of the agenda." Recent work by Rienks et al. (2006) also shows that topic change, among other structural

patterns discussed above, is the most robust feature in detecting influencers in small group meetings.

3.1.3 Topic Segmentation to Capture Conversational Structures

Whether in an informal situation or in more formal settings such as a political debate or business meeting, a conversation is often not about just one thing: topics evolve and are replaced as the conversation unfolds. Discovering this *hidden structure* is a key problem to understand conversations to build conversational assistants (Tur et al., 2010) and develop tools that summarize (Murray et al., 2005) and display (Ehlen et al., 2007) conversational data. Understanding when and how the topics change also helps us study human conversational behaviors such as individuals' agendas (Boydston et al., 2013a), patterns of agreement and disagreement (Hawes et al., 2009; Abbott et al., 2011), relationships among conversational participants (Ireland et al., 2011), and dominance and influence among participants (Palmer, 1989; Rienks et al., 2006).

One of the most natural ways to capture conversational structure is *topic segmentation*—the task of “automatically dividing single long recordings or transcripts into shorter, topically coherent segments” (Purver, 2011; Joty et al., 2013). There are broadly two basic approaches previous works have used to tackle this problem. The first approach focuses on identifying *discourse markers* which distinguish topical boundaries in the conversations. There are certain cue phrases such as *well*, *now*, *that reminds me*, etc. that explicitly indicate the end of one topic or the be-

ginning of another (Hirschberg and Litman, 1993; Passonneau and Litman, 1997). These markers can also serve as features for a discriminative classifier (Galley et al., 2003) or observed variables in generative model (Dowman et al., 2008). However, in practice the discourse markers that are most indicative of topic change often depend heavily on the domain of the data (Purver, 2011). This drawback makes methods solely relying on these markers difficult to adapt to new domains or settings.

The second general approach, which our model is based on, relies on the insight that topical segments evince *lexical cohesion* (Halliday and Hasan, 1976). Intuitively, words within a segment will look more like their neighbors than like words in other segments. This has been a key idea in previous work. Morris and Hirst (1991) determine the structure of text by finding “lexical chains” which consist of units of text that are about the same thing. The often-used text segmentation algorithm TextTiling (Hearst, 1997) exploits this insight to compute the lexical similarity between adjacent sentences. More recent improvements to this approach include using different lexical similarity metrics like LSA (Choi et al., 2001; Olney and Cai, 2005) and improving feature extraction for supervised methods (Hsueh et al., 2006). It also inspires unsupervised models using bags of words (Purver et al., 2006), language models (Eisenstein and Barzilay, 2008), and shared structure across documents (Chen et al., 2009).

3.1.4 Chapter Structure

We describe the model SITS in detail, together with an MCMC algorithm for perform posterior inference in Section 3.2. Applying SITS on real-world conversational data (Section 3.3), we show that this modeling approach is not only more effective than previous methods on traditional topic segmentation (Section 3.6), but also more intuitive in that it is able to capture an important behavior of individual speakers during conversations. More specifically, we analyze qualitatively the agenda control behaviors of candidates in 2008 U.S. election debates and 2012 Republican primary debates, as well as those of political pundits participating in CNN’s *Crossfire* TV show (Section 3.4). We then show that using SITS to model agenda control improves influencer detection (Section 3.5). In Section 3.7, we describe an interactive visualization that can be used to leverage SITS’s output to effectively analyze the dynamics of topics of a conversation. We conclude by summarizing the work and discuss some directions for future work in Section 3.8.

3.2 SITS: Speaker Identity for Topic Segmentation

In this section, we describe SITS, a nonparametric Bayesian model for topic segmentation that takes into consideration *speaker identities*, allowing us to characterize speakers’ agenda control behavior over the course of the conversation. We begin by providing an overview of our approach and highlighting the differences between SITS and previous approaches. We then describe the generative process and inference technique that we use to estimate the model.

3.2.1 Overview of our Approach

We follow the lexical cohesion approach described in Section 3.1.3 by using a probabilistic topic modeling method which we review in Chapter 2. The approach we take is unsupervised, so it requires few resources and is applicable in many domains without extensive training. Following the literature on topic modeling, we define each topic as a multinomial distribution over the vocabulary. Like previous generative models proposed for topic segmentation (Purver et al., 2006), each turn is considered a bag of words generated from an admixture of topics and topics are shared across different turns within a conversation or across different conversations.² In addition, we take a Bayesian nonparametric approach (Müller and Quintana, 2004) to allow the number of topics to be unbounded, in order to better represent the observed data.

In general, SITS takes as input a set of C multi-party conversations. A conversation c has T_c turns, each of which is a maximal uninterrupted utterance by one speaker.³ In each turn $t \in [1, T_c]$, a speaker $a_{c,t}$ utters $N_{c,t}$ words $\mathbf{w}_{c,t} = \{w_{c,t,n} \mid n \in [1, N_{c,t}]\}$. Each word is from a vocabulary of size V , and there are M distinct speakers. This setting is still consistent with those in popular topic models such as LDA (Blei et al., 2003b) or HDP (Teh et al., 2006), in which turns in a conversation are considered independent. In practice, however, this is not the case.

²The “bag of words” treatment of linguistic utterances is widely used, but of course a gross simplification. Previous research has investigated nonparametric models capturing arbitrary-length phrases (Hardisty et al., 2010) and syntactic topic models (Boyd-Graber and Blei, 2008b); integrating linguistically richer models with SITS is a topic for future work.

³Note the distinction from phonetic utterances, which by definition are bounded by silence.

Obviously, the topics of a turn at time t are highly correlated with those of the turn at $t + 1$. To address this issue, recent works have been proposed to capture the temporal dynamics within a document. For example, [Du et al. \(2010\)](#) introduce Sequential LDA to study how topics within a document evolve over its structure. It uses the nested two-parameter Poisson Dirichlet process (PDP) to model the progressive dependency between consecutive part of a document, which can capture the continuity of topical flow in a document nicely but does not capture the topic change explicitly. [Fox et al. \(2008\)](#) propose Sticky HDP-HMM, which is an extension of HDP-HMM ([Teh et al., 2006](#)) for the problem of speaker diarization involving segmenting an audio recording into intervals associated with individual speakers. Applying to the conversational setting, Sticky HDP-HMM associates each turn with a single topic; this is a strong assumption since people tend to talk about more than one thing in a turn, especially in political debates. We will, however, use it as one of the baselines in our topic segmentation experiment (Section 3.6). Other more recent works introduce model to perform topic segmentation in various settings including emails ([Joty et al., 2010, 2011](#)), book chapters and novels ([Du et al., 2012, 2013](#)).

However, many of these methods do not explicitly model the changes of the topics within a document or conversation. To address this, we endow each turn with a binary latent variable $l_{c,t}$, called the *topic shift indicator* ([Purver et al., 2006](#)). This latent variable signifies whether in this turn the speaker changed the topic of the conversation. In addition, to capture the agenda control behavior of the speakers across multiple conversations in the corpus, we further associate each

speaker m with a latent *topic shift tendency* denoted by π_m . Intuitively, this variable is intended to capture the propensity of a speaker to effect a topic shift. Formally, it represents the probability that the speaker m will change the topic (distribution) of a conversation. In the remainder of this section, we will describe the model in more detail together with the inference techniques we use.

3.2.2 Generative Process of SITS

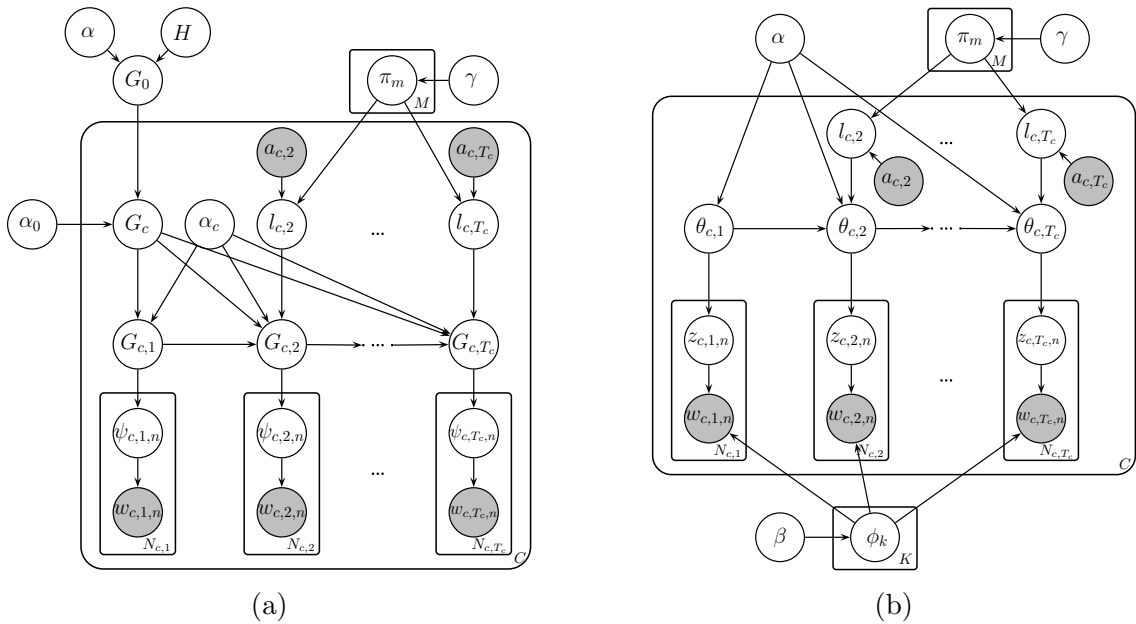


Figure 3.1: Plate diagrams of our proposed models: (a) nonparametric SITS; (b) parametric SITS. Nodes represent random variables (shaded nodes are observed); lines are probabilistic dependencies. Plates represent repetition. The innermost plates are turns, grouped in conversations.

As in the HDP (Teh et al., 2006), we allow an unbounded number of topics to be shared among the turns of the corpus. Topics are drawn from a base distribution H over multinomial distributions over the vocabulary of size V ; H is a finite Dirichlet distribution with symmetric prior λ . Unlike HDP, where every document (here,

every turn) independently draws a new multinomial distribution from a Dirichlet process, the social and temporal dynamics of a conversation, as specified by the binary topic shift indicator $l_{c,t}$, determine when new draws happen. The hierarchy of Dirichlet processes allows statistical strength to be shared across contexts; within a conversation and across conversations. The per-speaker topic shift tendency π_m allows *speaker identity* to influence the evolution of topics.

Generative process: The formal generative process (Figure 3.1) is

1. For speaker $m \in [1, M]$, draw speaker topic shift probability $\pi_m \sim \text{Beta}(\gamma)$
2. Draw the global topic distribution $G_0 \sim \text{DP}(\alpha, H)$
3. For each conversation $c \in [1, C]$
 - (a) Draw a conversation-specific topic distribution $G_c \sim \text{DP}(\alpha_0, G_0)$
 - (b) For each turn $t \in [1, T_c]$ with speaker $a_{c,t}$
 - i. If $t = 1$, set the topic shift indicator $l_{c,t} = 1$. Otherwise, draw $l_{c,t} \sim \text{Bernoulli}(\pi_{a_{c,t}})$.
 - ii. If $l_{c,t} = 1$, draw $G_{c,t} \sim \text{DP}(\alpha_c, G_c)$. Otherwise, set $G_{c,t} \equiv G_{c,t-1}$.
 - iii. For each word index $n \in [1, N_{c,t}]$
 - Draw a topic $\psi_{c,t,n} \sim G_{c,t}$
 - Draw a token $w_{c,t,n} \sim \text{Multinomial}(\psi_{c,t,n})$

Intuitively, SITS generates a conversation as follows: At the beginning of a conversation c , the first speaker $a_{c,1}$ draws a distribution over topics $G_{c,1}$ from the base distribution, and uses that topic distribution to draw a topic $\psi_{c,1,n}$ for each

token $w_{c,1,n}$. Subsequently, at turn t , speaker $a_{c,t}$ will first flip a speaker-specific biased coin $\pi_{a_{c,t}}$ to decide whether $a_{c,t}$ will change the topic of the conversation. If the coin comes up tails ($l_{c,t} = 0$), $a_{c,t}$ will not change the conversation topic and uses the previous turn’s topic distribution $G_{c,t-1}$ to generate turn t ’s tokens. If, on the other hand, the coin comes up heads ($l_{c,t} = 1$), $a_{c,t}$ will change the topic by drawing a new topic distribution $G_{c,t}$ from the conversation-specific collection of topics $\text{DP}(\alpha_c, G_c)$.

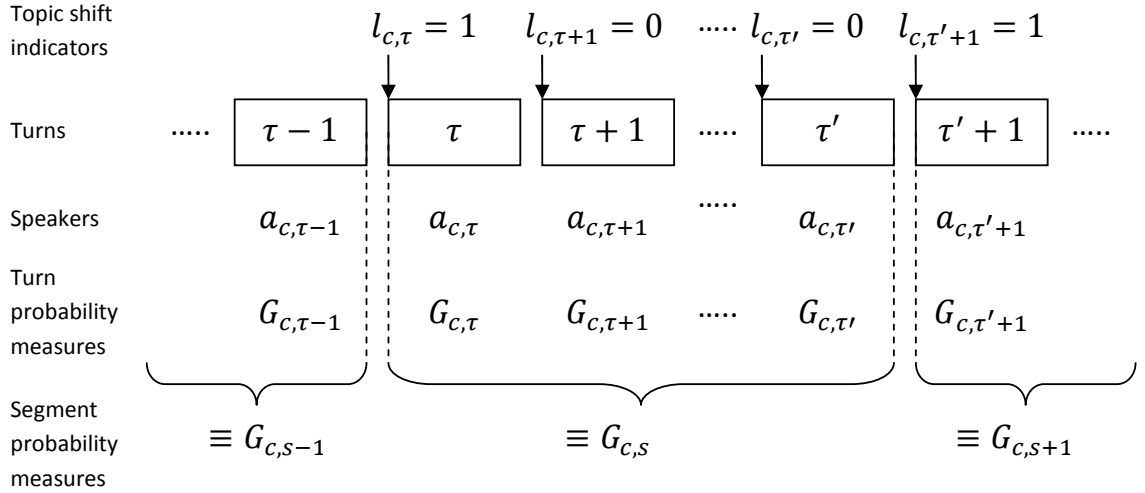


Figure 3.2: Diagram of notation for topic shift indicators and conversation segments: Each turn is associated with a latent binary variable *topic shift indicator* l specifying whether the topic of the turn is shifted. In this example, topic shifts occur in turns τ and $\tau' + 1$. As a result, the topic shift indicators of turn τ and $\tau' + 1$ are equal to 1 (i.e. $l_{c,\tau} = l_{c,\tau'+1} = 1$) and the topic shift indicators of all turns in between are 0 (i.e. $l_{c,t} = 0, \forall t \in [\tau + 1, \tau']$). Turns $[\tau, \tau']$ form a segment s in which all topic distributions $G_{c,\tau}, G_{c,\tau+1}, \dots, G_{c,\tau'}$ are the same and are denoted collectively as $G_{c,s}$.

Segmentation Notation: To make notation more concrete and to connect our model with topic segmentation, we introduce the notion of *segments* in a conversation. A

segment s of conversation c is a sequence of turns $[\tau, \tau']$ such that

$$\begin{cases} l_{c,\tau} = l_{c,\tau'+1} = 1 \\ l_{c,t} = 0, \forall t \in [\tau + 1, \tau'] \end{cases}$$

When $l_{c,t} = 0$, $G_{c,t}$ is the same as $G_{c,t-1}$ and all topics (i.e. multinomial distributions over words) $\{\psi_{c,t,n} \mid n \in [1, N_{c,t}]\}$ that generate words in turn t and the topics $\{\psi_{c,t-1,n'} \mid n' \in [1, N_{c,t-1}]\}$ that generate words in turn $t - 1$ come from the same distribution. Thus, all topics used in a segment s are drawn from a single segment-specific probability measure $G_{c,s}$,

$$G_{c,s} \mid l_{c,1}, l_{c,2}, \dots, l_{c,T_c}, \alpha_c, G_c \sim \text{DP}(\alpha_c, G_c) \quad (3.1)$$

A visual illustration of these notations can be found in Figure 3.2. For notational convenience, S_c denotes the number of segments in conversation c , and s_t denotes the segment index of turn t . We emphasize that all segment-related notations are derived from the posterior over the topic shifts l and are not part of the model itself.

3.2.3 Posterior Inference for SITS

To find the latent variables that best explain observed data, we use Gibbs sampling (Neal, 2000; Resnik and Hardisty, 2010). The state space in our Gibbs sampler consists of the latent variables for topic indices assigned to all tokens $\mathbf{z} = \{z_{c,t,n}\}$ and topic shifts assigned to turns $\mathbf{l} = \{l_{c,t}\}$. We marginalize over all other latent variables. For each iteration of the sampling process, we loop over each turn

in each conversation. For a given turn t in conversation c , we first sample the topic shift indicator variable $l_{c,t}$ and then sample the topic assignment $z_{c,t,n}$ for each token in the turn.

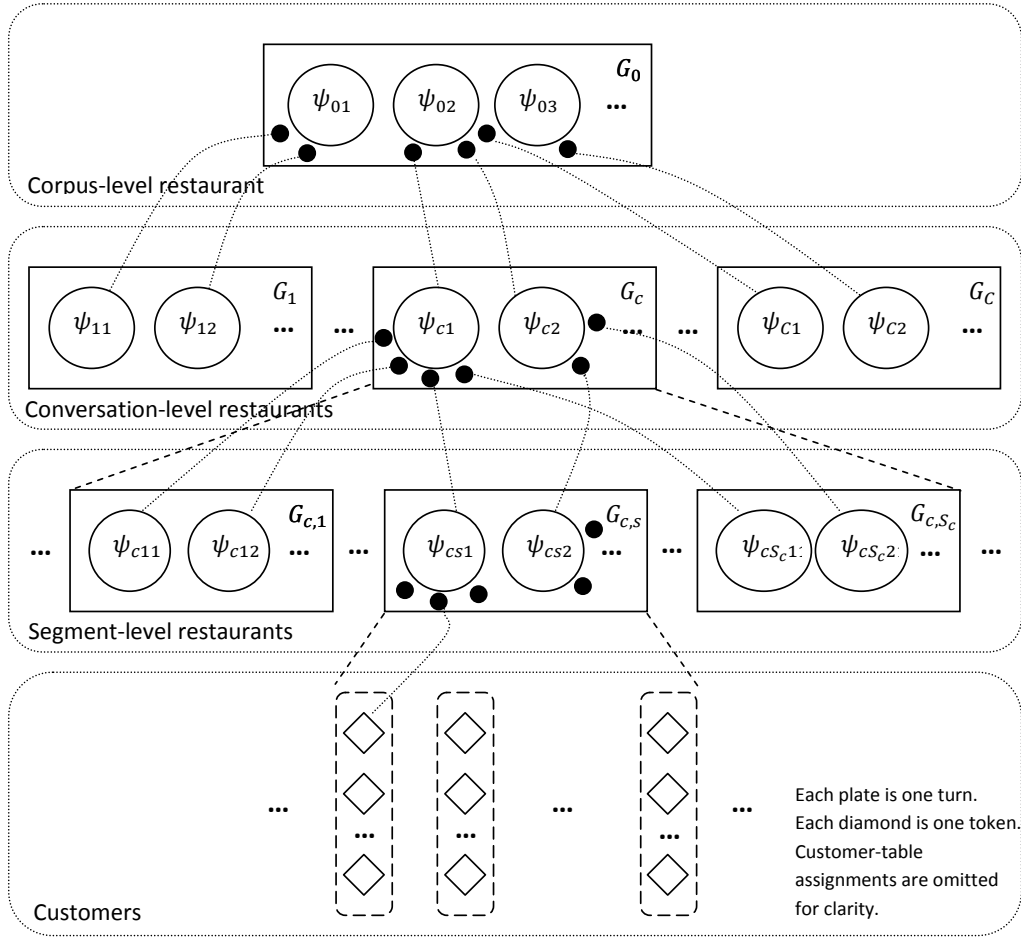


Figure 3.3: Illustration of topic assignments in our inference algorithm. Each solid rectangle represents a restaurant (i.e., a topic distribution) and each circle represents a table (i.e., a topic). To assign token n of turn t in conversation c to a table $z_{c,t,n}$ in the corpus-level restaurant, we need to sample a path assigning the token to a segment-level table, the segment-level table to a conversation-level table and the conversation-level table to a globally shared corpus-level table.

Sampling Topic Assignments In Bayesian nonparametrics, the Chinese restaurant process (CRP) metaphor is often used to explain the clustering effect of the Dirichlet process (Ferguson, 1973). As reviewed in Chapter 2, the CRP is an exchangeable

distribution over partitions of integers, which facilitates Gibbs sampling (Neal, 2000) (as we will see in Equation 3.2). When used in topic models, each Chinese restaurant consists of infinite number of tables, each of which corresponds to a topic. Customers, each of which is a token, are assigned to tables and if two tokens are assigned to the same table: they share the same topic.

The CRP has a “rich get richer” property, which means that tables with many customers will attract yet more customers—a new customer will sit at an existing table with probability proportional to the number of customers currently at the table. The CRP has no limit on the number of tables; when a customer needs to be seated, there is always a probability—proportional to the Dirichlet parameter α —that it will be seated at a new table. When a new table is formed, it is assigned a “dish”; this is a draw from the Dirichlet process’s base distribution. In a topic model, this atom associated with a new table is a multinomial distribution over word types. In a standard, non-hierarchical CRP, this multinomial distribution comes from a Dirichlet distribution.

But it doesn’t have to—hierarchical nonparametric models extend the metaphor further by introducing a hierarchy of restaurants (Teh et al., 2006; Teh, 2006), where the base distribution of one restaurant can be another restaurant. This is where things can get tricky. Instead of having a seating assignment, a customer now has a seating *path* and is potentially responsible for spawning new tables in every restaurant. In SITS there are restaurants for the current segment, the conversation, and the entire corpus, as shown in Figure 3.3.

To sample $z_{c,t,n}$, the index of the shared topic assigned to token n of turn

t in conversation c , we need to sample the *path* assigning each word token to a segment-level table, each segment-level table to a conversation-level table and each conversation-level table to a shared dish. Before describing the sampling equations, we introduce notation denoting the counts:

- $N_{c,s,k}$: number of tokens in segment s in conversation c assigned to dish k
- $N_{c,k}$: number of segment-level tables in conversations c assigned to dish k
- N_k : number of conversation-level tables assigned to dish k

Note that we use k to index the global topics shared across the corpus, each of which corresponds to a dish in the corpus-level restaurant. In general, computing the exact values of these counts makes bookkeeping rather complicated. Since there might be multiple tables at a lower-level restaurant assigned to the same table at the higher-level restaurant, to compute the correct counts, we need to sum the number of customers over all these tables. For example, in Figure 3.3, since both $\psi_{c,1}$ and $\psi_{c,2}$ are assigned to $\psi_{0,2}$ (i.e., $k = 2$), to compute $N_{c,k}$ we have to sum over the number of customers currently assigned to $\psi_{c,1}$ and $\psi_{c,2}$ (which are 4 and 2 respectively in this example).

To mitigate this problem of bookkeeping and to speed up the sampling process, we use the *minimal path assumption* (Cowans, 2006; Wallach, 2008) to generate the path assignments.⁴ Under the minimal path assumption, a new table in a restaurant is created only when there is no table already serving the dish. In other words in a

⁴We also investigated using the maximal assumption and fully sampling assignments. We found the minimal path assumption worked as well as explicitly sampling seating assignments and that the maximal path assumption worked less well. Another, more complicated, sampling method is to sample the counts $N_{c,k}$ and N_k according to their corresponding Antoniak distributions (Antoniak, 1974), similar to the direct assignment sampling method described in Teh et al. (2006).

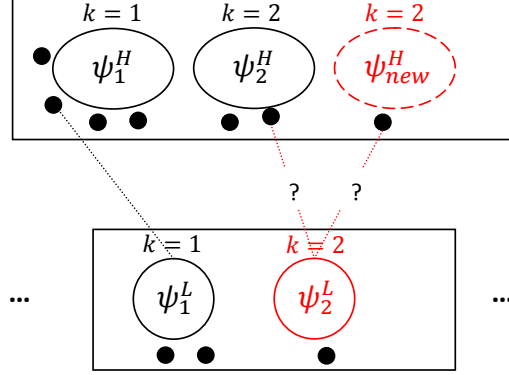


Figure 3.4: Illustration of minimal path assumption. This figure shows an example of the seating assignments in a hierarchy of Chinese restaurants of a higher-level restaurant and a lower-level restaurant. Each table in the lower restaurant is assigned to a table in the higher restaurant and tables on the same path serve the same dish k . When sampling the assignment for table ψ_2^L in the lower restaurant, given that dish $k = 2$ is assigned to this table, there are two options for how the table in the higher restaurant could be selected. It could be an existing table ψ_2^H or a new table ψ_{new}^H , both serving dish $k = 2$. Under the minimal path assumption, it is always assigned to an existing table (if possible) and only assigned to a new table if there is no table with the given dish. In this case, the minimal path assumption will assign ψ_2^L to ψ_2^H .

restaurant, there is at most one table serving a given dish. A more detailed example of the minimal path assumption is illustrated in Figure 3.4. Using this assumption, in the example shown in Figure 3.3, $\psi_{c,1}$ and $\psi_{c,2}$ will be merged together since they are both assigned to $\psi_{0,2}$.

Now that we have introduced our notations, the conditional distribution for $z_{c,t,n}$ is $P(z_{c,t,n} \mid w_{c,t,n}, \mathbf{z}^{-c,t,n}, \mathbf{w}^{-c,t,n}, \mathbf{l}, *) \propto$

$$P(z_{c,t,n} \mid \mathbf{z}^{-c,t,n})P(w_{c,t,n} \mid z_{c,t,n}, \mathbf{w}^{-c,t,n}, \mathbf{l}, *) \quad (3.2)$$

The first factor is the prior probability of assigning to a path according to the

minimal path assumption (Wallach, 2008, p. 60),

$$P(z_{c,t,n} = k \mid \mathbf{z}^{-c,t,n}) \propto \frac{N_{c,st,k}^{-c,t,n} + \alpha_c \frac{N_{c,k}^{-c,t,n} + \alpha_0 \frac{N_{c,t,n}^{-c,t,n} + \alpha \frac{1}{K^+}}{N_{c,t,n}^{-c,t,n} + \alpha}}{N_{c,st,\cdot}^{-c,t,n} + \alpha_c}, \quad (3.3)$$

where K^+ is the current number of shared topics.⁵ Intuitively, Equation 3.3 computes the probability of token $w_{c,t,n}$ being generated from a shared topic k . This probability is proportional to $N_{c,st,k}$ —the number of customers sitting at table serving dish k at restaurant $G_{c,st}$, smoothed by the probability of generating this token from the table serving dish k at the higher-level restaurant (i.e., restaurant G_c). This smoothing probability is computed in the same hierarchical manner until the top restaurant is reached, where the base distribution over topics is uniform and the probability of picking a topic is equal to $1/K^+$. Equation 3.3 also captures the case where a table is empty; when the number of customers on that table is zero, the probability of generating the token from the corresponding topic relies entirely on the smoothing probability from the higher-level restaurant’s table.

The second factor is the data likelihood. After integrating out all ψ ’s, we have

$$P(w_{c,t,n} = w \mid z_{c,t,n} = k, \mathbf{w}^{-c,t,n}, \mathbf{l}, *) \propto \begin{cases} \frac{M_{k,w}^{-c,t,n} + \lambda}{M_{k,\cdot}^{-c,t,n} + V\lambda}, & \text{if } k \text{ exists;} \\ \frac{1}{V}, & \text{if } k \text{ is new.} \end{cases} \quad (3.4)$$

Here, $M_{k,w}$ denotes the number of times word type w in the vocabulary is assigned to topic k ; marginal counts are represented with \cdot and $*$ represents all hyperparameters;

⁵The superscript $+$ is to denote that this number is unbounded and varies during the sampling process.

V is the size of the vocabulary, and the superscript $^{-c,t,n}$ denotes the same counts excluding $w_{c,t,n}$.

Sampling Topic Shift Indicators Sampling the topic shift variable $l_{c,t}$ requires us to consider merging or splitting segments. We define the following notation:

- $\mathbf{k}_{c,t}$: the shared topic indices of all tokens in turn t of conversation c .
- $S_{a_{c,t},x}$: the number of times speaker $a_{c,t}$ is assigned the topic shift $x \in \{0, 1\}$.
- $J_{c,s}^x$: the number of topics in segment s of conversation c if $l_{c,t} = x$
- $N_{c,s,j}^x$: the number of tokens assigned to the segment-level topic j when $l_{c,t} = x$.⁶

Again, the superscript $^{-c,t}$ denotes the exclusion of turn t of conversation c in the corresponding counts.

Recall that the topic shift is a binary variable. We use 0 to represent the “no shift” case, i.e. when the topic distribution is identical to that of the previous turn.

We sample this assignment with the following probability:

$$P(l_{c,t} = 0 \mid \mathbf{l}^{-c,t}, \mathbf{w}, \mathbf{k}, \mathbf{a}, *) \propto \frac{S_{a_{c,t},0}^{-c,t} + \gamma}{S_{a_{c,t},\cdot}^{-c,t} + 2\gamma} \times \frac{\alpha_c^{J_{c,s_t}^0} \prod_{j=1}^{J_{c,s_t}^0} (N_{c,s_t,j}^0 - 1)!}{\prod_{x=1}^{N_{c,s_t,\cdot}^0} (x - 1 + \alpha_c)} \quad (3.5)$$

In Equation 3.5, the first factor is proportional to the probability of assigning a topic shift of value 0 to speaker $a_{c,t}$ and the second factor is proportional to the

⁶Deterministically knowing the path assignments is the primary efficiency motivation for using the minimal path assumption. The alternative is to explicitly sample the path assignments, which is more complicated (for both notation and computation).

joint probability of all topics in segment s_t of conversation c when $l_{c,t} = 0$.⁷

The other alternative is for the topic shift to be 1, which represents the introduction of a new distribution over topics *inside* an existing segment. The probability of sampling this assignment is:

$$P(l_{c,t} = 1 \mid \mathbf{l}^{-c,t}, \mathbf{w}, \mathbf{k}, \mathbf{a}, *) \propto \frac{S_{a_{c,t},1}^{-c,t} + \gamma}{S_{a_{c,t},\cdot}^{-c,t} + 2\gamma} \times \left(\frac{\alpha_c^{J_{c,(s_t-1)}^1} \prod_{j=1}^{J_{c,(s_t-1)}^1} (N_{c,(s_t-1),j}^1 - 1)!}{\prod_{x=1}^{N_{c,(s_t-1),\cdot}^1} (x - 1 + \alpha_c)} \frac{\alpha_c^{J_{c,s_t}^1} \prod_{j=1}^{J_{c,s_t}^1} (N_{c,s_t,j}^1 - 1)!}{\prod_{x=1}^{N_{c,s_t,\cdot}^1} (x - 1 + \alpha_c)} \right) \quad (3.6)$$

As above, the first factor in Equation 3.6 is proportional to the probability of assigning a topic shift of value 1 to speaker $a_{c,t}$; the second factor in the big bracket is proportional to the joint distribution of the topics in segments $s_t - 1$ and s_t . In this case, $l_{c,t} = 1$ means splitting the current segment, which results in two joint probabilities for two segments.

3.3 Data Collections and Annotations

| Datasets | Speakers | Conversations | Annotations | Content |
|-----------------------|----------|---------------|-------------|-------------|
| 2008 Debates | 9 | 4 | topics | politics |
| 2012 Debates | 40 | 9 | none | politics |
| <i>Crossfire</i> | 2567 | 1134 | influencer | politics |
| ICSI Meetings | 60 | 75 | topics | engineering |
| Wikipedia discussions | 604 | 1991 | influencer | varied |

Table 3.1: Summary of datasets detailing how many distinct speakers are present, how many distinct conversations are in the corpus, the annotations available, and the general content of the dataset.

We validate our approach using five different datasets shown in Table 3.1.

⁷Refer to (Gershman and Blei, 2012) for a detailed derivation of this joint probability.

First, in Section 3.4, we qualitatively evaluate the effectiveness of SITS on capturing the agenda control behavior of candidates in two sets of debates: the 2008 presidential election debates and the 2012 Republican primary debates. We also analyze the behavior of participants in a large number of CNN’s *Crossfire* shows. We then quantitatively evaluate SITS on two computational tasks: influencer detection in Section 3.5 and topic segmentation in Section 3.6. For influencer detection, we collaborate with researchers in communication to annotate influencers in a set of Wikipedia discussion pages and *Crossfire* shows. For topic segmentation, we use the ICSI meeting corpus which is a commonly used dataset for topic segmentation, and the 2008 presidential election debates which was manually annotated by domain experts.

3.3.1 Datasets

We first describe the datasets that we use in our experiments. For all datasets, we tokenize texts using OpenNLP and remove common stopwords.⁸ After that, we remove turns that are very short since they do not contain much information content-wise and most likely there is no topic shift during these turns. We empirically remove turns that have fewer than 5 tokens after removing stopwords.

The 2008 Presidential Election Debates: Our first dataset contains three annotated presidential debates between Barack Obama and John McCain and a vice presidential debate between Joe Biden and Sarah Palin. Each turn is one of two types:

⁸<http://opennlp.apache.org/>

| Speaker | Type | Turn clauses | T_Q | T_R |
|---------|----------|--|-------|-------|
| Brokaw | <i>Q</i> | Sen. Obama, time for a discussion. I'm going to begin with you. Are you saying to Mr. Clark and to the other members of the American television audience that the American economy is going to get much worse before it gets better and they ought to be prepared for that? | 1 | N/A |
| Obama | <i>R</i> | No, I am confident about the American economy. | 1 | 1 |
| | | But most importantly, we're going to have to help ordinary families be able to stay in their homes, make sure that they can pay their bills, deal with critical issues like health care and energy, and we're going to have to change the culture in Washington so that lobbyists and special interests aren't driving the process and your voices aren't being drowned out. | 1 | 14 |
| Brokaw | <i>Q</i> | Sen. McCain, in all candor, do you think the economy is going to get worse before it gets better? | 1 | N/A |
| McCain | <i>R</i> | I think if we act effectively, if we stabilize the housing market—which I believe we can, | 1 | 14 |
| | | if we go out and buy up these bad loans, so that people can have a new mortgage at the new value of their home | 1 | 14 |
| | | I think if we get rid of the cronyism and special interest influence in Washington so we can act more effectively. | 1 | 20 |

Table 3.2: Example turns from the 2008 election debates annotated by [Boydston et al. \(2013a\)](#). Each clause in a turn is manually coded with a *Question Topic Code* (T_Q) and a *Response Topic Code* (T_R). The topic codes (T_Q and T_R) are from the Policy Agendas Topics Codebook. In this example, the following topic codes are used: Macroeconomics (1), Housing & Community Development (14), Government Operations (20).

questions (Q) from the moderator or *responses* (R) from a candidate. Each clause in a turn is coded by [Boydston et al. \(2013a\)](#) with a *Question Topic Code* (T_Q) and a *Response Topic Code* (T_R). Thus, a turn has a list of T_Q 's and T_R 's both of length equal to the number of clauses in the turn. Topics are from the Policy Agendas Topics Codebook, a widely used inventory containing codes for 19 major topics and 225 subtopics.⁹ Table 3.2 shows an example annotation.

To obtain reference segmentations in debates, we assign each turn a real value from 0 to 1 indicating how much a turn changes the topic. For a question-typed turn, the score is the fraction of clause topic codes not appearing in the previous turn; for response-typed turns, the score is the fraction of clause topic codes that do not appear in the corresponding question. This results in a set of *non-binary* reference segmentations. For evaluation metrics that require binary segmentations, we create a binary segmentation by labeling a turn as a segment boundary if the computed score is 1. This threshold is chosen to include only true segment boundaries. After preprocessing, this dataset contains 9 unique speakers and the vocabulary contains 1,761 non-stopword tokens.

The 2012 Republican Primary Debates: We also downloaded nine transcripts in the 2012 Republican Party presidential debates, whose information is shown in Table 3.3. Since the transcripts are pulled from different sources, we perform a simple entity resolution step using edit distance to merge duplicate participants' names. For example, "Romney", "Mitt Romney" are resolved into "Romney"; "Paul", "Rep.

⁹<http://www.policyagendas.org/page/topic-codebook>

Paul”, “Representative Ron Paul R-TX” are resolved into “Paul” etc. We also merge anonymous participants such as “Unidentified Female”, “Unidentified Male”, “Question”, “Unknown” etc into a single participant named “Audience”. After pre-processing, there are 40 unique participants in these 9 debates including candidates, moderators and audience members. This dataset is not annotated and we only use it for qualitative evaluation.

| Date | Place | Sponsor | Participants |
|--------------|------------------|----------------|---|
| 13 Jun. 2011 | Goffstown, NH | CNN | Bachmann, Cain, Gingrich, Paul, Pawlenty, Romney, Santorum |
| 12 Sep. 2011 | Tampa, FL | CNN | Bachmann, Cain, Gingrich, Huntsman, Paul, Perry, Romney, Santorum |
| 18 Oct. 2011 | Las Vegas, NV | CNN | Bachmann, Cain, Gingrich, Paul, Perry, Romney, Santorum |
| 09 Nov. 2011 | Rochester, MI | CNBC | Bachmann, Cain, Gingrich, Huntsman, Paul, Perry, Romney, Santorum |
| 22 Nov. 2011 | Washington, DC | CNN | Bachmann, Cain, Gingrich, Huntsman, Paul, Perry, Romney, Santorum |
| 19 Jan. 2012 | Charleston, SC | CNN | Gingrich, Paul, Romney, Santorum |
| 23 Jan. 2012 | Tampa, FL | NBC | Gingrich, Paul, Romney, Santorum |
| 26 Jan. 2012 | Jacksonville, FL | CNN | Gingrich, Paul, Romney, Santorum |
| 22 Feb. 2012 | Mesa, AZ | CNN | Gingrich, Paul, Romney, Santorum |

Table 3.3: List of the 9 Republican Party presidential debates used.

CNN’s *Crossfire*: *Crossfire* is a weekly U.S. television “talking heads” program engineered to incite heated arguments (hence the name). Each episode features two recurring hosts, two guests, and clips from the week’s news. Our *Crossfire* dataset contains 1134 transcribed episodes aired between 2000 and 2004.¹⁰ There are 2,567 unique speakers and the vocabulary size is 16,791. Unlike the previous two datasets, *Crossfire* does not have explicit topic segmentations, so we use it to explore speaker-specific characteristics (Section 3.4.2).

¹⁰<http://www.cnn.com/TRANSCRIPTS/cf.html>

The ICSI Meeting Corpus: The ICSI Meeting Corpus consists of 75 transcribed meetings at the International Computer Science Institute in Berkeley, California (Janin et al., 2003). Among these, 25 meetings were annotated with reference segmentations (Galley et al., 2003). Segmentations are *binary*, i.e., each point in the document is either a segment boundary or not, and on average each meeting has 8 segment boundaries. We use this dataset for evaluating topic segmentation (Section 3.6). After preprocessing, there are 60 unique speakers and the vocabulary contains 3346 non-stopword tokens.

| | |
|----|--|
| A: | The current lead sentence has been agreed upon by many - I know, I was embroiled in the huge debate that developed into the current lead. However, the sentence is still kinda awkward - even though it captures the broader essence of evolutionary theory. I would like to propose an alternate (below), because there is a problem with the way that the term change is used, as Kirk J. Fitzhugh has noted: "Change is not the pertinent quality of interest in evolution". Hence: Evolution is the gradual departure across successive generations in the constituency of the inherited characteristics of organisms in biological populations. |
| B: | No thank you, this is just more obscurantism. |
| A: | It's wp:V, not obscurantism, consistent with the history of the science. Not much thought goes into conceiving that "Evolution is change", but if you are asked to think past this and call it obscurantism in your critique, it is a strange response. Obscurantism: "is the practice of deliberately preventing the facts or the full details of some matter from becoming known" - ironic that this applies more aptly to your rejection. |
| B: | Your obsession with providing the most scientifically accurate and current definition of evolution prevents the average reader from having a chance at understanding this article. That is obscurantism. It is not WPV, because that definition is not by a longshot the most commonly used, and specifically it is entirely unsuited for works meant to be read by lay readers. |
| C: | This is a general encyclopedia, not a graduate level evolutionary biology course. Keeping it simple so that people can understand what we write without having an advanced degree is a good thing. So no, let's keep the lead as is. |

Table 3.4: Example of a Wikipedia discussion in our dataset.

Wikipedia Discussions: Each article on Wikipedia has a related discussion page so that the individuals writing and editing the article can discuss the content, editorial

decisions, and the application of Wikipedia policies (Butler et al., 2008). Unlike the other situations, Wikipedia discussions are not spoken conversations that have been transcribed. Instead, these conversations are written asynchronously.

However, Wikipedia discussions have much of the same properties as our other corpora. Contributors have different levels of responsibility and prestige, and many contributors are actively working to persuade the group to accept their proposed policies (for an example, see Table 3.4), other contributors are attempting to maintain civility, and other contributors are attacking their ostensible collaborators. Unlike spoken conversations, Wikipedia discussions lack social norms that prevent an individual from writing as often or as much as they want. This makes common techniques such as counting turns or turn lengths less helpful measures to discover who influencers are.

3.4 Evaluating Agenda Control

In this section, we focus on the ability of SITS to capture the extent to which individual speakers affect topic shifts in conversations. Recall that SITS associates with each speaker a topic shift tendency π that represents the probability of changing the topic in the conversation. While topic segmentation is a well studied problem, hence the evaluation in Section 3.6, there are no established quantitative measurements of an individual’s ability to control a conversation. To evaluate whether the tendency is capturing meaningful characteristics of speakers, we look qualitatively at the behavior of the model.

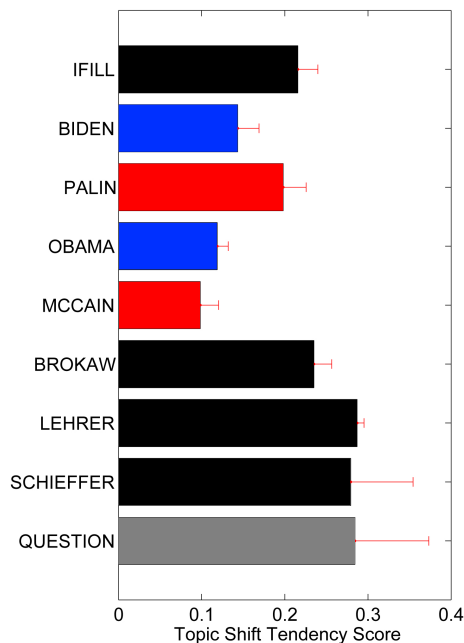


Figure 3.5: Topic shift tendency π of speakers in the 2008 Presidential Election Debates (larger means greater tendency). IFILL was the moderator in the vice presidential debate between BIDEN and PALIN; BROKAW, LEHRER and SCHIEFFER were the moderators in the three presidential debates between OBAMA and MCCAIN; QUESTION collectively refers to questions from the audiences. Colors denote **Republicans**, **Democrats**, **Moderators**, and **Audiences**.

3.4.1 2008 Election Debates

To obtain a posterior estimate of π (Figure 3.5) we create 10 chains with hyperparameters sampled from the uniform distribution $U(0, 1)$ and average π over 10 chains (as described in Section 3.6.1). In these debates, Ifill is the moderator of the debate between Biden and Palin; Brokaw, Lehrer and Schieffer are the three moderators of the three debates between Obama and McCain. Here “Question” denotes questions from audiences in “town hall” debate. The role of this “speaker” can be considered equivalent to the debate moderator.

The topic shift tendencies of moderators are generally much higher than for candidates. In the three debates between Obama and McCain, the moderators—Brokaw, Lehrer and Schieffer—have significantly higher scores than both candidates. This is a useful reality check, since in a debate the moderators are the ones asking questions and literally controlling the topical focus. Similarly, the “Question”

speaker had a relatively high variance, consistent with that “participant” in the model as an amalgamation of many distinct speakers.

Interestingly, however, in the vice-presidential debate, the score of moderator Ifill is higher than the candidates’ scores only by a small margin, and it is indistinguishable from the degree of topic control displayed by Palin. Qualitatively, the assessment of the model is consistent with widespread perceptions and media commentary at the time that characterized Ifill as a weak moderator. For example, *Harper’s Magazine’s* Horton (2008) discusses the context of the vice-presidential debate, in particular the McCain campaign’s characterization of Ifill as a biased moderator because she “was about to publish a book entitled *The Breakthrough* that discusses Barack Obama, and a number of other black politicians, achieving national prominence”. According to Horton:

“First, the charges against Ifill would lead to her being extremely passive in her questioning of Palin and permissive in her moderating the debate. Second, the charge of bias against Ifill would enable Palin to simply skirt any questions she felt uncomfortable answering and go directly to a pre-rehearsed and nonresponsive talking point. This strategy succeeded on both points.”

Similarly, Fallows (2008) of *The Atlantic* included the following in his “quick guide” remarks on the debate:

Ifill, moderator: *Terrible*. Yes, she was constrained by the agreed debate rules. But she gave not the slightest sign of chafing against them or looking

for ways to follow up the many unanswered questions or self-contradictory answers. This was the big news of the evening . . .

Palin: “*Beat expectations.*” In every single answer, she was obviously trying to fit the talking points she had learned to the air time she had to fill, knowing she could do so with impunity from the moderator.

That said, our quantitative modeling of topic shift tendency suggests that all candidates managed to succeed at some points in setting and controlling the topic of conversation in the debates. In the presidential debates, our model gives Obama a slightly higher score than McCain, consistent with social science claims that Obama had the lead in setting the agenda over McCain (Boydston et al., 2013a). Table 3.5 shows some examples of SITS-detected topic shifts.

3.4.2 *Crossfire*

The *Crossfire* dataset has many more speakers than the presidential and vice-presidential debates. This allows us to examine more closely what we can learn about speakers’ topic shift tendency and ask additional questions; for example, assuming that changing the topic is useful for a speaker, how can we characterize who does so effectively? In our analysis, we take advantage of properties of the *Crossfire* data to examine the relationship between topic shift tendency, social roles, and political ideology.

In order to focus on frequent speakers, we filter out speakers with fewer than 30 turns. Most speakers have relatively small π , with the mode around 0.3. There

| | Previous turn | Turn detected as shifting topic |
|----------------------|---|--|
| 2008 Debates Dataset | BIDEN: Well, mortgage-holders didn't pay the price [...] Barack Obama pointed out two years ago that there was a <i>subprime mortgage</i> [...] | PALIN: That is not so, but because that's just a quick answer, I want to talk about, again, my record on <i>energy</i> ... When we talk about energy, we need to consider the need to do all that we can to allow this nation to become energy independent [...] |
| | PALIN: Your question to him was whether he supported <i>gay marriage</i> and my answer is the same as his and it is that I do not. | IFILL: Wonderful. You agree. On that note, let's move to <i>foreign policy</i> . You both have sons who are in Iraq or on their way to Iraq. You, Governor Palin, have said that you would like to see a real clear plan for an exit strategy. [...] |
| | MCCAIN: I think that Joe Biden is qualified in many respects. ... | SCHIEFFER: [...] Let's talk about <i>energy</i> and <i>climate control</i> . Every president since Nixon has said what both of you [...] |
| | IFILL: So, Governor, as vice president, there's nothing that you have promised [...] that you wouldn't take off the table because of this <i>financial crisis</i> we're in? | BIDEN: Again, let me—let's talk about those <i>tax breaks</i> . [Obama] voted for an energy bill because, for the first time, it had real support for alternative energy. [...] on eliminating the tax breaks for the oil companies, Barack Obama voted to eliminate them. [...] |
| Crossfire Dataset | PRESS: But what do you say, governor, to Governor Bush and [...] your party who would let politicians and not medical scientists decide what <i>drugs</i> are distributed [...] | WHITMAN: Well I disagree with them on this particular issues [...] that's important to me that George Bush stands for <i>education</i> of our children [...] I care about <i>tax policy</i> , I care about the <i>environment</i> . I care about all the issues where he has a proven record in Texas [...] |
| | WEXLER: [...] They need a Medicare prescription drug plan [...] Talk about schools, [...] Al Gore has got a real plan. George Bush offers us vouchers. Talk about the environment. [...] Al Gore is right on in terms of the majority of Americans, but George Bush [...] | KASICH: [...] I want to talk about choice. [...] George Bush believes that, if schools fail, parents ought to have a chance and an opportunity for success. Gore says “no way” [...] Social Security. George Bush says [...] direct it the way federal employees do [...] Al Gore says “No way” [...] That's real choice. That's real bottom-up, not a bureaucratic approach, the way we run this country. |
| | PRESS: Senator, Senator Breaux mentioned that it's President Bush's aim to start on <i>education</i> [...] [McCain] [...] said he was going to do introduce the legislation the first day of the first week of the new administration. [...] | MCCAIN: After one of closest elections in our nation's history, there is one thing the American people are unanimous about. They want their <i>government</i> back. We can do that by ridding politics of large, unregulated contributions that give special interests a seat at the table while average Americans are stuck in the back of the room. |

Table 3.5: Example of turns designated as a topic shift by SITS. We chose turns to highlight speakers with high topic shift tendency π . Some keywords are manually italicized to highlight the topics discussed.

| Rank | Speaker | π | Rank | Speaker | π |
|------|-----------------------------|-------|------|-----------------------------|-------|
| 1 | ANNOUNCER | .884 | 10 | JOHN KASICH | .570 |
| 2 | MALE | .876 | 11 | JAMES CARVILLE [†] | .550 |
| 3 | QUESTION | .755 | 12 | TUCKER CARLSON [†] | .550 |
| 4 | GEORGE W. BUSH [‡] | .751 | 13 | PAUL BEGALA [†] | .545 |
| 5 | BILL PRESS [†] | .651 | 14 | CHRISTINE T. WHITMAN | .533 |
| 6 | FEMALE | .650 | 15 | TERRY MCAULIFFE | .529 |
| 7 | AL GORE [‡] | .650 | 16 | MARY MATALIN [†] | .527 |
| 8 | NARRATOR [‡] | .642 | 17 | JOHN MCCAIN | .524 |
| 9 | ROBERT NOVAK [†] | .587 | 18 | ARI FLEISCHER | .522 |

Table 3.6: Top speakers by topic shift tendencies from our *Crossfire* dataset. We mark hosts ([†]) and “speakers” who often (but not always) appeared in video clips ([‡]). ANNOUNCER makes announcements at the beginning and at the end of each show; NARRATOR narrates video clips; MALE and FEMALE refer to unidentified male and female respectively; QUESTION collectively refers to questions from the audience across different shows. Apart from those groups, speakers with the highest tendency were political moderates.

are, however, speakers with very high topic shift tendencies. Table 3.6 shows the speakers having the highest values according to SITS.

We find that there are three general patterns for who influences the course of a conversation in *Crossfire*. First, there are structural “speakers” that the show uses to frame and propose new topics. These are audience questions, news clips (e.g. many of Gore’s and Bush’s turns from 2000), and voiceovers. That SITS is able to recover these is reassuring, similar to what it has to say about moderators in the 2008 debates. Second, the stable of regular hosts receives high topic shift tendencies, which is again reasonable given their experience with the format and ostensible moderation roles (though in practice they also stoke lively discussion).

The third category is more interesting. The remaining non-hosts with high topic shift tendency appear to be relative moderates on the political spectrum:

- John Kasich, one of few Republicans to support the assault weapons ban and

who was elected in 2010 as the governor of Ohio, a swing state

- Christine Todd Whitman, former Republican governor of New Jersey, a very Democratic state
- John McCain, who before 2008 was known as a “maverick” for working with Democrats (e.g. Russ Feingold)

Although these observations are at best preliminary and require further investigation, we would conjecture that in *Crossfire*’s highly polarized context, it was the political moderates who pushed back, exerting more control over the agenda of the discussion, rather than going along with the topical progression and framing as posed by the show’s organizers. Table 3.5 shows several detected topic shifts from these speakers. In two of these examples, McCain and Whitman are Republicans disagreeing with President Bush. In the other, Kasich is defending a Republican plan (school vouchers) popular with traditional Democratic constituencies.

3.4.3 2012 Republican Primary Debates

As another qualitative data point, we include in Figure 3.6 the model’s topic shift tendency scores for a subset of nine 2012 Republican primary debates. Although we do not have objective measures to compare against, nor clearly stated contemporary commentary as in the case of Ifill’s performance as moderator, we would argue that the model displays quite reasonable face validity in the context of the Republican race.

For example, among the Republican candidates, Ron Paul is known for tight

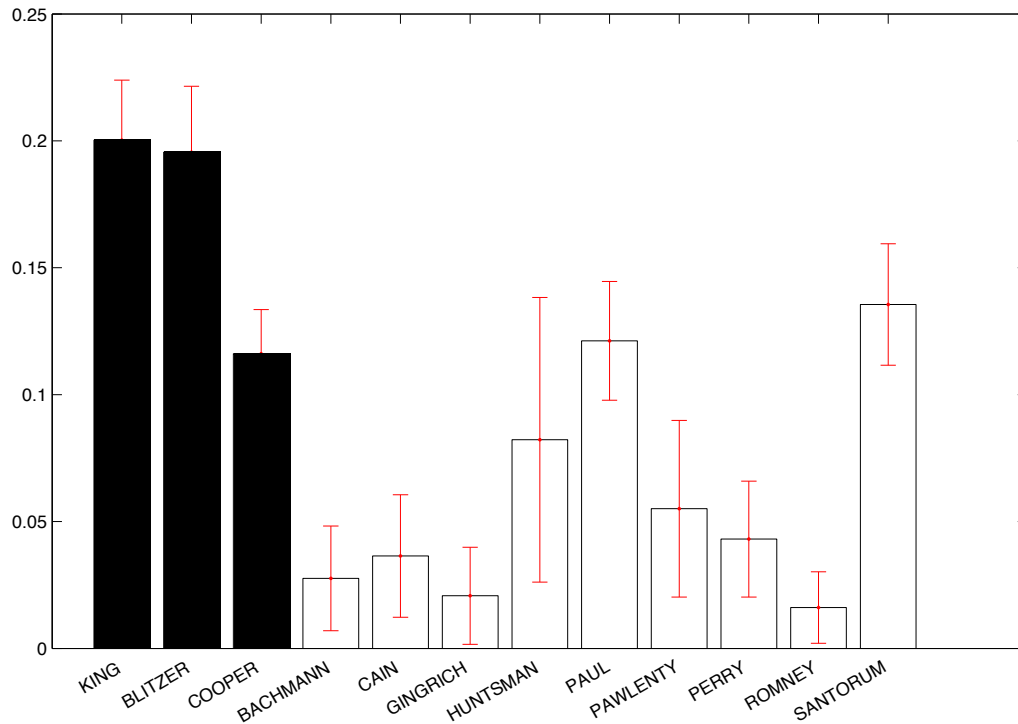


Figure 3.6: Topic shift tendency π of speakers in the 2012 Republican Primary Debates (larger means greater tendency). KING, BLITZER and COOPER are moderators in these debates; the rest are candidates.

focus on a discrete set of arguments associated with his position that “the proper role for government in America is to provide national defense, a court system for civil disputes, a criminal justice system for acts of force and fraud, and little else” (Paul, 2007), often regardless of the specific question that was asked. Similarly, Rick Santorum’s performance in the primary debates tended to include strong rhetoric on social issues. In contrast, Mitt Romney tended to be less aggressive in his responses, arguably playing things safer in a way that was consistent with his general position throughout the primaries as the front-runner.

3.5 Detecting Influencers in Conversations

As motivated in Section 3.1.2, prior research in communication has shown qualitatively that agenda control and management is one of the most effective ways to control the conversation and influence other participants. In this section, we use SITS to quantify how influential each participant is in a conversation, which is then used to detect influencers in the conversation. We collaborate with researchers in communication to annotate influencers in two datasets: *Crossfire* and Wikipedia discussions. In the remaining of this chapter, we first described the process which our collaborators follow to annotate influencers, review related work on influencer detection, and report empirical results showing SITS is more effective than traditional approach in detecting influencers in the two annotated datasets.

3.5.1 Influencer Annotation

In most research on persuasion and power, an influencer attempts to gain compliance from others or uses tactics to shape the opinions, attitudes, or behaviors of others (Scheer and Stern, 1992; Schlenker et al., 1976). In research on social media, such as blogs and Twitter, measurements such as the number of followers or readers serve as a proxy for influence (Alarcon-del Amo et al., 2011; Booth and Matic, 2011; Trammell and Keshelashvili, 2005). Others have studied what influencers say; Drake and Moberg (1986) demonstrated that linguistic influence differs from attempts to influence that rely on power and exchange relationships. In interactions with targets, influencers may rely more on linguistic frames and language than on

resources offered, which is proposed as the requirement for influence by exchange theorists (Blau, 1964; Foa and Foa, 1972; Emerson, 1981).

Our goal in this experiment is to discover who are the influencers in these discussions. We define an influencer as someone who has persuasive ability over where an interaction is headed, what topics are covered, and what positions are espoused within that interaction. In the same way that persuasion shapes, reinforces, or changes attitudes or beliefs, an influencer shapes, reinforces, or changes the direction of the interaction. An influencer within an interaction is someone who may introduce new ideas or arguments into the conversation that others pick up on and discuss (shapes new directions through topic shift), may express arguments about an existing topic that others agree to and further in the discussion (i.e., reinforces the direction), or may provide counter-arguments that others agree to and perpetuate, thereby redirecting where the topic of conversation is headed (i.e., changes the direction of the conversation).

To assess the ability of SITS to discover influencers, our collaborators annotated randomly selected documents from both Wikipedia and *Crossfire* datasets.¹¹ This process proceeded as follows. First, the annotation guidelines for influencers proposed by Bender et al. (2011) is used for Wikipedia discussion. A discussant is considered an influencer if he or she initiated a topic shift that steered the conversation in a different direction, convinced others to agree to a certain viewpoint, or used an authoritative voice that caused others to defer to or reference that person’s expertise. A discussant is not identified as an influencer if he or she merely initi-

¹¹The annotation was done by the last three co-authors in (Nguyen et al., 2014b)

ated a topic at the start of a conversation, did not garner any support from others for the points he or she made, or was not recognized by others as speaking with authority. After annotating an initial set of documents, the annotation guidelines were revised and two independent annotators were retrained until an inter-coder reliability Cohen’s Kappa (Artstein and Poesio, 2008) of 0.8 is reached.¹²

Wikipedia Discussions: Coders first learned to annotate transcripts using data from Wikipedia discussion pages. The two coders annotated over 400 English Wikipedia discussion transcripts for influencer in batches of 20 to 30 transcripts each week. For the English transcripts, each coder annotated the transcripts independently, then annotations were compared for agreement; any discrepancies in the annotations were resolved through discussion of how to apply the coding scheme. After the first four sets of 20 to 30 transcripts, the coders were able to code the transcripts with acceptable intercoder reliability (Cohen’s Kappa > 0.8). Once the coders reached acceptable intercoder reliability for two sets of English data in a row, the coders began independently coding the remaining set of transcripts. Intercoder reliability was maintained at an acceptable level (Cohen’s Kappa > 0.8) for the English transcripts over the subsequent weeks of coding.

Crossfire: After Wikipedia, *Crossfire* data are annotated. Each *Crossfire* episode is split into smaller segments using the “COMMERCIAL_BREAK” tags and each seg-

¹²Kappa was measured based on whether the two annotators agreed on (a) whether there was an influencer, (b) who the primary influencer was, and (c) if there was a secondary influencer. When discrepancies occurred between the annotators, they were resolved through discussion between the annotators and with the supervising researcher. So decisions were not “yes or no” about each speaker; instead, they were about whether or not there was an influencer in each overall interaction, and if so, who the primary and secondary influencers were in a particular interaction.

ment is used as a unit of conversation. The same two coders annotated the *Crossfire* data. To prepare for annotating the *Crossfire* interactions, the coders both annotated the same set of 20 interactions. First the intercoder reliability Cohen’s Kappa was calculated for the agreement between the coders, then any disagreements between the coders were resolved through discussion about the discrepant annotations. The first set of 20 transcripts was coded with a Cohen’s Kappa of 0.65 (before discussion). This procedure was repeated twice; each time the coders jointly annotated 20 transcripts, reliability was calculated, and any discrepancies were resolved through discussion. The third set achieved an acceptable Cohen’s Kappa of 0.8. The remaining transcripts were then split and annotated separately by the two coders. In all, 105 *Crossfire* episode segments were annotated.

3.5.2 Computational Methods for Influencer Detection

Even though influence in conversations has been studied for decades in communication and social psychology, computational methods have only emerged in recent years, thanks to improvements in both quantity and quality of conversational data. As one example, an early computational model to quantify influence between conversational participants (Basu et al., 2001) modeled interactions among a conversational group in a multi-sensor lounge room where people played interactive debating games. In these games, each participant can be in two states: speaker or silent. The model equates each participant with a Markov model. Each participant is allowed to be in either speaking state or silent state at each time step

and the transition from one state to another of an individual is influenced by other participants' states. This allows the model to capture pair-wise interactions among participants in the conversation. [Zhang et al. \(2005\)](#) then extended the work by proposing a model with two-level structure: the participant level, representing the actions of individual participants, and the group level, representing group-level actions. In this setting, the influence of each participant on the actions of the whole group is explicitly captured by the model. These models use expensive features such as prosody and visual cues.

Another popular approach is to treat influencer detection as a supervised classification problem that separates influential individuals from non-influential ones. [Rienks and Heylen \(2006\)](#) focus on extracting a set of structural features that can predict participants' involvement using Support Vector Machines ([Cortes and Vapnik, 1995](#), SVM). Later, [Rienks et al. \(2006\)](#) improved their previous work by extending the set of features to include features capturing topic changes as well as those derived from audio and speech. Again, we do not use any features extracted from audio or visual data, which makes our approach more generalizable. The two most relevant and most useful features extracted from the meeting textual transcripts are number of turns and length of turns, which we use as the baseline in our experiments described in Section [3.5.3](#). [Biran et al. \(2012\)](#) also follow a similar approach to detecting influencers in written online conversations by extracting features to capture different conversational behaviors such as persuasion, agreement/disagreement and dialog patterns.

In this work, we are interested in determining who are the influencers in a

conversation using only the conversation transcripts. We tackle this problem by using an unsupervised ranking approach. It is worth mentioning that, even though we are focused on studying how conversational influence expressed in textual data, there has also been a body of work approaching this problem by studying audio data (Hung et al., 2011), visual data (Otsuka et al., 2006) and both audio-visual activity cues (Jayagopi et al., 2009; Aran and Gatica-Perez, 2010).

Our main purpose in this experimentation is to assess how effective SITS can be in detecting influencers in conversations, especially in comparison with methods based on structural patterns of conversations. We focus on the *influencer detection problem*: *given a speaker in a multi-party conversation, predict whether the speaker is influential*. In the remaining of this section, we describe in details the approach we take, the experimental setups, and the results.

3.5.3 Influencer Detection Problem

The influencer detection problem can be tackled using different methods that can be broadly classified into *classification* and *ranking* approaches. Most previous work follows the classification approach, in which different sets of features are proposed and a classifier is used (Rienks and Heylen, 2006; Rienks et al., 2006; Biran et al., 2012). In this work, we follow the ranking approach.

The ranking approach allows us to focus on individual functions that take a set of individuals and produce an ordering over those individuals from most influential to least influential. The function that produces this ordering is called a *ranking*

method. More specifically, given a speaker a in a conversation c , each ranking method will provide an *influence score* $\mathcal{I}_{a,c}$ that indicates how influential speaker a is in conversation c . We emphasize that, unlike most classification approaches (Rienks and Heylen, 2006; Rienks et al., 2006; Biran et al., 2012), the ranking approach we are focusing on is entirely unsupervised and thus requires no training data.

The ranking approach has a straightforward connection to the classification approach, as each ranking function can be turned into a feature in the supervised classification framework. However, viewing the ranking methods (features) independently allows us to compare and interpret the effectiveness of each feature in isolation. This is useful as an evaluation method because it is independent of the choice of classifier and is less sensitive to the size of training data, which is often a limiting factor in computational social science.

We consider two sets of ranking methods: (1) structure-based methods, which use structural features and (2) topic-change-based methods, which use features extracted from the outputs of SITS.

Structure-based methods score each instance based on features extracted from the structure of the conversation. We use T_c to denote the number of turns in conversation c ; $a_{c,t}$ to denote the speaker that utters turn t in conversation c ; and $N_{c,t}$ to denote the number of tokens in turn t in conversation c .

1. *Number of turns:* assumes that the more turns a speaker has during a conversation, the more influential he or she is. The influence score of this method

is

$$\mathcal{I}_{a,c} = |\{t \in [1, T_c] : a_{c,t} = a\}| \quad (3.7)$$

2. *Total turn lengths*: instead of the number of turns, this method uses the total length of turns uttered by the speaker.

$$\mathcal{I}_{a,c} = \sum_{t \in [1, T_c] : a_{c,t} = a} N_{c,t} \quad (3.8)$$

The two structural features used here capture the activeness of the speakers during a conversation and have been shown to be among the most effective features to detect influencers. These two structure-based methods are appropriate baselines in our experiment since, although being simple, they have been proven to be very effective in detecting influencers, both qualitatively (Bales, 1970) and quantitatively (Rienks et al., 2006; Biran et al., 2012).

Topic-change-based methods score each instance based on features extracted from the posterior distributions of SITS.

1. *Total topic shifts* is the total number of expected topic shifts speaker a makes in conversation c ,

$$\mathcal{I}_{a,c} = \sum_{t \in [1, T_c] : a_{c,t} = a} \bar{l}_{c,t}. \quad (3.9)$$

Recall that in SITS, each turn t in conversation c is associated with a binary latent variable $l_{c,t}$, which indicates whether the topic of turn t is changed or not (these latent variables are introduced in Section 3.2). This expectation is

computed through the empirical average of samples from the Gibbs sampler, $\bar{l}_{c,t}$, after a burn-in period.¹³ Intuitively, the higher $\bar{l}_{c,t}$ is, the more successful the speaker $a_{c,t}$ is in changing the topic of the conversation at this turn t .

2. *Weighted topic shifts* also quantify the topic changes a speaker makes by using the average topic shift indicator $\bar{l}_{c,t}$ but weighted by $(1 - \pi_a)$, where π_a is the topic shift tendency score of the speaker a . The basic idea here is that not all topic shifts should be counted equally. A successful topic shift by a speaker with small topic shift tendency score should be weighted higher than a successful topic by a speaker with high topic shift tendency score. The influence score of this ranking method is defined as

$$\mathcal{I}_{a,c} = (1 - \pi_a) \cdot \sum_{t \in [1, T_c]: a_{c,t} = a} \bar{l}_{c,t} \quad (3.10)$$

3.5.4 Experimental Setup

Datasets: In this experiment, we use two datasets annotated for influencers: *Cross-fire* and Wikipedia discussion pages. These two datasets and the annotation procedures are described in detail in Section 3.5.1. Table 3.7 shows dataset statistics.

Parameter settings and implementation: As before, we use Gibbs sampling with 10 randomly initialized chains for inference. Initial hyperparameter values are sampled from $U(0, 1)$ and statistics are collected after 200 burn-in iterations with a

¹³For more details on how to compute this value, refer to Section 3 of (Resnik and Hardisty, 2010)

| Statistics | <i>Crossfire</i> | Wikipedia |
|---|------------------|-----------|
| Number of conversations | 3391 | 604 |
| Number of unique speakers | 2381 | 1991 |
| Average number of turns per conversation | 38.2 | 12.8 |
| Average number of speakers per conversation | 5 | 7 |
| Number of conversations annotated | 85 | 48 |
| Number of positive instances | 197 | 57 |
| Number of negative instances | 182 | 338 |

Table 3.7: Statistics of the two datasets *Crossfire* and Wikipedia discussions that we annotated influencers. We use these two datasets to evaluate SITS on influencer detection.

lag of 20 iterations over a total of 1000 iterations. Slice sampling optimizes the hyperparameters.

Evaluation measurements: To evaluate the effectiveness of each ranking method in detecting the influencers, we use three standard evaluation measurements. The first measurement is F_1 , the harmonic mean of precision and recall,

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.11)$$

Even though F_1 is widely used, an important disadvantage is that it only examines a subset of top instances with highest scores, which might be the “easiest” cases. This phenomenon might lead to biased results when comparing the performance of different ranking methods. To overcome this problem, we also use AUC-ROC and AUC-PR, which measure the area under the Receiver-Operating-Characteristic (ROC) curve and the Precision-Recall (PR) curve. Using these two measurements, we can compare the performances of ranking methods using the full ranked lists. [Davis and Goadrich \(2006\)](#) point out that PR curve is more appropriate than ROC

for skewed datasets.

3.5.5 Results and Analysis

| | Ranking methods | F_1 | AUC-ROC | AUC-PR |
|------------------|-------------------------------|------------------|------------------|------------------|
| <i>Crossfire</i> | Num. of turns \diamond | .736 | .795 | .726 |
| | Total turn lengths \diamond | .716 | .782 | .730 |
| | Total topic shifts \star | .806 \pm .0122 | .858 \pm .0068 | .865 \pm .0063 |
| | Weighted topic shifts \star | .828 \pm .0100 | .869 \pm .0078 | .873 \pm .0057 |
| Wikipedia | Num. of turns \diamond | .367 | .730 | .291 |
| | Total turn lengths \diamond | .306 | .732 | .281 |
| | Total topic shifts \star | .552 \pm .0353 | .752 \pm .0144 | .377 \pm .0284 |
| | Weighted topic shifts \star | .488 \pm .0295 | .749 \pm .0149 | .379 \pm .0307 |

Table 3.8: Influencer detection results on *Crossfire* and Wikipedia discussion pages. For both datasets, topic-change-based methods (\star) outperform structure-based methods (\diamond) by large margins. For all evaluation measurements, higher is better.

Table 3.8 shows the results of the four ranking methods using *Crossfire* and Wikipedia discussion datasets. Since we run our Gibbs samplers multiple times, the results of the two topic-change-based methods are reported with standard deviations (across different chains).

For both datasets, the two topic-change-based methods outperform the two structure-based methods by a large margin for all three evaluation measurements. The standard deviations in all three measurements of the two topic-change-based methods are relatively small. This shows the effectiveness of features based on topic changes in detecting influencers in conversations. In addition, the weighted topic shifts ranking method generally performs better than the total topic shifts method. This provides strong evidence that SITS is capable of capturing the speakers’ propensity to change the topic. The improvement (if any) in the performance of

the weighted topic shifts ranking method over the total topic shifts method is more obvious in the *Crossfire* dataset than in Wikipedia discussions. We argue that this is because conversations in Wikipedia discussion pages are generally shorter and contain more speakers than those in *Crossfire* debates. This leaves less evidence about the topic change behavior of the speakers in Wikipedia and thus SITS struggles to capture the speakers' behavior.

3.6 Evaluating Topic Segmentation

In this section, we examine how well SITS identifies when new topics are introduced, i.e., how well it can segment conversations. We discuss metrics for evaluating an algorithm's segmentation relative to a gold annotation, describe our experimental setup, and report those results.

3.6.1 Experiment Setups

Evaluation Metrics: To evaluate the performance on topic segmentation, we use P_k (Beeferman et al., 1999) and WindowDiff (WD) (Pevzner and Hearst, 2002). Both metrics measure the probability that two points in a document will be incorrectly separated by a segment boundary. Both techniques consider all windows of size k in the document and count whether the two endpoints of the window are (im)properly segmented against the gold segmentation. More formally, given a reference segmentation \mathcal{R} and a hypothesized segmentation \mathcal{H} , the value of P_k for a

given window size k is defined as follow:

$$P_k = \frac{\sum_{i=1}^{N-k} \delta_{\mathcal{H}}(i, i+k) \oplus \delta_{\mathcal{R}}(i, i+k)}{N-k} \quad (3.12)$$

where $\delta_{\mathcal{X}}(i, j)$ is 1 if the segmentation \mathcal{X} assigns i and j to the same segment and 0 otherwise; \oplus denotes the XOR operator; N is the number of candidate boundaries.

WD improves P_k by considering how many boundaries lie between two points in the document, instead of just looking at whether the two points are separated or not. WD of size k between two segmentations \mathcal{H} and \mathcal{R} is defined as:

$$\text{WD} = \frac{\sum_{i=1}^{N-k} [|b_{\mathcal{H}}(i, i+k) - b_{\mathcal{R}}(i, i+k)| > 0]}{N-k} \quad (3.13)$$

where $b_{\mathcal{X}}(i, j)$ counts the number of boundaries that the segmentation \mathcal{X} puts between two points i and j .

However, these metrics have a major drawback. They require both hypothesized and reference segmentations to be binary. Many algorithms (e.g., probabilistic approaches) give *non-binary* segmentations where candidate boundaries have real-valued scores (e.g., probability or confidence). Thus, evaluation requires arbitrary thresholding to binarize soft scores. In previous work, to be fair for all methods, thresholds are usually set so that the number of segments is equal to a predefined value (Galley et al., 2003; Purver et al., 2006). In practice, this value is usually unknown.

To overcome these limitations, we also use $\widehat{\text{EMD}}$ (Pele and Werman, 2008),

| | Model | \widehat{EMD} | P_k | | | WindowDiff | | |
|--------------|--------------------|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | | | $k = 5$ | 10 | 15 | $k = 5$ | 10 | 15 |
| 2008 Debates | TextTiling | 2.821 | .433 | .548 | .633 | .534 | .674 | .760 |
| | P-NoSpeaker-single | 2.822 | .426 | .543 | .653 | .482 | .650 | .756 |
| | P-NoSpeaker-all | 2.712 | .411 | .522 | .589 | .479 | .644 | .745 |
| | P-SITS | 2.269 | .380 | .405 | .402 | .482 | .625 | .719 |
| | NP-HMM | 2.132 | .362 | .348 | .323 | .486 | .629 | .723 |
| | NP-SITS | 1.813 | .332 | .269 | .231 | .470 | .600 | .692 |
| ICSI | TextTiling | 2.507 | .289 | .388 | .451 | .318 | .477 | .561 |
| | P-NoSpeaker-single | 1.949 | .222 | .283 | .342 | .269 | .393 | .485 |
| | P-NoSpeaker-all | 1.935 | .207 | .279 | .335 | .253 | .371 | .468 |
| | P-SITS | 1.807 | .211 | .251 | .289 | .256 | .363 | .434 |
| | NP-HMM | 2.189 | .232 | .257 | .263 | .267 | .377 | .444 |
| | NP-SITS | 2.126 | .228 | .253 | .259 | .262 | .372 | .440 |

Table 3.9: Results on the topic segmentation task. Lower is better. The parameter k is the window size of the metrics P_k and WindowDiff chosen to replicate previous results.

a variant of the *Earth Mover’s Distance* (EMD). Originally proposed by [Rubner et al. \(2000\)](#), EMD is a metric that measures the distance between two normalized histograms. Intuitively, it measures the minimal cost that must be paid to transform one histogram into the other. EMD is a true metric only when the two histograms are normalized (e.g., two probability distributions). \widehat{EMD} relaxes this restriction to define a metric for non-normalized histograms by adding or subtracting masses so that both histograms are of equal size.

Applied to our segmentation problem, each segmentation can be considered a histogram where each candidate boundary point corresponds to a bin. The probability of each point being a boundary is the mass of the corresponding bin. We use $|i - j|$ as the ground distance between two points i and j .¹⁴ To compute \widehat{EMD} we use the FastEMD implementation ([Pele and Werman, 2009](#)).

¹⁴The ground distance is the distance between two bins in a histogram. Please refer to ([Pele and Werman, 2008](#)) for a more formal definition of \widehat{EMD} .

Experimental Methods: We applied the following methods to discover topic segmentations in a conversation:

- **TextTiling** (Hearst, 1997) is one of the earliest and most widely used general-purpose topic segmentation algorithms, sliding a fixed-width window to detect major changes in lexical similarity.
- **P-NoSpeaker-single**: parametric version of SITS without speaker identity, run individually on each conversation (Purver et al., 2006).
- **P-NoSpeaker-all**: parametric version of SITS without speaker identity run on all conversations.
- **P-SITS**: the parametric version of SITS with speaker identity run on all conversations.
- **NP-HMM**: the HMM-based nonparametric model with speaker identity. This model uses the same assumption as the Sticky HDP-HMM (Fox et al., 2008), where a single topic is associated with each turn.
- **NP-SITS**: the nonparametric version of SITS with speaker identity run on all conversations.

Parameter Settings and Implementation: In our experiment, all parameters of TextTiling are the same as in (Hearst, 1997). For statistical models, Gibbs sampling with 10 randomly initialized chains is used. Initial hyperparameter values are sampled from $U(0, 1)$ to favor sparsity; statistics are collected after 500 burn-in iterations with a lag of 25 iterations over a total of 5000 iterations; and slice sampling (Neal, 2003) optimizes hyperparameters. Parametric models are run with 25,

50 and 100 topics and the best results (averaged over 10 chains) are reported.

3.6.2 Results and Analysis

Table 3.9 shows the performance of various models on the topic segmentation problem, using the ICSI corpus and the 2008 election debates. Consistent with previous results in the literature, probabilistic models outperform TextTiling. In addition, among the probabilistic models, the models that had access to speaker information consistently segment better than those lacking such information. Furthermore, NP-SITS outperforms NP-HMM in both experiments, suggesting that using a distribution over topics for turns is better than using a single topic. This is consistent with the parametric models in (Purver et al., 2006).

The contribution of speaker identity seems more valuable in the debate setting. Debates are characterized by strong rewards for setting the agenda; dodging a question or moving the debate toward an opponent’s weakness can be useful strategies (Boydston et al., 2013a). In contrast, meetings (particularly low-stakes ICSI meetings, technical discussions in R&D group) tend to have pragmatic rather than strategic topic shifts. In addition, agenda-setting roles are clearer in formal debates; a moderator is tasked with setting the agenda and ensuring the conversation does not wander too much.

In the last three sections, we have reported the empirical evidences to show the effectiveness of SITS in (1) capturing agenda control behavior of participants, (2) detecting influencers, and (3) performing topic segmentation in conversations.

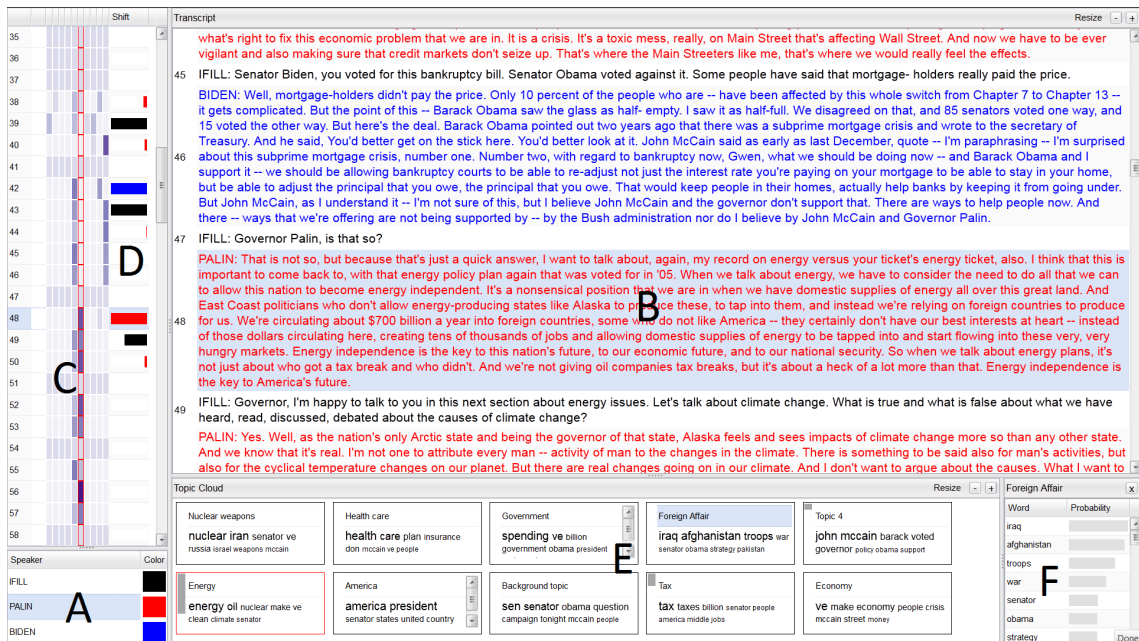


Figure 3.7: The *Argviz* user interface consists of *speaker panel* (A), *transcript panel* (B), *heatmap* (C), *topic shift column* (D), *topic cloud panel* (E), *selected topic panel* (F).

However, the latent structures that SITS extracts from conversational data are much richer. To help analyst leverage SITS's outputs to analyze conversations more effectively, we build an interactive visualization called *Argviz*, which we describe in detail in the next section.

3.7 *Argviz*: Interactive Visualization of Topic Dynamics in Conversations

Uncovering the structure of conversations often requires close reading by a human expert to be effective. As motivated at the beginning of this chapter, political scientists often use manual content analysis to analyze what gets said in debate to explore how candidates shape the debate's agendas and frame issues (Boydston

et al., 2013a,b), or how answers subtly (or not so subtly) shift the conversation by dodging the question that was asked (Rogers and Norton, 2011). In previous sections, we have introduced SITS, a Bayesian nonparametric model which can automatically discover the topics discussed in a conversation and when these topics change. In this section, we introduce *Argviz*, an interactive visualization which can leverage SITS’s outputs to help domain expert analyze the topical dynamics of multi-party conversations. *Argviz*’s interface allows users to quickly grasp the topical flow of the conversation, discern when the topic changes and by whom, and interactively visualize the conversation’s details on demand.

Argviz is a web-based application, built using Google Web Toolkit (GWT),¹⁵ which allows users to visualize and manipulate SITS’s outputs entirely in their browser after a single server request. Given the limited screen of a web browser, *Argviz* follows the multiple coordinated views approach (Wang Baldonado et al., 2000; North and Shneiderman, 2000) successfully used in Spotfire (Ahlberg, 1996), Improvise (Weaver, 2004), and SocialAction (Perer and Shneiderman, 2006). *Argviz* supports three main coordinated views: TRANSCRIPT, OVERVIEW and TOPIC.

- TRANSCRIPT occupies the prime real estate for a close reading. It has a *transcript panel* and a *speaker panel*. The *transcript panel* displays the original transcript. Each conversational turn is numbered and color-coded by speaker. The color associated with each speaker can be customized using the *speaker panel*, which lists all the speakers.
- OVERVIEW shows how topics gain and lose prominence during the conversa-

¹⁵<https://developers.google.com/web-toolkit/>

tion. SITS’s outputs include a topic distribution and a topic shift probability for each turn in the conversation. In *Argviz* these are represented using a *heatmap* and *topic shift column*.

In the *heatmap*, each turn-specific topic distribution is displayed by a heatmap row (Sopan et al., 2013). There is a cell for each topic, and the color intensity of each cell is proportional to the probability of the corresponding topic of a particular turn. Thus, users can see the topical flow of the conversation through the vertical change in cells’ color intensities as the conversation progresses. In addition, the *topic shift column* shows the topic shift probability (inferred by SITS) using color-coded bar charts, helping users discern large topic changes in the conversation. Each row is associated with a turn in the conversation; clicking on one shifts the TRANSCRIPT view.

- TOPIC displays the set of topics learned by SITS, with font-size proportional to the words’ topic probabilities. The *selected topic panel* goes into more detail, with bar charts showing the topic-word distribution. For example, in Figure 3.7, the Foreign Affairs topic in panel E has high probability words “iraq”, “afghanistan”, “war”, etc. in panel F.

Figure 3.7 shows *Argviz* displaying the 2008 vice presidential debate between Joe Biden and Sarah Palin, moderated by Gwen Ifill. Users can start exploring the interface from any of the views described above to gain insight about the conversation. For example, an user may be interested in seeing how the “Economy” is discussed in the debates. Clicking on a topic in the *topic cloud panel* highlights

that column in the *heatmap*. The user can now see where the “Economy” topic is discussed in the debate. Next to the heatmap, the *topic shift column* when debate participants changed the topic. The red bar in turn 48 shows an interaction where Governor Palin dodged a question on the “bankruptcy bill” to discuss her “record on energy”. Clicking on this turn shows the interaction in the TRANSCRIPT view, allowing a closer reading.

Users might also want to contrast the topics that were discussed before and after the shift. This can be easily done with the coordination between the *heatmap* and the *topic cloud panel*. Clicking on a cell in the *heatmap* will select the corresponding topic to display in the *selected topic panel*. In our example, the topic of the conversation was shifted from “Economy” to “Energy” at turn 48.

3.8 Conclusions and Future Work

In this chapter, we focus on analyzing agendas and agenda control behaviors in political debates and other conversations. We introduce SITS, a Bayesian nonparametric topic model that jointly captures topics, topic shifts and individuals’ tendency to control the topic in conversations. Using SITS, we analyze the agenda control behaviors of candidates in the 2008 U.S. election debates and the 2012 Republican primary debates. We also apply SITS on a large-scale set of debate transcripts from CNN’s TV show *Crossfire*. To make the analysis process more effective, we build *Argviz*, an interactive visualization which leverages SITS’s outputs to allow users to quickly grasp the topical dynamics of the conversation, discover

when the topic changes and by whom, and interactively visualize the conversation’s details on demand. In addition to providing insights on agendas and agenda control in multi-party conversation, through extensive empirical experiments, we also show that SITS can effectively improve the performance of two quantitative tasks: influencer detection and topic segmentation.

Crucially, SITS models speaker-specific properties. As such, it improves performance on practical tasks such as unsupervised segmentation, but it also is attractive philosophically. Accurately modeling *individuals* is part of a broader research agenda that seeks to understand individuals’ values (Fleischmann et al., 2011), interpersonal relationships (Chang et al., 2009a), and perspective (Hardisty et al., 2010), which creates a better understanding of what people think based on what they write or say (Pang and Lee, 2008). One particularly interesting direction is to extend the model to capture how language is coordinated during the conversation and how it correlates with influence (Giles et al., 1991; Danescu-Niculescu-Mizil et al., 2012).

The problem of finding influencers in conversation has been studied for decades by researchers in communication, sociology, and psychology, who have long acknowledged qualitatively the correlation between the ability of a participant to control conversational topic and his or her influence on other participants during the conversation. With SITS, we now introduce a computational technique for modeling more formally who is controlling the conversation. Empirical results on two datasets (*Crossfire* TV show and Wikipedia discussion pages) show that methods based on SITS outperform previous methods that used conversational structure patterns in detecting influencers.

Chapter 4: Learning Agenda Hierarchy from Multi-labeled Political Text

4.1 Introduction

In the previous chapter, we focus our study on agendas in conversations, where multiple individuals take turn to discuss various topics. In this chapter, we move to a more general setting to study agendas in political text, using legislative text in the U.S. Congress. The U.S. Congress is the bicameral legislature of the U.S. federal government which consists of the Senate and the House of Representatives. The most important responsibility of the Congress is making the laws, which are proposed in the *congressional bills*. Discovering and analyzing agendas in congressional bills help shed light on understanding the political attentions of policymakers and answer important questions such as: What are the most important policy issues in the 112th Congress? How does the attention on those issues change over time?

We study policy agendas in congressional bill text using *multi-labeled data*, in which each bill is tagged with multiple agenda issues (i.e., labels) from a flexible list. Using this type of multi-labeled data provides two clear advantages: (1) it helps improve the interpretability of the learned topics, compared with using unlabeled

data, which reduces post-analysis cost, and (2) it reduces pre-analysis cost in comparison with single-labeled data using a predefined coding system. However, one major drawback of this type of multi-labeled data is that the label space is often much larger than that of single-labeled data, which makes learning and predicting relatively harder.

In this chapter, we present L2H—*Label to Hierarchy*—a hierarchical topic model that can induce a hierarchy of user-generated labels and the topics associated with those labels from a set of multi-labeled documents. The model is robust enough to account for missing labels from untrained, disparate annotators and provide an interpretable summary of an otherwise unwieldy label set. We apply L2H to study policy agendas and the relationships among different agenda issues using bill text from four Congresses (109th–112th). Empirical experiments shows the effectiveness of L2H in predicting held-out words and labels of unseen documents.

This chapter synthesizes and revises the work originally published in (Nguyen et al., 2013b, 2014c).

4.1.1 Analyzing Agendas in Legislative Text

As discussed in Chapter 1, the study of agendas in political text has been the focus of a large body of research in political science for decades. However, for a long period of time, agenda-setting research had largely been case studies: researchers analyzed agendas in various separated cases of interest, but there was no systematic way to compare and measure the differences in policy activities across

different settings (Baumgartner et al., 2006). In 1993, Bryan Jones and Frank Baumgartner initiated the Policy Agendas Project “in response to a clear need for better measurement of key concepts in the study of public policy”.¹ The project develops a coding scheme of 19 major topic and 225 subtopic codes, which provides an *exhaustive and consistent* set of references in analyzing agendas in political text. The codebook has been used extensively in numerous research and the project has become the model for various research programs on studying policy agendas (John, 2006).² Table 4.1 shows the list of the major topics in the 2014 Policy Agendas Topics codebook.

Using the Policy Agendas Topics codebook, the Congressional Bills project, led by Adler and Wilkerson (2006), labels over 400,000 public and private bills introduced in the U.S. Congress since 1947. Each bill is coded with one major topic and one subtopic, based on its title of the introduced version. Table 4.2 shows some examples of coded bills from the 100th Congress.

This set of coded congressional bills has provided an invaluable resource for scholars studying legislative politics (Adler and Wilkerson, 2013). However, using the Policy Agendas Topics codebook to analyze agendas and assign a single label to each bill also poses some drawbacks. In this chapter, we use the set of labeled data provided by the Congressional Research Service (CRS), in which each bill is coded with *multiple agenda issues*. In this set of data, each issue comes from a

¹See <http://www.policyagendas.org/page/about-project> for a description of the history and development of the project

²Examples of research programs modeled after the Policy Agendas Project include the *European Union Policy Agendas Project* (<http://www.policyagendas.eu/>), the *Comparative Agendas Project* (<http://www.comparativeagendas.info/>), and the *Pennsylvania Policy Database Project* (<http://www.cla.temple.edu/papolicy/>)

| Code | Major Topic |
|------|--|
| 1 | Macroeconomics |
| 2 | Civil Rights, Minority Issues, and Civil Liberties |
| 3 | Health |
| 4 | Agriculture |
| 5 | Labor and Employment |
| 6 | Education |
| 7 | Environment |
| 8 | Energy |
| 9 | Immigration [†] |
| 10 | Transportation |
| 12 | Law, Crime, and Family Issues |
| 13 | Social Welfare |
| 14 | Community Development and Housing Issues |
| 15 | Banking, Finance, and Domestic Commerce |
| 16 | Defense |
| 17 | Space, Science, Technology and Communication |
| 18 | Foreign Trade |
| 19 | International Affairs and Foreign Aid |
| 20 | Government Operations |
| 21 | Public Lands and Water Management |

Table 4.1: Major topics with their corresponding codes in the Policy Agendas Topics codebook. [†]The major topic “Immigration” was newly added to the codebook in 2014.

| Bill | Title | PA Topics | |
|---------|---|-----------|-------|
| | | Major | Minor |
| H.R.228 | A bill to amend title II of the Social Security Act to provide that an individuals entitlement to benefits thereunder shall continue through the month of his or her death (without affecting any other persons entitlement to benefits for that month), in order to provide such individuals family with assistance in meeting the extra death-related expenses. | 13 | 1303 |
| H.R.62 | A bill to establish a series of six regional Presidential primaries at which the public may express its preference for the nomination of an individual for election to the Office of President of the United States. | 20 | 2012 |
| H.R.364 | A bill to amend the Internal Revenue Code of 1954 to allow individuals a deduction for commuting expenses incurred on public mass transit. | 10 | 1001 |

Table 4.2: Examples of bills from the 100th Congress, coded by the Congressional Bills project. The mapping of the Policy Agenda (PA) major topic codes are provided in Table 4.1.

| Bill | CRS Index Terms |
|---------|--|
| | Social welfare |
| H.R.228 | Old age, survivors and disability insurance Social security eligibility |
| | Government operations and politics Campaign funds |
| H.R.62 | Presidential candidates Presidential elections Presidential primaries Voting |
| | Taxation Commuting Income tax Mass rapid transit |
| H.R.364 | Motor Vehicles and Driving Tax deductions Transportation and Travel Travel costs Urban affairs Urban transportation |

Table 4.3: Examples of multiple labels provided by the Congressional Research Service for the three bills shown in Table 4.2

relatively long list of *CRS Legislative Subject Terms*.³ Here are the advantages of using multi-labeled data over the single-labeled data coded using the Policy Agendas

Topic codebook:

- A Congressional bill can be about more than one agenda issue. Indeed, in the description about its coded data, the Congressional Bills project notes that “researchers should not assume that every bill relating to ‘air pollution’ (for example) will be found among the ‘705’ bills. A bill could address air pollution but be primarily among something else”, and explicitly refers to the Library of Congress, where we obtained the multi-labeled data, for coded data with more details.

³See <http://thomas.loc.gov/help/terms-subjects.html> for more details about the list of CRS subject terms.

- Compared to using a fixed and exhaustive coding system like the Policy Agendas Topics Codebook, it is relatively cheaper to maintain and use the list of labels such as the one provided by CRS. This list of labels needs not to be completely and exhaustively defined beforehand by domain experts and can be cumulatively extended as new labels arise. This flexibility is particularly useful given the fact that, despite being very well defined, the Policy Agendas Topic codebook has also been modified, both globally (e.g., adding major topic “Immigration” in 2014) or specifically for certain datasets (e.g., adding eight major topics for coding the New York Times and the Encyclopedia of Associations) or domains (e.g., adding major topic “EU Governance” to study agendas in the European Union).

These advantages help reduce the high pre-analysis cost of approaches using traditionally coded data. However, they do not come for free. The large number of labels poses new challenges to learning and interpreting the data. First, with the large label space, there are naturally dependencies among the labels, which any effective models should capture. For example, the model should be able to capture the fact that “Taxation” is, in general, more similar to “Income Tax” than to “Social Welfare”. In addition, the coded data are likely to be noisy and not exhaustive. For example, one might expect “Public Transit”—a valid CRS subject term—to appear in the list of labels for Bill H.R. 364 shown in Tables 4.2 and 4.3. In this chapter, the model we introduce, L2H, tackles these problems by learning a tree-structured hierarchy of labels, which provides interpretable results, and thus helps avoid a high

post-analysis cost.

4.1.2 Topic Models for Multi-labeled Documents

Even though we are motivated by analyzing agendas in political text, especially legislative text, that are tagged with multiple agenda issues, multi-labeled data are ubiquitous and exist in a wide range of settings and applications. Web pages are tagged with multiple directories,⁴ Ph.D. theses are associated with multiple key words,⁵ interdisciplinary research grants are assigned multiple research areas, and books can be labeled with more than one category. Topic models for multi-labeled documents in general fall into a branch of topic modeling research, which jointly captures the documents' text and their associated metadata. Chapter 2 surveys a collection of topic models using different types of metadata.

Specifically having multiple labels as the metadata, various topic models have been proposed. [Rosen-Zvi et al. \(2004\)](#) introduce the Author-Topic model which jointly captures the documents' text and their authorship information. [Ramage et al. \(2009\)](#) propose Labeled LDA whose generative process is similar to LDA except that tokens in a given document are generated only from a subset of topics defined by the document's set of labels. Various researchers have applied Labeled LDA to different settings including characterizing Twitter users' contents ([Ramage et al., 2010a](#)) and performing named entity recognition ([Ritter et al., 2011](#)) in microblogs. [Wang et al. \(2009\)](#) extend sLDA to jointly capture both the class label and multiple

⁴Open Directory Project (<http://www.dmoz.org/>)

⁵ProQuest service (<http://www.proquest.com/>)

annotated key words in a set of images. [Yang et al. \(2009\)](#) use logistic regression to indicate whether a label appears in a document.

These models have been applied effectively to a wide range of multi-labeled data. However, one common drawback of these models is that they assume that all labels are independent. This assumption is reasonable if the number of labels is relatively small. Unfortunately, in many real-world examples, it is not uncommon to have hundreds or thousands of unique labels, which makes the independence assumption too strong. When the label space is large, there are naturally relationships among the labels. This characteristic of large-scale multi-labeled data motivates various work to learn the label dependencies. For example, [Ramage et al. \(2011\)](#) propose Partially Labeled LDA (PLDA) and Partially Labeled Dirichlet Process (PLDP) to discover the label relationships in a hidden topic space. Also projecting labels onto some latent space, [Rubin et al. \(2012\)](#) introduce Prior-LDA and Dependency-LDA to capture the label dependencies and overcome the aforementioned problem.

Projecting labels onto a latent space has proven to be an effective approach to learn the label dependencies, especially when the main goal is to achieve good performances on predicting labels for unseen data ([Rubin et al., 2012](#)). However, by moving to a latent space, these models risk the interpretability that the labels provide, which motivates us to use labeled data to analyze agendas and reduce the post-analysis cost in the first place (Chapter 1). This motivates the model we present in this chapter, L2H, which captures *explicitly* label dependencies using a tree-structured hierarchy. Besides analyzing agendas in multi-labeled political text

in this thesis, interpretable label hierarchies in general have been shown to be a valuable resource which can provide an effective way of organizing data including text (Lewis et al., 2004) and images (Deng et al., 2009), and make annotating large-scale multi-labeled data more effective and scalable (Deng et al., 2014; Lin et al., 2014).

4.1.3 Chapter Structure

We describe the L2H model in detail next in Section 4.2. In Section 4.3, we present an MCMC algorithm to perform posterior inference efficiently. To evaluate L2H, we use multi-labeled legislative text in four Congresses (109th–112th). Section 4.4 qualitatively analyzes the hierarchy learned from the data by L2H. We also quantitatively evaluate L2H in two computational tasks: document modeling—measuring perplexity on a held-out set of document, and multi-label classification—predicting multiple labels for unseen documents in Section 4.5. Section 4.6 concludes the chapter and opens up some potential future directions.

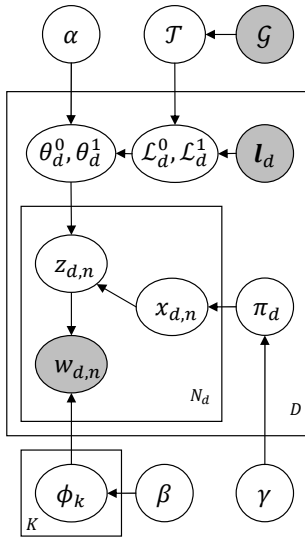
4.2 L2H: Capturing Label Dependencies using a Tree-structured Hierarchy

Discovering a topic hierarchy from text has been the focus of much research in the topic modeling community. One popular approach is to learn an *unsupervised* hierarchy of topics, which we survey in Chapter 2. As discussed in Chapter 1, unsupervised topic hierarchies provide a useful way to perform exploratory analysis, but

it usually requires an additional step of post hoc topic labeling to make them interpretable. This difficulty motivates work leveraging existing label taxonomies such as HSLDA (Perotte et al., 2011), hLLDA (Petinot et al., 2011) and SSDLDA (Mao et al., 2012a).

A second active area of research is constructing a taxonomy from multi-labeled data. For example, Heymann and Garcia-Molina (Heymann and Garcia-Molina, 2006) extract a tag hierarchy using the tag network centrality; similar work has been applied to protein hierarchies (Tibely et al., 2013). Hierarchies of concepts have come from seeded ontologies (Schmitz, 2006), crowdsourcing (Nikolova et al., 2012; Chilton et al., 2013; Bragg et al., 2013), and user-specified relations (Plangprasopchok and Lerman, 2009). More sophisticated approaches build domain-specific keyword taxonomies with adapting Bayesian Rose Trees (Liu et al., 2012). All of these approaches, however, concentrate on the tags, ignoring the *content* the tags describe.

In this chapter, we combine ideas from these two lines of research and introduce L2H, a hierarchical topic model that discovers a tree-structured hierarchy of concepts from a collection of multi-labeled documents. L2H takes as input a set of D documents $\{\mathbf{w}_d\}$, each tagged with a set of labels \mathbf{l}_d . The label set \mathcal{L} contains K unique, unstructured labels and the word vocabulary size is V . To learn an interpretable taxonomy, L2H associates each *label*—a user-generated word/phrase—with a *topic*—a multinomial distribution over the vocabulary—to form a *concept*, and infers a tree-structured hierarchy to capture the relationships among concepts. Figure 4.1 shows the plate diagram for L2H, together with its generative process.



1. Create label graph \mathcal{G} (Section 4.2.1)
2. Draw a spanning tree \mathcal{T} from \mathcal{G} (Section 4.2.2)
3. For each node $k \in [1, K]$ in \mathcal{T}
 - (a) If k is the root, draw background topic $\phi_k \sim \text{Dir}(\beta \mathbf{u})$
 - (b) Otherwise, draw topic $\phi_k \sim \text{Dir}(\beta \phi_{\sigma(k)})$ where $\sigma(k)$ is node k 's parent.
4. For each document $d \in [1, D]$ having labels \mathbf{l}_d
 - (a) Define \mathcal{L}_d^0 and \mathcal{L}_d^1 using \mathcal{T} and \mathbf{l}_d (Section 4.2.3)
 - (b) Draw $\theta_d^0 \sim \text{Dir}(\mathcal{L}_d^0 \times \alpha)$ and $\theta_d^1 \sim \text{Dir}(\mathcal{L}_d^1 \times \alpha)$
 - (c) Draw a switching variable $\pi_d \sim \text{Beta}(\gamma_0, \gamma_1)$
 - (d) For each token $n \in [1, N_d]$
 - i. Draw set indicator $x_{d,n} \sim \text{Bern}(\pi_d)$
 - ii. Draw topic indicator $z_{d,n} \sim \text{Mult}(\theta_d^{x_{d,n}})$
 - iii. Draw word $w_{d,n} \sim \text{Mult}(\phi_{z_{d,n}})$

Figure 4.1: Generative process and the plate diagram notation of L2H.

4.2.1 Creating the Label Graph

We assume an underlying directed graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ in which each node is a concept consisting of (1) a label—observable user-generated input, and (2) a topic—latent multinomial distribution over words.⁶ The prior weight of a directed edge from node i to node k is the fraction of documents tagged with label k which are also tagged with label i : $t_{i,k} = D_{i,k}/D_j$. We also assume an additional root node which is called the BACKGROUND node. Edges to the BACKGROUND node have prior zero weight and edges from the BACKGROUND node to node i have prior weight $t_{\text{root},i} = D_i/\max_k D_k$. Here, D_i is the number of documents tagged with label i , and $D_{i,k}$ is the number of documents tagged with both labels i and k .

Figure 4.2a illustrates the weighted directed graph \mathcal{G} , created from a simple multi-labeled data set having three unique labels: HEALTH CARE, HEALTH CARE

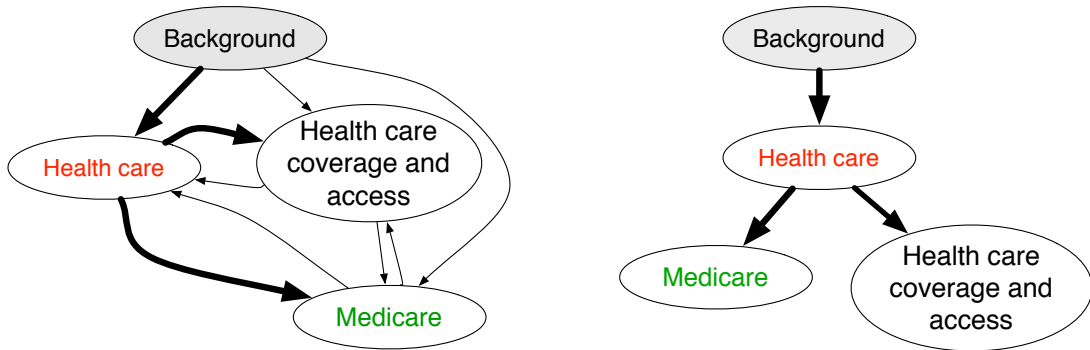
⁶We use *node* when emphasizing the structure discovered by the model. Each node corresponds to a *concept* which consists of a *label/topic* pair.

COVERAGE AND ACCESS, and MEDICARE. In this example, HEALTH CARE is more general than MEDICARE, and thus the weight of the directed edge from HEALTH CARE to MEDICARE is greater than that of the reciprocal edge.

4.2.2 Generating Tree-structured Hierarchy

The tree \mathcal{T} is a spanning tree generated from \mathcal{G} . The probability of a tree given the graph \mathcal{G} is thus the product of all its edge prior weights $p(\mathcal{T} | \mathcal{G}) = \prod_{e \in \mathcal{E}} t_e$. To capture the intuition that child nodes in the hierarchy specialize the concepts of their parents, we model the topic ϕ_k at each node k using a Dirichlet distribution whose mean is centered at the topic of node k 's parent $\sigma(k)$, i.e., $\phi_k \sim \text{Dir}(\beta\phi_{\sigma(k)})$. The topic at the root node is drawn from a symmetric Dirichlet $\phi_{\text{root}} \sim \text{Dir}(\beta\mathbf{u})$, where \mathbf{u} is a uniform distribution over the vocabulary (Adams et al., 2010; Ahmed et al., 2013a). This is similar to the idea of “back-off” in language models where more specific contexts inherit the ideas expressed in more general contexts; i.e., if we talk about “pedagogy” in EDUCATION, there’s a high likelihood we’ll also talk about it in UNIVERSITY EDUCATION (Mackay and Peto, 1995; Teh, 2006).

Figure 4.2b illustrates an example of a spanning tree that can be generated from the label graph shown in Figure 4.2a. This spanning tree is also the maximum spanning tree that can be generated from \mathcal{G} , based on the example edge weights illustrated, which we will use to initialize the hierarchy during posterior inference (Section 4.3).



(a) The weighted directed graph \mathcal{G} (b) A spanning tree \mathcal{T} generated from \mathcal{G}

Figure 4.2: Example of the weighted directed graph \mathcal{G} and a spanning tree \mathcal{T} generated from \mathcal{G} , created from a set of multi-labeled data having three unique labels: HEALTH CARE, HEALTH CARE COVERAGE AND ACCESS, and MEDICARE. The thickness of an directed edge represents its weight.

4.2.3 Generating Documents

As in LDA, each word in a document is generated by one of the latent topics. L2H, however, also uses the labels and topic hierarchy to restrict the topics a document uses. The document’s label set \mathbf{l}_d identifies which nodes are more likely to be used. *Restricting* tokens of a document in this way—to be generated only from a subset of the topics depending the document’s labels—creates specific, focused, labeled topics (Ramage et al., 2009, Labeled LDA).

Unfortunately, \mathbf{l}_d is unlikely to be an exhaustive enumeration: particularly when the label set is large, users often forget or overlook relevant labels. We therefore depend on the learned topology of the hierarchy to *fill in* the gaps of what users forget by expanding \mathbf{l}_d into a broader set, \mathcal{L}_d^1 , which is the union of nodes on the paths from the root node to any of the document’s label nodes. We call this the document’s *candidate set*. The candidate set also induces a *complementary set*

$\mathcal{L}_d^0 \equiv \mathcal{L} \setminus \mathcal{L}_d^1$. Previous approaches such as LPAM (Bakalov et al., 2012) and Tree labeled LDA (Slutsky et al., 2013b) also leverage the label hierarchy to expand the original label set. However, these previous models require that the label hierarchy is given rather than inferred as in our L2H.

Figure 4.3 illustrated the candidate set and the complementary set of a document tagged with two labels: HIGHER EDUCATION and MEDICARE. The candidate set consists of nodes on all the paths from the BACKGROUND root node to any of the document’s label nodes. This allows an imperfect label set to include topics that the document *should* be associated with even if they were not explicitly enumerated.

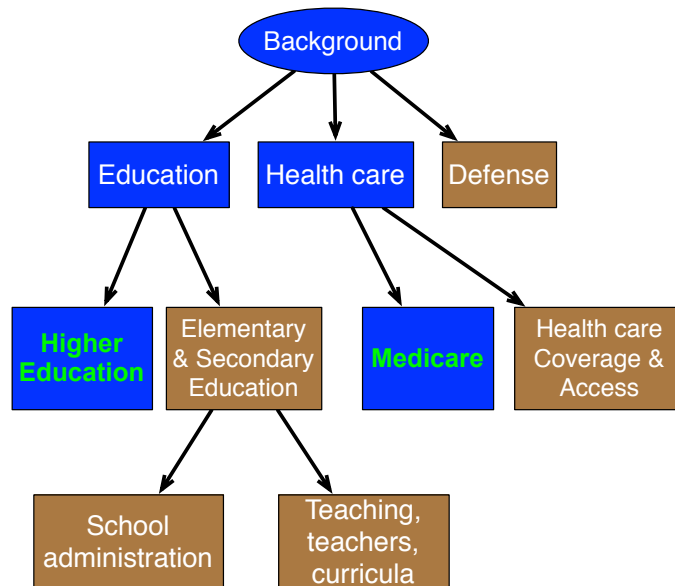
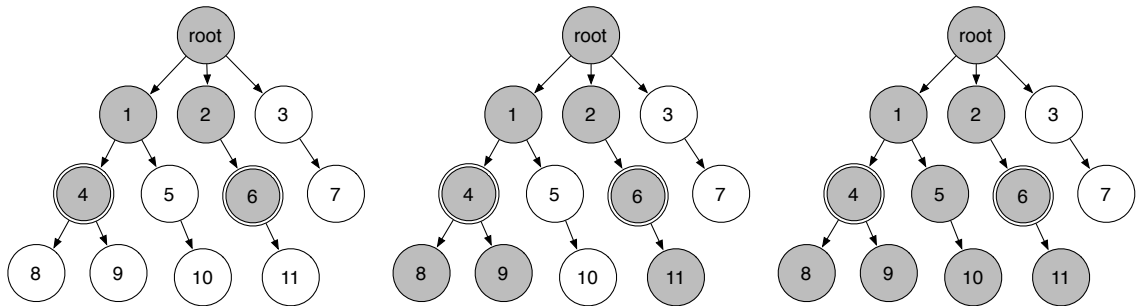


Figure 4.3: Illustration of the *candidate set* and the *complementary set* of a document tagged with two labels: HIGHER EDUCATION and MEDICARE.

L2H replaces Labeled LDA’s absolute *restriction* to specific topics to a soft *preference*. To achieve this, each document d has a *switching variable* π_d drawn from $\text{Beta}(\gamma_0, \gamma_1)$, which effectively decides how likely tokens in d are to be generated from \mathcal{L}_d^1 versus \mathcal{L}_d^0 . Token n in document d is generated by first flipping the

biased coin π_d to choose the set indicator $x_{d,n}$. Given $x_{d,n}$, the label $z_{d,n}$ is drawn from the corresponding label distribution $\theta_d^{x_{d,n}}$ and the token is generated from the corresponding topic $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$.

In [Nguyen et al. \(2013b\)](#), we also explored different ways to define the candidate set of a document given its set of labels (Figure 4.4). Preliminary results show that using these candidate sets yields relatively similar predictive results. However, understanding the effects of different candidate sets in more detail is an important research question that requires further investigation in future work. Moreover, to keep the posterior inference tractable (Section 4.3), we use a relatively straightforward prior distribution over the nodes in the candidate/complementary sets. Incorporating more sophisticated prior distributions which take into account of the tree-structured hierarchy is another interesting research direction.



(a) The candidate set $\mathcal{L}^{1,a}$ contains all nodes on the paths from the root to any of the document's label nodes (explored in detail in this chapter). (b) The candidate set $\mathcal{L}^{1,b}$ contains nodes in $\mathcal{L}^{1,a}$ and all the nodes in the subtree rooted at any of the document's label nodes. (c) The candidate set $\mathcal{L}^{1,c}$ contains nodes in $\mathcal{L}^{1,b}$ and all the nodes in the subtree rooted at the first-level nodes included in $\mathcal{L}^{1,a}$ (Fig. 4.4a).

Figure 4.4: Example of different ways to define the candidate set \mathcal{L}_1 (shaded nodes) and the complementary set \mathcal{L}_0 (white nodes) for a document tagged with two labels (double-circled nodes).

4.3 Posterior Inference

Given a set of documents with observed words $\{\mathbf{w}_d\}$ and labels $\{\mathcal{L}_d\}$, we develop an MCMC algorithm to infer the posterior distribution over the latent variables. Each iteration of our algorithm, after the initialization, consists of the following steps: (1) sample a set indicator $x_{d,n}$ and topic assignment $z_{d,n}$ for each token, (2) sample a word distribution ϕ_k for each node k in the tree, and (3) update the structure of the label tree.

4.3.1 Initialization

With the large number of labels, the space of hierarchical structures that MCMC needs to explore is huge. Initializing the tree-structure hierarchy is crucial to help the sampler focus on more important regions of the search space and help the sampler converge more quickly. We initialize the hierarchy with the maximum *a priori* probability tree by running Chu-Liu/Edmonds' algorithm to find the *maximum spanning tree* on the graph \mathcal{G} starting at the BACKGROUND node (Chu and Liu, 1965; Edmonds, 1967).

4.3.2 Sampling Assignments $x_{d,n}$ and $z_{d,n}$

For each token, we need to sample whether it was generated from the label set or not, $x_{d,n}$. We choose label set i with probability $\frac{C_{d,i}^{-d,n} + \gamma_i}{C_{d,\cdot}^{-d,n} + \gamma_0 + \gamma_1}$ and we sample a node in the chosen set i with probability $\frac{N_{d,k}^{-d,n} + \alpha}{C_{d,i}^{-d,n} + \alpha|\mathcal{L}_d^i|} \cdot \phi_{k,w_{d,n}}$. Here, $C_{d,i}$ is the number of times tokens in document d are assigned to label set i ; $N_{d,k}$ is the number

of times tokens in document d are assigned to node k . Marginal counts are denoted by \cdot , and $^{-d,n}$ denotes the counts excluding the assignment of token $w_{d,n}$.

After we have the label set, we can sample the topic assignment. This is more efficient than sampling jointly, as most tokens are in the label set, and there are a limited number of topics in the label set. The probability of assigning node k to $z_{d,n}$ is $p(x_{d,n} = i, z_{d,n} = k \mid \mathbf{x}^{-d,n}, \mathbf{z}^{-d,n}, \phi, \mathcal{L}_d^i)$

$$\propto \frac{C_{d,i}^{-d,n} + \gamma_i}{C_{d,\cdot}^{-d,n} + \gamma_0 + \gamma_1} \cdot \frac{N_{d,k}^{-d,n} + \alpha}{C_{d,i}^{-d,n} + \alpha |\mathcal{L}_d^i|} \cdot \phi_{k,w_{d,n}} \quad (4.1)$$

4.3.3 Sampling Topics ϕ

As discussed in Section 4.2.2, topics on each path in the hierarchy form a cascaded Dirichlet-multinomial chain where the multinomial ϕ_k at node k is drawn from a Dirichlet distribution with the mean vector being the topic $\phi_{\sigma(k)}$ at the parent node $\sigma(k)$. Given assignments of tokens to nodes, we need to determine the conditional probability of a word given the token. This can be done efficiently in two steps: bottom-up smoothing and top-down sampling (Ahmed et al., 2013a).

- Bottom-up smoothing: This step estimates the counts $\tilde{M}_{k,v}$ of node k propagated from its children. This can be approximated efficiently by using the minimal/maximal path assumption (Cowans, 2006; Wallach, 2008). For the minimal path assumption, each child node k' of k propagates a value of 1 to $\tilde{M}_{k,v}$ if $M_{k',v} > 0$. For the maximal path assumption, each child node k' of k propagates the full count $M_{k',v}$ to $\tilde{M}_{k,v}$ (Chapter 2).

- Top-down sampling: After estimating $\tilde{M}_{k,v}$ for each node from leaf to root, we sample the word distributions top-down using its actual counts \mathbf{m}_k , its children’s propagated counts $\tilde{\mathbf{m}}_k$ and its parent’s word distribution $\phi_{\sigma(k)}$ as $\phi_k \sim \text{Dirichlet}(\mathbf{m}_k + \tilde{\mathbf{m}}_k + \beta\phi_{\sigma(k)})$.

4.3.4 Updating tree structure \mathcal{T}

We update the tree structure by looping through each non-root node, proposing a new parent node and either accepting or rejecting the proposed parent using the Metropolis-Hastings algorithm. More specifically, given a non-root node k with current parent i , we propose a new parent node j by sampling from all incoming nodes of k in graph \mathcal{G} , with probability proportional to the corresponding edge weights. If the proposed parent node j is a descendant of k , we reject the proposal to avoid a cycle. If it is not a descendant, we accept the proposed move with probability $\min\left(1, \frac{Q(i \prec k) P(j \prec k)}{Q(j \prec k) P(i \prec k)}\right)$, where Q and P denote the proposal distribution and the model’s joint distribution respectively, and $i \prec k$ denotes the case where i is the parent of k .

Since we sample the proposed parent using the edge weights, the proposal probability ratio is

$$\frac{Q(i \prec k)}{Q(j \prec k)} = \frac{t_{i,k}}{t_{j,k}} \tag{4.2}$$

The joint probability of L2H’s observed and latent variables is:

$$P = \prod_{e \in \mathcal{E}} p(e | \mathcal{G}) \prod_{d=1}^D p(\mathbf{x}_d | \gamma) p(\mathbf{z}_d | \mathbf{x}_d, \mathbf{l}_d, \alpha) p(\mathbf{w}_d | \mathbf{z}_d, \phi) \prod_{l=1}^K p(\phi_l | \phi_{\sigma(l)}, \beta) p(\phi_{\text{root}} | \beta) \quad (4.3)$$

When node k changes its parent from node i to j , the candidate set \mathcal{L}_d^1 changes for any document d that is tagged with any label in the subtree rooted at k . Let Δ_k denote the subtree rooted at k and $\mathcal{D}_{\Delta_k} = \{d | \exists l \in \Delta_k \wedge l \in \mathbf{l}_d\}$ the set of documents whose candidate set might change when k ’s parent changes. Canceling unchanged quantities, the ratio of the joint probabilities is:

$$\frac{P(j \prec k)}{P(i \prec k)} = \frac{t_{j,k}}{t_{i,k}} \prod_{d \in \mathcal{D}_{\Delta_k}} \frac{p(\mathbf{z}_d | j \prec k) p(\mathbf{x}_d | j \prec k) p(\mathbf{w}_d | j \prec k)}{p(\mathbf{z}_d | i \prec k) p(\mathbf{x}_d | i \prec k) p(\mathbf{w}_d | i \prec k)} \prod_{l=1}^K \frac{p(\phi_l | j \prec k)}{p(\phi_l | i \prec k)} \quad (4.4)$$

We now expand each factor in Equation 4.4. The probability of node assignments \mathbf{z}_d for document d is computed by integrating out the document-topic multinomials θ_d^0 and θ_d^1 (for the candidate set and its inverse):

$$p(\mathbf{z}_d | \mathbf{x}_d, \mathcal{L}_d^0, \mathcal{L}_d^1, \alpha) = \prod_{x \in \{0,1\}} \frac{\Gamma(\alpha |\mathcal{L}_d^x|)}{\Gamma(C_{d,x} + \alpha |\mathcal{L}_d^x|)} \prod_{l \in \mathcal{L}_d^x} \frac{\Gamma(N_{d,l} + \alpha)}{\Gamma(\alpha)} \quad (4.5)$$

Similarly, we compute \mathbf{x}_d for each document d , integrating out π_d ,

$$p(\mathbf{x}_d | \gamma) = \frac{\Gamma(\gamma_0 + \gamma_1)}{\Gamma(C_{d,\cdot} + \gamma_0 + \gamma_1)} \prod_{x \in \{0,1\}} \frac{\Gamma(C_{d,x} + \gamma_x)}{\Gamma(\gamma_x)} \quad (4.6)$$

Since we explicitly sample the topic ϕ_l at each node l , we need to re-sample all topics for the case that j is the parent of i to compute the ratio $\prod_{l=1}^K \frac{p(\phi_l | j \prec k)}{p(\phi_l | i \prec k)}$.

Given the sampled ϕ , the word likelihood is $p(\mathbf{w}_d | \mathbf{z}_d, \phi) = \prod_{n=1}^{N_d} \phi_{z_{d,n}, w_{d,n}}$. However, re-sampling the topics for the whole hierarchy for every node proposal is inefficient. To avoid that, we keep all ϕ 's fixed and approximate the ratio as:

$$\prod_{d \in \mathcal{D}_{\Delta_k}} \frac{p(\mathbf{w}_d | j \prec k)}{p(\mathbf{w}_d | i \prec k)} \prod_{l=1}^K \frac{p(\phi_l | j \prec k)}{p(\phi_l | i \prec k)} \approx \frac{\int_{\phi_k} p(\mathbf{m}_k + \tilde{\mathbf{m}}_k | \phi_k) p(\phi_k | \phi_j) d\phi_k}{\int_{\phi_k} p(\mathbf{m}_k + \tilde{\mathbf{m}}_k | \phi_k) p(\phi_k | \phi_i) d\phi_k} \quad (4.7)$$

where \mathbf{m}_k is the word counts at node k and $\tilde{\mathbf{m}}_k$ is the word counts propagated from children of k . Since ϕ is fixed and the node assignments \mathbf{z} are unchanged, the word likelihoods cancel out except for tokens assigned at k or any of its children. The integration in Equation 4.7 is

$$\int_{\phi_k} p(\mathbf{m}_k + \tilde{\mathbf{m}}_k | \phi_k) p(\phi_k | \phi_j) d\phi_k = \frac{\Gamma(\beta)}{\Gamma(M_{k,\cdot} + \tilde{M}_{k,\cdot} + \beta)} \prod_{v=1}^V \frac{\Gamma(M_{k,v} + \tilde{M}_{k,v} + \beta\phi_{i,v})}{\Gamma(\beta\phi_{i,v})} \quad (4.8)$$

Using Equations 4.2 and 4.4, we can compute the Metropolis-Hastings acceptance probability.

4.4 Analyzing Political Agendas in U.S. Congresses

To study the U.S. Congress's policy agendas, we obtain both the title and the legislative text of bills in four Congresses (109th–112th) from THOMAS—the online archive of the Library of Congress.⁷ Each bill is tagged with multiple subject terms issued by the Congressional Research Service. Examples of bills' titles and their assigned labels are in Table 4.2.

⁷<https://www.govtrack.us/data/us/>

For each bill, we merge its title and legislative text together. We then perform standard pre-processing steps including tokenization, removing stopwords, stemming, adding bigrams and filtering using TF-IDF to obtain a vocabulary of 10,000 words. The statistics of the data after pre-processing are in Figure 4.5.⁸ We ignore labels associated with fewer than 100 bills.

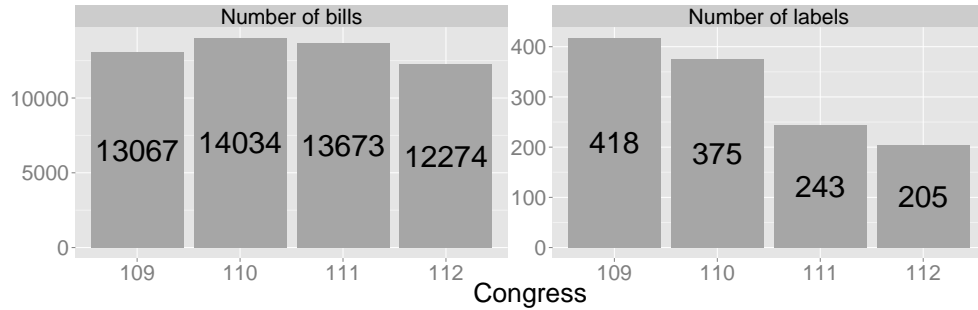


Figure 4.5: Number of bills and unique labels in our dataset after pre-processing for each Congress.

We first qualitatively analyze the hierarchy learned by our model. Figure 4.6 shows a subtree whose root is about INTERNATIONAL AFFAIRS, obtained by running L2H on bills in the 112th U.S. Congress. The learned topic at INTERNATIONAL AFFAIRS shows the focus of 112th Congress on the Arab Spring—a revolutionary wave of demonstrations and protests in Arab countries like Libya and Bahrain. The concept is then split into two distinctive aspects of international affairs: *military* and *diplomacy*. In addition, Figures 4.7 and 4.8 show the subtree in the learned hierarchy rooted at ENVIRONMENTAL ASSESSMENT, MONITORING, RESEARCH and HEALTH respectively.

When analyzing the learned hierarchy, political scientist Kristina Miler (per-

⁸We find bigram candidates that occur at least ten times in the training set and use a χ^2 test to filter out those having a χ^2 value less than 5.0. We then treat selected bigrams as single word types in the vocabulary.

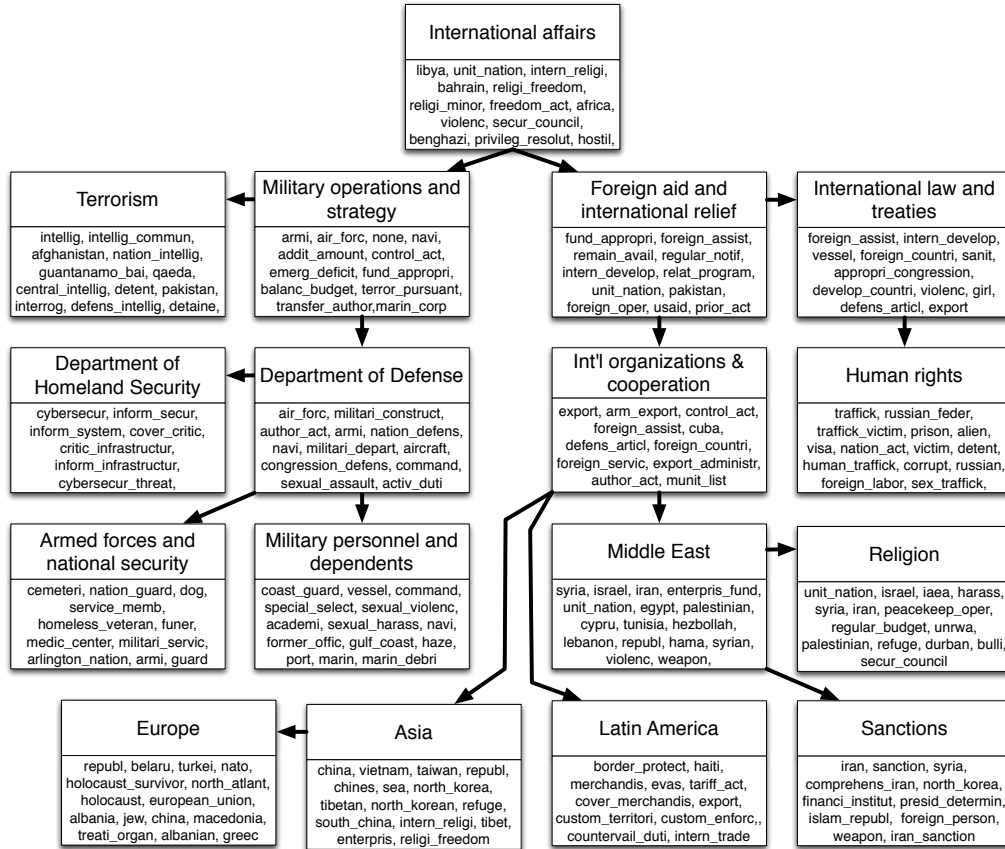


Figure 4.6: A subtree rooted at INTERNATIONAL AFFAIRS in the hierarchy learned by L2H using data from the 112th Congress.

sonal communication) has the following comments:

- On Figure 4.6: “The international affairs topic does an excellent job of capturing the key distinction between military/defense and diplomacy/aid. Even more impressive is that it then also captures the major policy areas within each of these issues: the distinction between traditional military issues and terrorism-related issues, and the distinction between thematic policy (e.g., human rights) and geographic/regional policy.”
- On Figure 4.7: “The environmental policy sub-tree nicely recognizes the importance of water policy within broader environmental policy. Even more

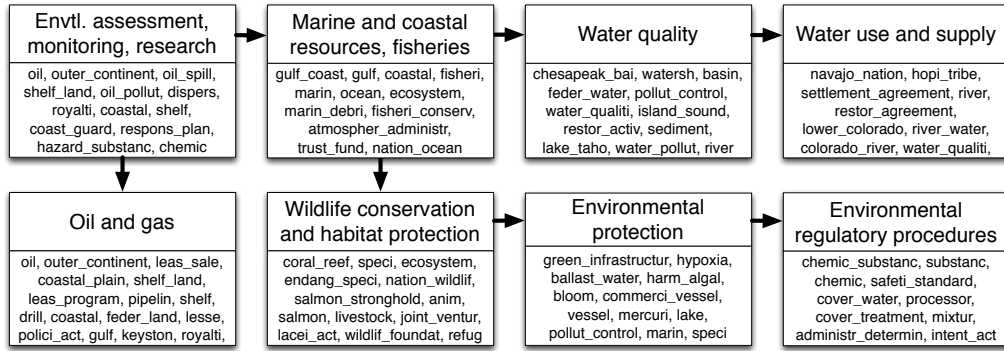


Figure 4.7: A subtree rooted at ENVIRONMENTAL ASSESSMENT, MONITORING, RESEARCH in the hierarchy learned by L2H using data from the 112th Congress.

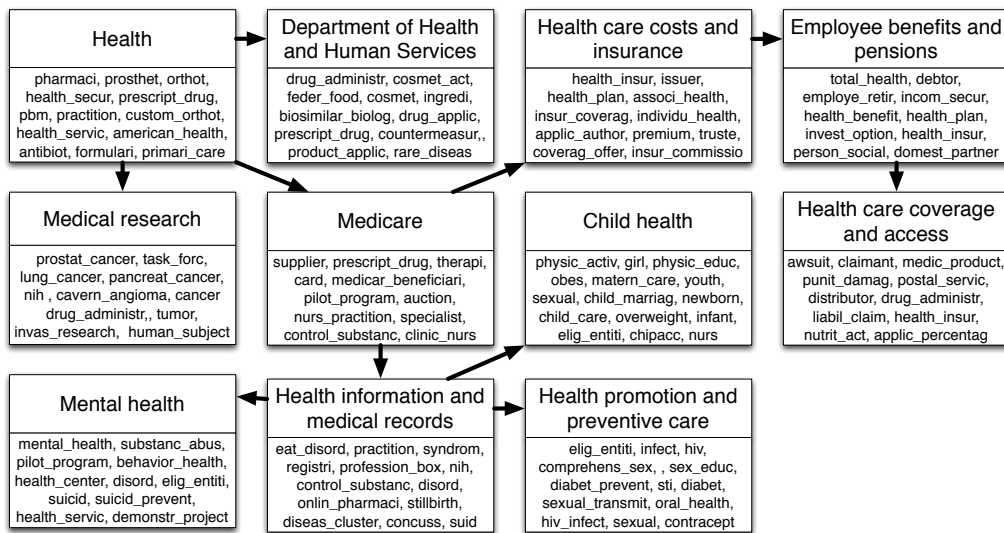


Figure 4.8: A subtree rooted at HEALTH in the hierarchy learned by L2H using data from the 112th Congress.

notable is that the sub-tree captures the unique water needs of the Western states, which is a hallmark of the politics of water (and environmental policy more generally).”

- On Figure 4.8: “Within health policy, there has been increasing attention to issues of childhood obesity and health as well as to mental health issues (both children and adults), and the subtree does a nice job of mirroring the recent attention to these two issue areas.”

4.5 Document Modeling and Classification

To quantitatively evaluate the effectiveness of L2H, we conduct experiments on two computational tasks: document modeling—measuring perplexity on a held-out set of documents, and multi-label classification—predicting multiple labels for unseen documents. For both tasks, we perform 5-fold cross-validation. For each fold, we perform standard pre-processing steps as described in Section 4.4. After building the vocabulary from training documents, we discard all out-of-vocabulary words in the test documents.

4.5.1 Document modeling

In the first quantitative experiment, we focus on the task of predicting the words in held-out test documents, given their labels. This is measured by *perplexity*, a widely-used evaluation metric (Blei et al., 2003b; Wallach et al., 2009). To compute perplexity, we follow the “*estimating θ* ” method described in Wallach et al. (Wallach et al., 2009, Sec. 5.1) and split each test document d into $\mathbf{w}_d^{\text{TE1}}$ and $\mathbf{w}_d^{\text{TE2}}$. During training, we estimate all topics’ distributions over the vocabulary $\hat{\phi}$. During test, first we run Gibbs sampling using the learned topics on $\mathbf{w}_d^{\text{TE1}}$ to estimate the topic proportions $\hat{\theta}_d^{\text{TE}}$ for each test document d . Then, we compute the perplexity on the held-out words $\mathbf{w}_d^{\text{TE2}}$ as

$$\exp \left\{ - \frac{\sum_d \log \left(p(\mathbf{w}_d^{\text{TE2}} | \mathbf{l}_d, \hat{\theta}_d^{\text{TE}}, \hat{\phi}) \right)}{N^{\text{TE2}}} \right\}$$

where N^{TE_2} is the total number of tokens in $\mathbf{w}_d^{\text{TE}_2}$.

Setup: We compare our proposed model L2H with the following methods:

- LDA (Blei et al., 2003b): unsupervised topic model with a flat topic structure. In our experiments, we set the number of topics of LDA equal to the number of labels in each dataset.
- L-LDA (Ramage et al., 2009): associates each topic with a label, and a document is generated using the topics associated with the document’s labels only.
- L2F (Label-to-Flat structure): a simplified version of L2H with a fixed, flat topic structure. The major difference between L2F and L-LDA is that L2F allows tokens to be drawn from topics that are not in the document’s label set via the use of the switching variable (Section 4.2.3). Improvements of L2H over L2F show the importance of the *hierarchical* structure.

For all models, the number of topics is the number of labels in the dataset. We run for 1,000 iterations on the training data with a burn-in period of 500 iterations. After the burn-in period, we store ten sets of estimated parameters, one after every fifty iterations. During test time, we run ten chains using these ten learned models on the test data and compute the perplexity after 100 iterations. The perplexity of each fold is the average value over the ten chains as described in Chapter 2.

Results: Figure 4.9 shows the perplexity of the four models averaged over five folds on the four datasets. LDA naturally outperforms the other models with labels when

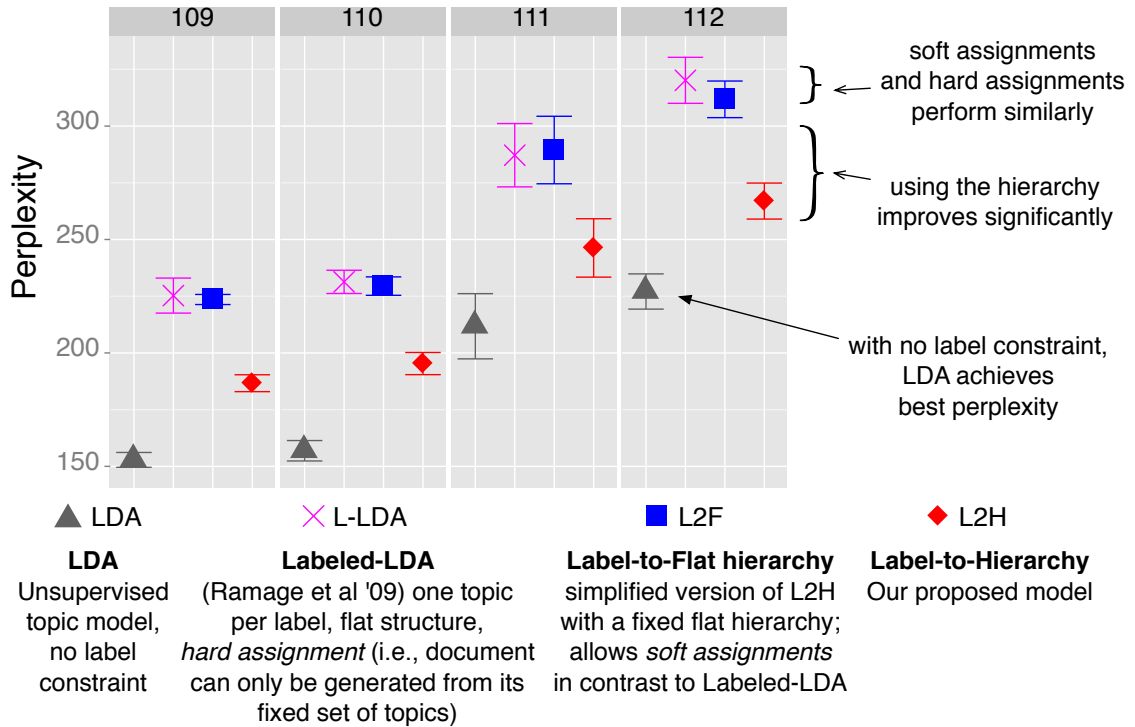


Figure 4.9: Perplexity on held-out documents, averaged over 5 folds (lower is better).

measuring perplexity since it can freely optimize the likelihood without additional constraints. LDA does that, however, at the cost of learning less interpretable set of topics, each of which is not associated with a predefined label.

Among the models using labels, L-LDA and L2F are comparable which shows that soften the constraint of assigning tokens of a document to only labels associated with that document does not help. However, L2H significantly outperforms both L-LDA and L2F which shows that when incorporating labels, using an additional topic hierarchy improves the predictive power and generalizability of L-LDA.

4.5.2 Multi-label Classification

Multi-label classification is predicting a set of labels for a test document given its text (Tsoumakias et al., 2010; Madjarov et al., 2012; Zhang and Zhou, 2014). The prediction is from a set of pre-defined K labels and each document can be tagged with any of the 2^K possible subsets. In this experiment, we use M3L—an efficient max-margin multi-label classifier (Hariharan et al., 2012)—to study how features extracted from our L2H improve classification.

We use F_1 as the evaluation metric. The F_1 score is first computed for each document d as $F_1(d) = 2 P(d) R(d)/(P(d) + R(d))$, where $P(d)$ and $R(d)$ are the precision and recall for document d . After $F_1(d)$ is computed for all documents, the overall performance can be summarized by micro-averaging and macro-averaging to obtain Micro- F_1 and Macro- F_1 respectively. In macro-averaging, F_1 is first computed for each document using its own confusion matrix and then averaged. In micro-averaging, on the other hand, only a single confusion matrix is computed for all documents, and the F_1 score is computed based on this single confusion matrix (Rubin et al., 2012).

Setup: We use the following sets of features:

- TF: Each document is represented by a vector of term frequency of all word types in the vocabulary.
- TF-IDF: Each document is represented by a vector $\psi_d^{\text{TF-IDF}}$ of TF-IDF of all word types.

- L-LDA&TF-IDF: [Ramage et al. \(2010a\)](#) combine L-LDA features and TF-IDF features to improve the performance on recommendation tasks. Likewise, we extract a K -dimensional vector $\hat{\theta}_d^{\text{L-LDA}}$ and concatenate with TF-IDF vector $\psi_d^{\text{TF-IDF}}$ to form the feature vector of L-LDA&TF-IDF.⁹
- L2H&TF-IDF: Similarly, we concatenate TF-IDF with the features $\hat{\theta}_d^{\text{L2H}} = \{\hat{\theta}_d^0, \hat{\theta}_d^1\}$ extracted using L2H (same MCMC setup as L-LDA).

One complication for L2H is the candidate label set \mathcal{L}_d^1 , which is not observed during test time. Thus, during test time, we estimate \mathcal{L}_d^1 using TF-IDF. Let \mathcal{D}_l be the set of documents tagged with label l . For each l , we compute a TF-IDF vector $\phi_l^{\text{TF-IDF}} = \text{avg}_{d \in \mathcal{D}_l} \psi_d^{\text{TF-IDF}}$. Then for each document d , we generate the k nearest labels using cosine similarity, and add them to the candidate label set \mathcal{L}_d^1 of d . Finally, we expand this initial set by adding all labels on the paths from the root of the learned hierarchy to any of the k nearest labels (Figure 4.3). We explored different values of $k \in \{3, 5, 7, 9\}$, with similar results; the results in this section are reported with $k = 5$.

Results: Figure 4.10 shows classification results. For both Macro- F_1 and Micro- F_1 , TF-IDF, L-LDA&TF-IDF and L2H&TF-IDF significantly outperform TF. Also, L-LDA&TF-IDF performs better than TF-IDF, which is consistent with [Ramage et al. \(2010a\)](#). L2H&TF-IDF performs better than L-LDA&TF-IDF, which in turn performs better than TF-IDF. This shows that features extracted from L2H are

⁹We run L-LDA on the training data for 1,000 iterations and store ten models, each 50 iterations apart, after 500 burn-in iterations. For each model, we sample assignments for all tokens using 100 iterations and average over chains to estimate $\hat{\theta}_d^{\text{L-LDA}}$.

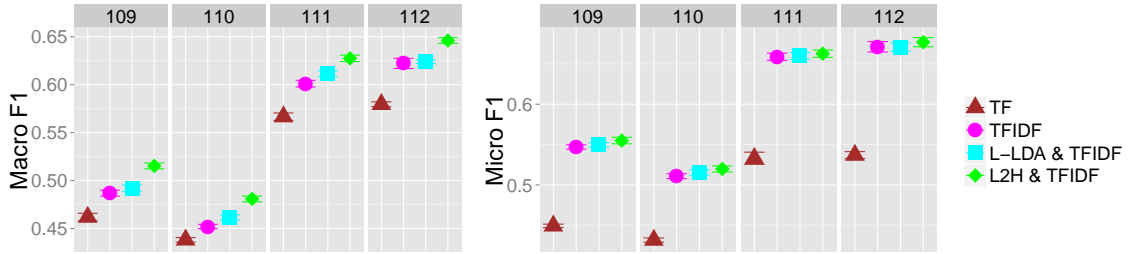


Figure 4.10: Multi-label classification results. The results are averaged over 5 folds.

more predictive than those extracted from L-LDA, and both improve classification. The improvements of L2H&TF-IDF and L-LDA&TF-IDF over TF-IDF are clearer for Macro- F_1 compared with Micro- F_1 , which shows that features from both topic models help improve prediction, regardless of the frequencies of their tagged labels.

4.6 Summary

In this chapter, we focus on analyzing policy agendas in legislative text using the topic modeling approach. To improve the model’s interpretability and thus reduce the post-analysis cost of analyzing and interpreting the results, we leverage a set of multi-labeled data, in which each legislative text is tagged with multiple policy agenda issues from a flexible list of labels. There are two major advantages of using this type of labeled data over traditional single-labeled data using a fixed coding system: (1) it captures the multi-faceted nature of many Congressional bills, and (2) it helps reduce the pre-analysis cost of creating and maintaining the well-defined coding system. However, the large label space also incurs new challenges for the learning techniques. Any effective automated methods should be able to (1) capture the dependencies among the labels and (2) handle missing annotated labels.

We introduce L2H, a hierarchical topic model, to address these problems. L2H captures the dependencies among labels using an interpretable tree-structured hierarchy, in which each node is associated with a label—a pre-defined word or phrase from the label list, and a topic—a multinomial distribution over the vocabulary. We apply L2H on a set of legislative bill text from four U.S. Congresses. Qualitative analysis of the results shows that L2H can learn interpretable label hierarchies, which helps provide insights about the political attentions that policymakers focus on, and how those policy issues relate to each other. Empirical results also show the effectiveness of L2H on two computational tasks: predicting held-out words and predicting multiple labels for unseen documents.

Specifically for the problem of studying agendas in political text, although in this chapter we focus on using multi-labeled data with the policy agenda issues provided by the Congressional Research Service, we are not suggesting using this to replace the Policy Agendas Topics codebook. Instead, we want to suggest a complementary resource with relatively cheaper pre-analysis cost for agenda-setting research and also present a computational method to address some specific challenges that these data incur.

As discussed in Section 4.1.2, besides the multi-labeled legislative text which is the focus of this work, multi-labeled data are ubiquitous and can be found in various settings. One potential future direction is to apply L2H to other settings to help improve the multi-label classification performance, which is the focus of much recent in machine learning (Tsoumakas et al., 2010; Madjarov et al., 2012; Rubin et al., 2012; Zhang and Zhou, 2014). In addition, another major advantage of L2H is

the interpretable label hierarchy, which can be used to make *searching* and *browsing* large-scale data collection more effective.

Chapter 5: Discovering Agendas and Frames in Ideologically Polarized Text

5.1 Introduction

In the two previous chapters, we have focused on developing effective topic models to discover and analyze agenda issues from political text. In this chapter, we go beyond agenda-setting (i.e., what topics people talk about) and expand our focus to *framing* (i.e., how they talk about different issues). We introduce SHLDA—*Supervised Hierarchical Latent Dirichlet Allocation*—which can discover a hierarchy of topics from a collection of documents, each associated with the ideological score of the author. In the learned hierarchy, first-level nodes map to agenda issues while second-level nodes represent ideologically polarized frames of the corresponding issue.

Although inspired by the study of political discourse, SHLDA is applicable to a much broader setting: a collection of text, in which each document is associated with a continuous response variable of interest. Examples of this type of data include product reviews with their corresponding ratings (Pang and Lee, 2005), online status updates with their corresponding geo-tagged latitudes and longitudes (Eisenstein

et al., 2010), and students' essays with their accompanying depression scores (Resnik et al., 2013). Our model extends the nested Chinese restaurant processes to discover tree-structured topic hierarchies and uses both per-topic hierarchical and per-word lexical regression parameters to model the response variable. Experiments on political text and product reviews show that SHLDA is able to discover meaningful topic hierarchies to provide insight into how issues under discussion are framed; while improving the performance in predicting political ideologies and online review ratings.

This chapter revises and extends the work originally published in (Nguyen et al., 2013c).

5.1.1 Framing: Going beyond Agenda-setting to Understand How Things are Talked About

How do liberal-leaning bloggers talk about immigration in the U.S.? What do conservative politicians have to say about education? How do Fox News and MSNBC differ in their language about the gun debate? Such questions concern not only *what*, but *how* things are talked about. In the previous two chapters, we have focused on *agenda-setting*, which concerns *what* issues are introduced into political discourses (e.g., political debates, legislative proceedings etc) and their influence over public priorities (McCombs and Shaw, 1993) and policy agendas (Baumgartner and Jones, 1993b). The question of *how* concerns *framing*: the way presentation of an issue reflects or encourages a particular perspective or interpretation (McCombs

and Ghanem, 2001). “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, italics originally by the author).

By highlighting a particular perspective or interpretation and deemphasizing others, it is widely accepted that framing can have significant influence on public opinions towards important policy issues (Scheufele, 1999; Chong and Druckman, 2007). For example, when covering news or events leading toward the application of the death penalty, mass media might use the “morality frame” to discuss whether it is right or wrong to kill as punishment. On the other hand, the use of the “innocence frame” which emphasizes the irreversible consequence of mistaken convictions, has led to a sharp decline in the use of capital punishment in the U.S. (Dardis et al., 2008; Baumgartner et al., 2008). As another example, when discussing the issue of legalizing marijuana, news articles might focus on different frames such as the “economic frame” (e.g., stories emphasizing the cost of the drug war and the potential revenue through legalizing and regulating the market), the “health frame” (e.g., stories on the health benefits that medical marijuana can provide), and the “legal frame” (e.g., stories on the conflicts between federal and state regularization when marijuana becomes legal) (Boydston and Gross, 2014).

5.1.2 Framing as Second-level Agenda-setting

Despite its omnipresence in social sciences, framing research lacks “a commonly shared theoretical model” (Scheufele, 1999). Boydstun et al. (2013c) argue that “the very definition of framing has been notoriously slippery”, for which Entman (1993) refers framing as a “scattered conceptualization” and a “fractured paradigm”. Interestingly, one line of communication theory seeks to unify agenda setting and framing by viewing frames as *second-level agendas*: just as agenda setting is about which objects of discussion are salient, framing is about the salience of *attributes* of those objects (McCombs, 2004). The key is that what communications theorists consider an attribute in a discussion can itself be an object, as well. For example, “mistaken convictions” is one attribute of the death penalty discussion, but it can also be viewed as an object of discussion in its own right.

This two-level view leads naturally to the idea of using a *hierarchical topic model* to formalize both agenda-setting and framing within a uniform setting. In previous chapters, we have used topic models to capture agendas in which each topic (i.e., a distribution over the vocabulary) maps to an agenda issue. In this chapter and the next chapter, we introduce models to discover hierarchy of topics, in which higher-level nodes in the hierarchy map to agenda issues while lower-level nodes map to issue-specific frames.

To learn politically meaningful topics to discover frames, we jointly model the text and its author’s *ideological score*—a continuous response variable. Figure 5.1 illustrates the hierarchical output that SHLDA learns. In this hierarchy, the first-

level nodes are agenda issues while each second-level node represents a frame of the corresponding issue. For example, the hierarchy illustrates three central political agenda issues: “Health care”, “Environment”, and “Economy”. When discussing the environmental issue, policymakers and the media might use different frames such as the “nature frame” (e.g., supporting policies that prevent global warming and opposing), the “externalities frame” (e.g., studying the cost that the air pollution from motor vehicles incurs on the society), and the “industry frame” (e.g., analyzing how extreme Environmental Protection Agency regulations would hurt the industry and create job losses) (Lakoff, 2010).

In addition to the topic, SHLDA also learns for each node regression parameter, which indicates where on the ideological spectrum that the node falls onto. For example, Figure 5.1 illustrates that liberals are more likely to talk about the environmental issue using the “nature frame”, while conservatives often focus on discussing the “industry frame”.

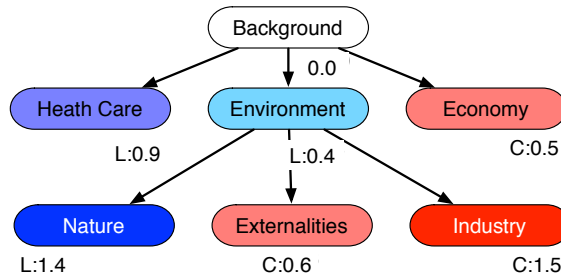


Figure 5.1: Example hierarchy with ideologically polarized topics that SHLDA learns. First-level nodes map to agenda issues, while second-level nodes map to issue-specific frames. Each node is associated with a topic (i.e., a multinomial distribution over words) and an ideological score.

5.1.3 Framing Research: Traditional vs. Data-driven Approach

Like many concepts in political science in particular and social science in general, traditional approach for studying framing is heavily based on close reading and manual content analysis. However, due to its complex and abstract nature, research on framing lacks a general framework or a shared coding system, like the Policy Agendas Topics Codebook for studying agenda-setting, that can be used across different studies. In a recent study, [Matthes and Kohring \(2008\)](#) survey five common methods that researchers have been using to study framing and argue “a frame is a quite abstract variable that is hard to identify and hard to code in content analysis”.

Due to this “lack of a commonly shared theoretical model underlying framing research” ([Scheufele, 1999](#)), much of prior research on framing is *issue-specific*, in which researchers focus specifically on an issue or an event and analyze different frames qualitatively based on relatively small samples. Although in this type of studies, frames are analyzed extensively, it is often difficult to generalize the process from which the frames are extracted from the materials, which makes it very challenging to apply the same approach for different issues or events. On the other hand, despite the challenge, [Boydston and Gross \(2014\)](#) have started an ambitious project to define a Policy Frames Codebook to examine framing both within and across issues. Just like the Policy Agendas Topics Codebook provides a system for categorizing policy agenda issues, the Policy Frames Codebook aims to provide a system for categorizing frames across different policy issues ([Boydston et al., 2013c](#)).

Different from these traditional methods, in this thesis, we take a *data-driven*

approach to study framing using automated content analysis. We follow the line of framing research, considering framing as second-level agenda-setting, to design probabilistic topic models to discover hierarchy of topics in which higher-level nodes map to policy agenda issue and lower-level nodes map to issue-specific frames. To capture more interesting, interpretable frames, we jointly model the text with additional information of political actors such as their pre-estimated ideological score or their voting records.

Using this data-driven approach to study framing has several advantages. First, just like other automated content analysis, it enjoys relatively low pre-analysis cost, which allows us to study and analyze a large collection of text (Chapter 1). Second, the discovered hierarchy of topics provides a natural way to study framing for multiple issues discussed in the text at the same time. However, one major drawback of using automated content analysis methods for framing study is the interpretability of the discovered topic hierarchy. Learning a coherent topic hierarchy from text is a challenging task and has been an active research area in topic modeling community. In this chapter we present an attempt to overcome this challenge.

5.1.4 Chapter Structure

We describe the model SHLDA in detail next in Section 5.2. In Section 5.3, we describe an inference algorithm using stochastic EM, which learns SHLDA's posterior distribution given the observed data. To evaluate the effectiveness of SHLDA, we use three datasets described in Section 5.4. Section 5.5 qualitatively

analyzes the topic hierarchies discovered by SHLDA and Section 5.6 shows SHLDA’s improvements over various baseline methods in predicting the documents’ responses. Section 5.7 concludes the chapter and opens up some directions for future work.

5.2 SHLDA: Capturing Text and Continuous Response using Hierarchical Topic Structure

Jointly capturing supervision and hierarchical topic structure falls under a class of models called *supervised hierarchical latent Dirichlet allocation*. These models take as input a set of D documents, each of which is associated with a response variable y_d , and output a hierarchy of topics which is informed by y_d . Figure 5.2 shows the plate notation diagram of SHLDA, whose generative process is:¹

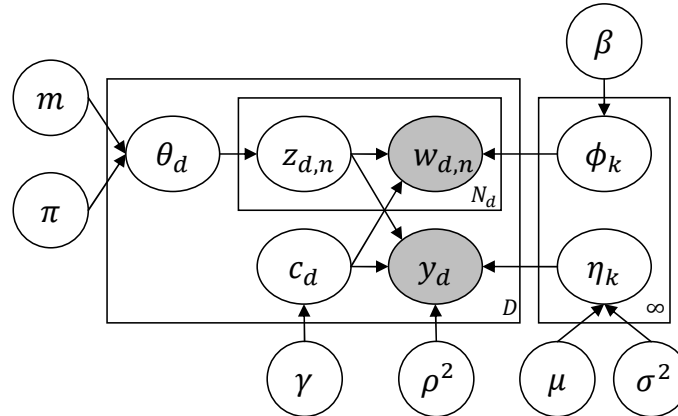


Figure 5.2: Plate notation diagram of our SHLDA model.

1. For each node $k \in [1, \infty)$ in the tree

¹Zhang (2012) also introduces a model called supervised hierarchical latent Dirichlet allocation for modeling a set of documents, each of which is associated with categorical response. In this thesis, we refer to Zhang (2012)’s model as *multi-class SHLDA*, which is consistent with the way prior works named the non-hierarchical counterparts: sLDA by Blei and McAuliffe (2007) for continuous response and multi-class sLDA by Wang et al. (2009) for categorical response.

- (a) Draw topic $\phi_k \sim \text{Dir}(\beta_k)$
 - (b) Draw regression parameter $\eta_k \sim \mathcal{N}(\mu, \sigma)$
2. For each word $v \in [1, V]$, draw $\tau_v \sim \text{Laplace}(0, \omega)$
3. For each document $d \in [1, D]$
- (a) Draw level distribution $\theta_d \sim \text{GEM}(m, \pi)$
 - (b) Draw table distribution $\psi_d \sim \text{GEM}(\alpha)$
 - (c) For each table $t \in [1, \infty)$, draw a path $c_{d,t} \sim \text{nCRP}(\gamma)$
 - (d) For each sentence $s \in [1, S_d]$, draw a table indicator $t_{d,s} \sim \text{Mult}(\psi_d)$
 - i. For each token $n \in [1, N_{d,s}]$
 - A. Draw level $z_{d,s,n} \sim \text{Mult}(\theta_d)$
 - B. Draw word $w_{d,s,n} \sim \text{Mult}(\phi_{c_{d,t_{d,s}}, z_{d,s,n}})$
 - (e) Draw response $y_d \sim \mathcal{N}(\boldsymbol{\eta}^T \bar{\mathbf{z}}_d + \boldsymbol{\tau}^T \bar{\mathbf{w}}_d, \rho)$:
 - i. $\bar{z}_{d,k} = \frac{1}{N_{d,\cdot}} \sum_{s=1}^{S_d} \sum_{n=1}^{N_{d,s}} \mathbb{I}[k_{d,s,n} = k]$
 - ii. $\bar{w}_{d,v} = \frac{1}{N_{d,\cdot}} \sum_{s=1}^{S_d} \sum_{n=1}^{N_{d,s}} \mathbb{I}[w_{d,s,n} = v]$

5.2.1 Generating Text

At its core, SHLDA’s document generative process resembles a combination of hierarchical latent Dirichlet allocation (Blei et al., 2010b, hLDA) and the hierarchical Dirichlet process (Teh et al., 2006, HDP), both of which we review in Chapter 2. hLDA uses the nested Chinese restaurant process (nCRP(γ)), combined with an appropriate base distribution, to induce an unbounded tree-structured hierarchy of topics: each node contains one topic and general topics at the top, specific at the

bottom. A document is generated by traversing this tree, at each level creating a new child (hence a new path) with probability proportional to γ or otherwise respecting the “rich-get-richer” property of a CRP.

A drawback of hLDA, however, is that each document is restricted to only a *single path* in the tree. Since each path is designed to capture a consistent theme, from more general (i.e., at higher-level nodes) to more specific (i.e., at lower-level nodes), restricting a document to be about a theme is a relatively strong assumption, especially when modeling long documents. Recent work relaxes this restriction through different priors: nested hierarchical Dirichlet processes (Paisley et al., 2014, nHDP), nested Chinese restaurant franchises (Ahmed et al., 2013a, nCRF) or recursive Chinese restaurant processes (Kim et al., 2012, rCRP). In this chapter, we address this problem by allowing documents to have *multiple paths* through the tree by leveraging information at the sentence level using the two-level structure used in HDP. More specifically, in the HDP’s Chinese restaurant franchise metaphor, customers (i.e., tokens) are grouped by sitting at tables and each table takes a dish (i.e., topic) from a *flat* global menu. In our SHLDA, dishes are organized in a *tree-structured* global menu by using the nCRP as prior. Each path in the tree is a collection of L dishes (one for each level) and is called a *combo*. SHLDA groups sentences of a document by assigning them to tables and associates each table with a combo, and thus, models each document as a *distribution* over combos.²

In SHLDA’s metaphor, customers come in a restaurant and sit at a table in

²We emphasize that, unlike in HDP where each table is assigned to a single dish, each *table* in our metaphor is associated with a *combo*—a collection of L dishes. We also use *combo* and *path* interchangeably.

groups, where each group is a sentence. A sentence $w_{d,s}$ enters restaurant d and selects a table t (and its associated combo) with probability proportional to the number of sentences $S_{d,t}$ at that table; or, it sits at a new table with probability proportional to α . After choosing the table (indexed by $t_{d,s}$), if the table is new, the group will select a combo of dishes (i.e., a path, indexed by $c_{d,t}$) from the tree menu. Once a combo is in place, each token in the sentence chooses a “level” (indexed by $z_{d,s,n}$) in the combo, which specifies the topic ($\phi_{k_{d,s,n}} \equiv \phi_{c_{d,t_{d,s}}, z_{d,s,n}}$) producing the associated observation (Figure 5.3).

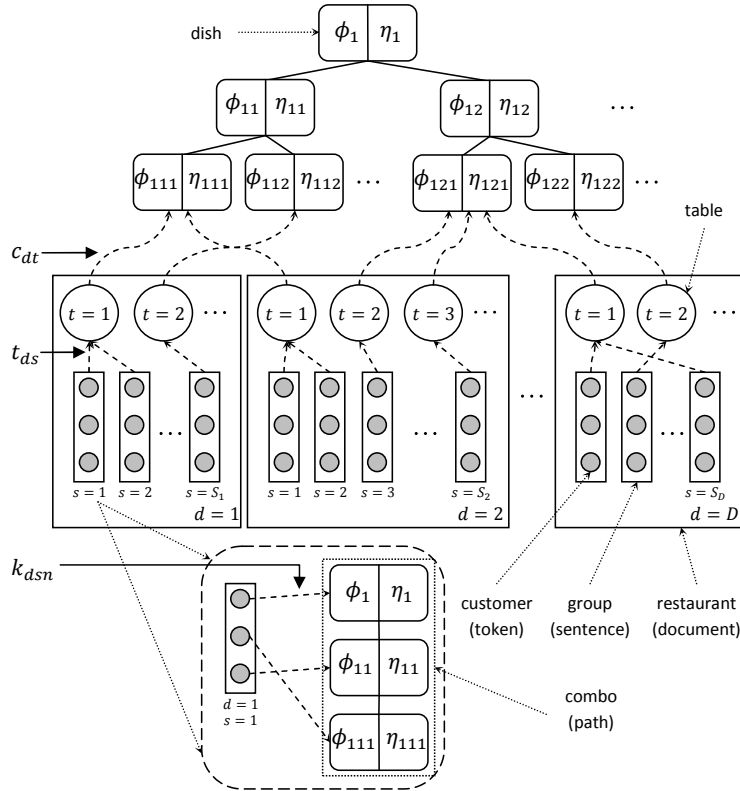


Figure 5.3: Illustration of SHLDA’s restaurant franchise metaphor.

5.2.2 Generating Responses

SHLDA also draws on supervised LDA (Blei and McAuliffe, 2007, sLDA) associating each document d with an observable continuous response variable y_d that represents the author’s perspective toward a topic, e.g., positive vs. negative sentiment, conservative vs. liberal ideology, etc. This lets us infer a multi-level topic structure informed by how topics are framed with respect to positions along the y_d continuum.

Unlike sLDA, we model the response variables using a Gaussian linear regression that contains *both* per-topic hierarchical and per-word lexical regression parameters. The *hierarchical regression parameters* are just like topics’ regression parameters in sLDA: each topic k (here, a tree node) has a parameter η_k , and the model uses the empirical distribution over the nodes that generated a document as the regressors. However, the hierarchy in SHLDA makes it possible to discover relationships between topics and the response variable that sLDA’s simple latent space obscures. Consider, for example, a topic model trained on Congressional debates. Vanilla LDA would likely discover a *healthcare* category. sLDA (Blei and McAuliffe, 2007) could discover a pro-Obamacare topic and an anti-Obamacare topic. SHLDA could do that *and* capture the fact that there are alternative perspectives, i.e., that the healthcare issue is being discussed from two ideological perspectives, along with characterizing *how* the higher level topic is discussed by those on both sides of that ideological debate.

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* to model the interaction between specific words and response variables. We denote the regression parameter associated with a word type v in the vocabulary as τ_v , and use the normalized frequency of v in the documents to be its regressor.

Including both hierarchical and lexical parameters is important. For detecting ideology in the U.S., “liberty” is an effective indicator of conservative speakers regardless of context; however, “cost” is a conservative-leaning indicator in discussions about environmental policy but liberal-leaning in debates about foreign policy. For sentiment, “wonderful” is globally a positive word; however, “unexpected” is a positive descriptor of books but a negative one of a car’s steering. SHLDA captures these properties in a single model.

5.3 Posterior Inference and Optimization

Given documents with observed words $\mathbf{w} = \{w_{d,s,n}\}$ and response variables $\mathbf{y} = \{y_d\}$, the inference task is to find the posterior distribution over: the tree structure including topic ϕ_k and regression parameter η_k for each node k , combo assignment $c_{d,t}$ for each table t in document d , table assignment $t_{d,s}$ for each sentence s in a document d , and level assignment $z_{d,s,n}$ for each token $w_{d,s,n}$. We approximate SHLDA’s posterior using stochastic EM, which alternates between a Gibbs sampling E-step and an optimization M-step. More specifically, in the E-step, we integrate out $\boldsymbol{\psi}$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ to construct a Markov chain over $(\mathbf{t}, \mathbf{c}, \mathbf{z})$ and alternate sampling

| | |
|----------------------|--|
| S_d | # sentences in document d |
| $S_{d,t}$ | # groups (i.e. sentences) sitting at table t in restaurant d |
| $N_{d,s}$ | # tokens $\mathbf{w}_{d,s}$ |
| $N_{d,\cdot,l}$ | # tokens in \mathbf{w}_d assigned to level l |
| $N_{d,\cdot,>l}$ | # tokens in \mathbf{w}_d assigned to level $> l$ |
| $N_{d,\cdot,\geq l}$ | $\equiv N_{d,\cdot,l} + N_{d,\cdot,>l}$ |
| $M_{c,l}$ | # tables at level l on path c |
| $C_{c,l,v}$ | # word type v assigned to level l on path c |
| $C_{d,x,l,v}$ | # word type v in $\mathbf{v}_{d,x}$ assigned to level l |
| ϕ_k | Topic at node k |
| η_k | Regression parameter at node k |
| τ_v | Regression parameter of word type v |
| $c_{d,t}$ | Path assignment for table t in restaurant d |
| $t_{d,s}$ | Table assignment for group $\mathbf{w}_{d,s}$ |
| $z_{d,s,n}$ | Level assignment for $w_{d,s,n}$ |
| $k_{d,s,n}$ | Node assignment for $w_{d,s,n}$ (i.e., node at level $z_{d,s,n}$ on path $c_{d,t_{d,s}}$) |
| L | Height of the tree |
| \mathcal{C}^+ | Set of all possible paths (including new ones) of the tree |

Table 5.1: Notation used in this chapter

each of them from their conditional distributions. In the M-step, we optimize the regression parameters $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ using L-BFGS (Liu and Nocedal, 1989). Before describing each step in detail, let us define the following quantities.

- First, define $\mathbf{v}_{d,x}$ as a set of tokens (e.g., a token, a sentence or a set of sentences) in document d . The conditional density of an arbitrary set of tokens $\mathbf{v}_{d,x}$ in document d being assigned to path c given all other assignments is

$$f_c^{-d,x}(\mathbf{v}_{d,x}) = \prod_{l=1}^L \frac{\Gamma(C_{c,l,\cdot}^{-d,x} + V\beta_l)}{\Gamma(C_{c,l,\cdot}^{-d,x} + C_{d,x,l,\cdot} + V\beta_l)} \prod_{v=1}^V \frac{\Gamma(C_{c,l,v}^{-d,x} + C_{d,x,l,v} + \beta_l)}{\Gamma(C_{c,l,v}^{-d,x} + \beta_l)} \quad (5.1)$$

where we use $\mathbf{v}_{d,x,l}$ to denote the set of tokens in $\mathbf{v}_{d,x}$ that are assigned to level l . We use $C_{c,l,v}$ to denote the number of times word type v is assigned to node at level l on path c , and $C_{d,x,l,v}$ to denote the number of times word type v in $\mathbf{v}_{d,x}$ is assigned to node at level l on path c . Superscript $^{-d,x}$ denotes the same count excluding the assignments of $\mathbf{v}_{d,x}$. Marginal counts are represented by

·'s. For a new path c^{new} , if the node does not exist, $C_{c^{new},l,v}^{-d,x} = 0$ for all word types v .

- Second, define the conditional density of the response variable y_d of document d given the set of $\mathbf{v}_{d,x}$ being assigned to path c and all other assignments as $g_c^{-d,x}(y_d) \equiv P(y_d | \mathbf{c}_{d,x}, \mathbf{c}^{-d,x}, \mathbf{z}, \mathbf{t})$ which is a Gaussian with variance ρ and mean

$$\frac{1}{N_{d,\cdot}} \left(\underbrace{\sum_{\mathbf{w}_{d,s,n} \in \{\mathbf{w}_d \setminus \mathbf{v}_{d,x}\}} \eta_{c_{d,t_d,s}, z_{d,s,n}}}_{\text{other words' topic regression}} + \underbrace{\sum_{l=1}^L \eta_{c,l} \cdot C_{d,x,l,\cdot}}_{\mathbf{v}_{d,x} \text{'s topic regression}} + \underbrace{\sum_{s=1}^{S_d} \sum_{n=1}^{N_{d,s}} \tau_{\mathbf{w}_{d,s,n}}}_{\text{doc. lexical regression}} \right) \quad (5.2)$$

where $N_{d,\cdot}$ is the total number of tokens in document d . For a new node at level l on a new path c^{new} , we integrate over all possible values of $\eta_{c^{new},l}$. For new node at level l on a new path c^{new} , we integrate over all possible values of $\eta_{c^{new},l}$ by using the following property of Gaussian distribution

$$\int \mathcal{N}(a + bx; y, \sigma_x) \mathcal{N}(y; \mu, \sigma_y) dy = \mathcal{N}(a + bx; \mu, b^2\sigma_x + \sigma_y) \quad (5.3)$$

Sampling \mathbf{t} : For each group $\mathbf{w}_{d,s}$ we need to sample a table $t_{d,s}$. The conditional distribution of a table t given $\mathbf{w}_{d,s}$ and other assignments is proportional to the number of sentences sitting at t times the probability of $\mathbf{w}_{d,s}$ and y_d being observed under this assignment. This is

$$P(t_{d,s} = t \mid \text{rest})$$

$$\begin{aligned} &\propto P(t_{d,s} = t \mid \mathbf{t}_d^{-s}) \cdot P(\mathbf{w}_{d,s}, y_d \mid t_{d,s} = t, \mathbf{w}^{-d,s}, \mathbf{t}^{-d,s}, \mathbf{z}, \mathbf{c}, \boldsymbol{\eta}) \\ &\propto \begin{cases} S_{d,t}^{-d,s} \cdot f_{c_{d,t}}^{-d,s}(\mathbf{w}_{d,s}) \cdot g_{c_{d,t}}^{-d,s}(y_d), & \text{for existing table } t; \\ \alpha \cdot \sum_{c \in \mathcal{C}^+} P(c_{d,t^{new}} = c \mid \mathbf{c}^{-d,s}) \cdot f_c^{-d,s}(\mathbf{w}_{d,s}) \cdot g_c^{-d,s}(y_d), & \text{for new table } t^{new}. \end{cases} \end{aligned}$$

For a new table t^{new} , we need to sum over all possible paths \mathcal{C}^+ of the tree, including new ones. For example, the set \mathcal{C}^+ for the tree shown in Figure 5.3 consists of four existing paths (ending at one of the four leaf nodes) and three possible new paths (a new leaf off of one of the three internal nodes). The prior probability of path c is:

$$P(c_{d,t^{new}} = c \mid \mathbf{c}^{-d,s}) \propto$$

$$\begin{cases} \prod_{l=2}^L \frac{M_{c,l}^{-d,s}}{M_{c,l-1}^{-d,s} + \gamma_{l-1}}, & \text{for an existing path } c; \\ \frac{\gamma_{l^*}}{M_{c^{new},l^*}^{-d,s} + \gamma_{l^*}} \prod_{l=2}^{l^*} \frac{M_{c^{new},l}^{-d,s}}{M_{c^{new},l-1}^{-d,s} + \gamma_{l-1}}, & \text{for a new path } c^{new} \text{ which consists of an existing path} \\ & \text{from the root to a node at level } l^* \text{ and a new node.} \end{cases} \quad (5.4)$$

Sampling \mathbf{z} : After assigning a sentence $\mathbf{w}_{d,s}$ to a table, we assign each token $w_{d,s,n}$ to a level to choose a dish from the combo. The probability of assigning $w_{d,s,n}$ to level l is

$$P(z_{d,s,n} = l \mid \text{rest}) \propto P(z_{d,s,n} = l \mid \mathbf{z}_d^{-s,n}) P(w_{d,s,n}, y_d \mid z_{d,s,n} = l, \mathbf{w}^{-d,s,n}, \mathbf{z}^{-d,s,n}, \mathbf{t}, \mathbf{c}, \boldsymbol{\eta}) \quad (5.5)$$

The first factor captures the probability that a customer in restaurant d is assigned to level l , conditioned on the level assignments of all other customers in restaurant d , and is equal to

$$P(z_{d,s,n} = l | \mathbf{z}_d^{-s,n}) = \frac{m\pi + N_{d,\cdot,l}^{-d,s,n}}{\pi + N_{d,\cdot,\geq l}^{-d,s,n}} \prod_{j=1}^{l-1} \frac{(1-m)\pi + N_{d,\cdot,>j}^{-d,s,n}}{\pi + N_{d,\cdot,\geq j}^{-d,s,n}}, \quad (5.6)$$

The second factor is the probability of observing $w_{d,s,n}$ and y_d , given that $w_{d,s,n}$ is assigned to level l : $P(w_{d,s,n}, y_d | z_{d,s,n} = l, \mathbf{w}^{-d,s,n}, \mathbf{z}^{-d,s,n}, \mathbf{t}, \mathbf{c}, \boldsymbol{\eta}) = f_{c_{d,t},s}^{-d,s,n}(w_{d,s,n}) \cdot g_{c_{d,t},s}^{-d,s,n}(y_d)$.

Sampling \mathbf{c} : After assigning customers to tables and levels, we also sample path assignments for all tables. This is important since it can change the assignments of all customers sitting at a table, which leads to a well-mixed Markov chain and faster convergence. The probability of assigning table t in restaurant d to a path c is

$$P(c_{d,t} = c | \text{rest}) \propto P(c_{d,t} = c | \mathbf{c}^{-d,t}) \cdot P(\mathbf{w}_{d,t}, y_d | c_{d,t} = c, \mathbf{w}^{-d,t}, \mathbf{c}^{-d,t}, \mathbf{t}, \mathbf{z}, \boldsymbol{\eta}) \quad (5.7)$$

where we slightly abuse the notation by using $\mathbf{w}_{d,t} \equiv \cup_{\{s|t_{d,s}=t\}} \mathbf{w}_{d,s}$ to denote the set of customers in all the groups sitting at table t in restaurant d . The first factor is the prior probability of a path given all tables' path assignments $\mathbf{c}^{-d,t}$, excluding table t in restaurant d and is given in Equation 5.4.

The second factor in Equation 5.7 is the probability of observing $\mathbf{w}_{d,t}$ and y_d given the new path assignments, $P(\mathbf{w}_{d,t}, y_d | c_{d,t} = c, \mathbf{w}^{-d,t}, \mathbf{c}^{-d,t}, \mathbf{t}, \mathbf{z}, \boldsymbol{\eta}) = f_{\mathbf{c}}^{-d,t}(\mathbf{w}_{d,t})$.

$g_c^{-d,t}(y_d)$.

Optimizing $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$: We optimize the regression parameters $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ via the likelihood,

$$\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\tau}) = -\frac{1}{2\rho} \sum_{d=1}^D (y_d - \boldsymbol{\eta}^T \bar{\mathbf{z}}_d - \boldsymbol{\tau}^T \bar{\mathbf{w}}_d)^2 - \frac{1}{2\sigma} \sum_{k=1}^{K^+} (\eta_k - \mu)^2 - \frac{1}{\omega} \sum_{v=1}^V |\tau_v|, \quad (5.8)$$

where K^+ is the number of nodes currently in the tree.³ This maximization is performed using L-BFGS (Liu and Nocedal, 1989).

5.4 Data: Congress, Products, Films

In this section, we describe the three datasets that we use to evaluate SHLDA. First, we discover and analyze agendas and frames used in a collection of floor debates in the 109th U.S. Congress. Second, to show the applicability of SHLDA in other settings, we use two online review datasets: Amazon product reviews and movie reviews. For all datasets, we perform similar preprocessing step as in experiments in Chapter 4, in which we remove stopwords, add bigrams to the vocabulary, and filter the vocabulary using TF-IDF.

5.4.1 U.S. congressional floor debates:

We download the transcripts of the floor debates in the 109th U.S. Congress from GovTrack and follow the preprocessing procedure described in Thomas et al.

³The superscript $+$ is to denote that this number is unbounded and varies during the sampling process.

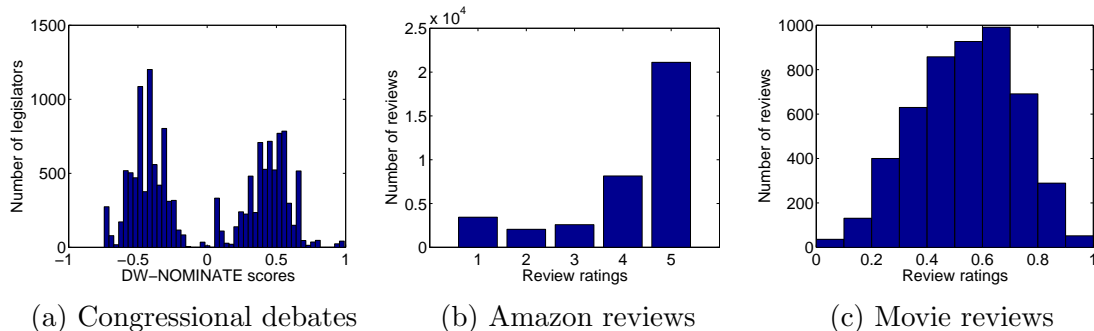


Figure 5.4: Distributions of the response variables in the three datasets.

(2006).⁴ First, we associate each page with a bill and each bill with a roll-call vote. If there is no association, the page is discarded. We then segment each page into *turns*, each of which is a continuous utterance by a legislator.⁵ Each turn is labeled with either “Yea” or “Nay” according to the voting records. Each set of turns corresponding to the same bill is then grouped into a *debate*. To consider only “interesting” debates, we keep only the ones in which at least 20% of the turns were labeled “Yea” and at least 20% were labeled “Nay”.

In using this corpus, we are interested in studying what agenda issues that legislators with different ideologies talk about and how they talk about those issues on the congressional floor. To approximate the ideological position of each legislator on a liberal-conservative spectrum, we use the first dimension of the DW-NOMINATE coordinate obtained from VoteView.⁶ Developed by (Poole and Rosenthal, 1997), DW-NOMINATE and other NOMINATE-based methods are procedures to estimate the positions of legislators, or often called *ideal point*, in a ideological space using their voting records. Although by design, DW-NOMINATE can capture the

⁴<http://www.govtrack.us/data/us/109/>

⁵A *turn* here is equivalent to a *speech segment* in Thomas et al. (2006).

⁶http://voteview.com/dwnomin_joint_house_and_senate.htm

ideal points in multi-dimensional space, only one or two dimensions are often used. VoteView provides pre-estimated two dimensional ideal points, in which the first coordinate approximately maps to positions on a liberal-conservative dimension.⁷

We then download information about legislators including their names, types (either Representative or Senator), parties, states and the ICPSR ID from GovTrack.⁸ Matching legislator records from GovTrack and VoteView is done using their ICPSR IDs. For records with missing ICPSR IDs, we manually matched using the legislator's name, type, state and party. After processing, our corpus contains 5,201 turns in the House, 3,060 turns in the Senate, and 5,000 words in the vocabulary. Figure 5.4a shows the distribution of the ideological score.⁹

5.4.2 Amazon product reviews

Our second dataset is a set of Amazon reviews on products such as computers, MP3 players, GPS devices etc, used in (Jindal and Liu, 2008; Lim et al., 2010). We focus our analysis on the most popular products by keeping only the top 50 products with the most reviews. After filtering, this corpus contains 37,349 reviews with a vocabulary of 5,000 words. We use the rating associated with each review as the response variable in our experiments. The values of these ratings range from 1 to 5 and their distribution is very skewed towards high ratings as shown in Figure 5.4b.

⁷More details about NOMINATE-based procedures and other ideal point models are discussed in Chapter 6.

⁸ICPSR IDs are identification numbers that are issued by the Inter-university Consortium for Political and Social Research for each Congressional member. The set of ICPSR IDs used in our corpus are from <http://www.voteview.com/icpsr.htm>, which have been corrected by Keith Poole and Howard Rosenthal.

⁹Data are available at <http://www.cs.umd.edu/~vietan/data/debates-109112.zip>.

5.4.3 Movie reviews

Our third corpus is a set of 5,006 reviews of movies (Pang and Lee, 2005), again using review ratings as the response variable y_d , although in this corpus the ratings are normalized to the range from 0 to 1 (Figure 5.4c). After preprocessing, the vocabulary contains 5,000 words.

5.5 Qualitative Analysis of Topic Hierarchies

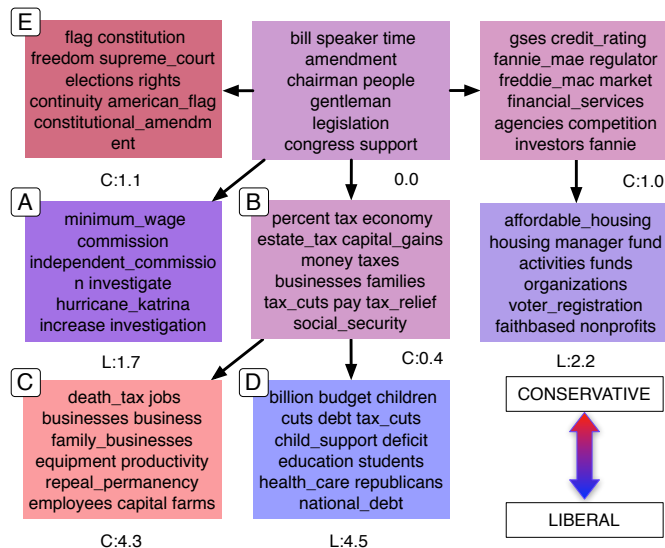


Figure 5.5: Topics discovered from Congressional floor debates. Many first-level topics are bipartisan (purple), while lower level topics are associated with specific ideologies (Democrats blue, Republicans red). For example, the “tax” topic (B) is bipartisan, but its Democratic-leaning child (D) focuses on social goals supported by taxes (“children”, “education”, “health care”), while its Republican-leaning child (C) focuses on business implications (“death tax”, “jobs”, “businesses”). The number below each topic denotes the magnitude of the learned regression parameter associated with that topic. Colors and the numbers beneath each topic show the regression parameter η associated with the topic.

We first qualitatively analyze the topic hierarchies learned by SHLDA. In Figure 5.5, a portion of the topic hierarchy induced from the Congressional debate cor-

pus, Nodes A and B illustrate agendas—issues introduced into political discourse—associated with a particular ideology: Node A focuses on the hardships of the poorer victims of hurricane Katrina and is associated with Democrats, and text associated with Node E discusses a proposed constitutional amendment to ban flag burning and is associated with Republicans. Nodes C and D, children of a neutral “tax” topic (value η is near zero), reveal how parties frame taxes as *gains* in terms of new social services (Democrats) and *losses* for job creators (Republicans).¹⁰ Although a formal coherence evaluation remains a goal for future work (Chang et al., 2009b), a qualitative look at the topic hierarchy uncovered by the model suggests that it is indeed capturing agenda/framing structure as discussed in Section 5.1

Figure 5.6 shows the topic structure discovered by SHLDA in the Amazon review corpus. Nodes at higher levels are relatively neutral, with relatively small regression parameters. These nodes have general topics with no specific polarity. However, the bottom level clearly illustrates polarized positive/negative perspective. For example, Node A concerns washbasins for infants, and has two polarized children nodes: reviewers take a positive perspective when their children enjoy the product (Node B: “loves”, “splash”, “play”) but have negative reactions when it leaks (Node C: “leak(s/ed/ing)”).

In addition to the per-topic regression parameters, SHLDA also associates each word with a lexical regression parameter τ . Table 5.2 shows the top ten words with highest and lowest τ . The results are unsurprising, although the lexical regression

¹⁰This relates to a different view of framing in term of cognitive bias (Tversky and Kahneman, 1981; Ledgerwood and Boydstun, 2014)

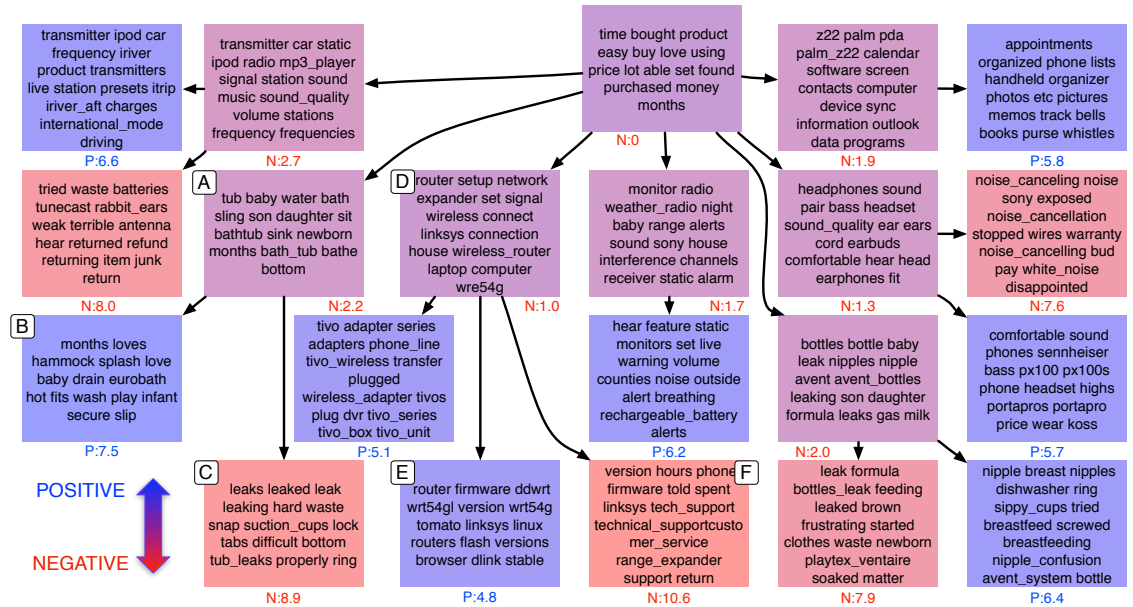


Figure 5.6: Topics discovered from Amazon reviews. Higher topics are general, while lower topics are more specific. The polarity of the review is encoded in the color: red (negative) to blue (positive). Many of the first-level topics have no specific polarity and are associated with a broad class of products such as “routers” (Node D). However, the lowest topics in the hierarchy are often polarized; one child topic of “router” focuses on upgradable firmware such as “tomato” and “ddwrt” (Node E, positive) while another focuses on poor “tech.support” and “customer.service” (Node F, negative). The number below each topic is the regression parameter learned with that topic.

for the Congressional debates is less clear-cut than other datasets. As we will see in Section 5.6, for similar datasets, SHLDA’s context-specific regression is more useful when global lexical weights do not readily differentiate documents.

5.6 Quantitative Prediction of Document Responses

For quantitative evaluation, we measure the effectiveness of SHLDA in predicting values of the response variables for unseen documents in the three datasets. For comparison we consider these baselines:

- Multiple linear regression (MLR) models the response variable as a linear

| Dataset | Top 10 words with positive weights | Top 10 words with negative weights |
|----------------|--|--|
| Debates | bringing, private_property, illegally, tax_relief, regulation, mandates, constitutional, committee_report, illegal_alien | bush_administration, strong_opposition, ranking, republicans, republican_leadership, secret, discriminate, majority, undermine |
| Amazon Reviews | highly_recommend, pleased, love, loves, perfect, easy, excellent, amazing, glad, happy | waste, returned, return, stopped, leak, junk, useless, returning, refund, terrible |
| Movie Reviews | hilarious, fast, schindler, excellent, motion_pictures, academy_award, perfect, journey, fortunately, ability | bad, unfortunately, supposed, waste, mess, worst, acceptable, awful, suppose, boring |

Table 5.2: Top words based on the global lexical regression coefficient, τ . For the floor debates, positive τ 's are Republican-leaning while negative τ 's are Democrat-leaning.

function of multiple features (or regressors). Here, we consider two types of features: topic-based features and lexically-based features. Topic-based MLR, denoted by MLR-LDA, uses the topic distributions learned by vanilla LDA as features (Blei and McAuliffe, 2007), while lexically-based MLR, denoted by MLR-VOC, uses the normalized frequencies of words in the vocabulary as features. MLR-LDA-VOC uses both features.

- Support vector regression (SVM) is a discriminative method (Joachims, 1999) that uses LDA topic distributions (SVM-LDA), word frequencies (SVM-VOC), and both (SVM-LDA-VOC) as features.¹¹
- Supervised topic model (sLDA): we implemented sLDA using Gibbs sampling. The version of sLDA we use is slightly different from the original sLDA described in (Blei and McAuliffe, 2007), in that we place a Gaussian prior $\mathcal{N}(0, 1)$ over the regression parameters to perform L2-norm regularization.¹²

¹¹<http://svmlight.joachims.org/>

¹²This performs better than unregularized sLDA in our experiments.

| Models | Floor Debates | | | | Amazon Reviews | | Movie Reviews | |
|-----------------------|----------------|------------------|----------------|------------------|----------------|------------------|----------------|------------------|
| | House-Senate | | Senate-House | | PCC \uparrow | MSE \downarrow | PCC \uparrow | MSE \downarrow |
| | PCC \uparrow | MSE \downarrow | PCC \uparrow | MSE \downarrow | | | | |
| SVM-LDA ₁₀ | 0.173 | 0.861 | 0.08 | 1.247 | 0.157 | 1.241 | 0.327 | 0.970 |
| SVM-LDA ₃₀ | 0.172 | 0.840 | 0.155 | 1.183 | 0.277 | 1.091 | 0.365 | 0.938 |
| SVM-LDA ₅₀ | 0.169 | 0.832 | 0.215 | 1.135 | 0.245 | 1.130 | 0.395 | 0.906 |
| SVM-VOC | 0.336 | 1.549 | 0.131 | 1.467 | 0.373 | 0.972 | 0.584 | 0.681 |
| SVM-LDA-VOC | 0.256 | 0.784 | 0.246 | 1.101 | 0.371 | 0.965 | 0.585 | 0.678 |
| MLR-LDA ₁₀ | 0.163 | 0.735 | 0.068 | 1.151 | 0.143 | 1.034 | 0.328 | 0.957 |
| MLR-LDA ₃₀ | 0.160 | 0.737 | 0.162 | 1.125 | 0.258 | 1.065 | 0.367 | 0.936 |
| MLR-LDA ₅₀ | 0.150 | 0.741 | 0.248 | 1.081 | 0.234 | 1.114 | 0.389 | 0.914 |
| MLR-VOC | 0.322 | 0.889 | 0.191 | 1.124 | 0.408 | 0.869 | 0.568 | 0.721 |
| MLR-LDA-VOC | 0.319 | 0.873 | 0.194 | 1.120 | 0.410 | 0.860 | 0.581 | 0.702 |
| sLDA ₁₀ | 0.154 | 0.729 | 0.090 | 1.145 | 0.270 | 1.113 | 0.383 | 0.953 |
| sLDA ₃₀ | 0.174 | 0.793 | 0.128 | 1.188 | 0.357 | 1.146 | 0.433 | 0.852 |
| sLDA ₅₀ | 0.254 | 0.897 | 0.245 | 1.184 | 0.241 | 1.939 | 0.503 | 0.772 |
| SHLDA | 0.356 | 0.753 | 0.303 | 1.076 | 0.413 | 0.891 | 0.597 | 0.673 |

Table 5.3: Regression results for Pearson’s correlation coefficient (PCC, higher is better (\uparrow)) and mean squared error (MSE, lower is better (\downarrow)). Results on Amazon product reviews and movie reviews are averaged over 5 folds. Subscripts denote the number of topics for parametric models. For SVM-LDA-VOC and MLR-LDA-VOC, only best results across $K \in \{10, 30, 50\}$ are reported. Best results are in **bold**.

For parametric models (LDA and sLDA), which require the number of topics K to be specified beforehand, we use $K \in \{10, 30, 50\}$. We use symmetric Dirichlet priors in both LDA and sLDA, initialize the Dirichlet hyperparameters to 0.5, and use slice sampling (Neal, 2003) for updating hyperparameters. For sLDA, the variance of the regression is set to 0.5. For SHLDA, we use trees with maximum depth of three. We slice sample m , π , β and γ , and fix $\mu = 0$, $\sigma = 0.5$, $\omega = 0.5$ and $\rho = 0.5$. We found that the following set of initial hyperparameters works reasonably well for all the datasets in our experiments: $m = 0.5$, $\pi = 100$, $\vec{\beta} = (1.0, 0.5, 0.25)$, $\vec{\gamma} = (1, 1)$, $\alpha = 1$. We also set the regression parameter of the root node to zero, which speeds inference (since it is associated with every document) and because it is reasonable to assume that it would not change the response variable.

To compare the performance of different methods, we compute Pearson’s correlation coefficient (PCC) and mean squared error (MSE) between the true and predicted values of the response variables and average over 5 folds. For the Congressional debate corpus, following [Yu et al. \(2008\)](#), we use documents in the House to train and test on documents in the Senate and vice versa.

Results and analysis: Table 5.3 shows the performance of all models on our three datasets. Methods that only use topic-based features such as SVM-LDA and MLR-LDA do poorly. Methods only based on lexical features like SVM-VOC and MLR-VOC outperform methods that are based only on topic features significantly for the two review datasets, but are comparable or worse on congressional debates. This suggests that reviews have more highly discriminative words than political speeches (Table 5.2). Combining topic-based and lexically-based features improves performance, which supports our choice of incorporating both per-topic and per-word regression parameters in SHLDA.

In all cases, SHLDA achieves strong performance results. For the two cases where SHLDA was second best in MSE score (Amazon reviews and House-Senate), it outperforms other methods in PCC. Doing well in PCC for these two datasets is important since achieving low MSE is relatively easier due to the response variables’ bimodal distribution in the floor debates and positively-skewed distribution in Amazon reviews. For the floor debate dataset, the results of the House-Senate experiment are generally better than those of the Senate-House experiment, which is consistent with previous results and is explained by the greater number of debates

in the House (Yu et al., 2008).

5.7 Conclusion

5.7.1 Summary

In this chapter, we go beyond studying agendas—the focus of Chapters 3 and 4 and expand our focus to discovering and analyzing framing—tackling the question of how agenda issues are talked about. We are particularly interested in understanding how policymakers with different ideologies talk about various policy issues in U.S. Congress. We present SHLDA, a supervised hierarchical nonparametric topic model, to discover a tree-structured topic hierarchy, in which top-level topics map to agenda issues and bottom-level topics represents issue-specific frames that policymakers with different ideologies.

Although motivated by studying framing in political discourse, the model we introduce, SHLDA, is applicable to a broader setting of text, which are associated with continuous responses of interest. Besides congressional floor debates with the accompanying ideological scores of the speakers, we also apply SHLDA on modeling online reviews with their associated rating scores. We show qualitatively that the topic hierarchies learned by SHLDA indeed capture the two-level agenda/framing structure in line with the theory that motivates the work. Experimental results on both political and review data show that SHLDA can improve the performance of predicting political ideologies and review ratings over commonly used baselines.

5.7.2 Discussions and Future Directions

SHLDA extends state-of-the-art hierarchical nonparametric topic models to learn topic hierarchies. Without requiring any topic labels, SHLDA enjoys a low pre-analysis cost which is similar to traditional unsupervised topic models. Qualitative analysis reveals that the discovered topic hierarchies are meaningful and can potentially provide insights to discover and study agenda-setting and framing in political text. However, due to the complex and abstract nature of framing as described in Section 5.1.2, hierarchical topics modeled by multinomial distributions over the vocabulary, which are usually represented by lists of most probable words as shown in Figures 5.5 and 5.6, still incur moderately high post-analysis cost. In this chapter, we alleviate the problem by jointly modeling the text and pre-estimated authors’ ideological scores—readily available metadata with no additional cost. However, improving the interpretability of the discovered topic hierarchies in general is still a challenging problem. In Chapter 6, we address this problem by leveraging existing labeled data to learn prior distributions for topics representing agenda issues that help discover more interpretable topic hierarchies.

In this chapter, we use a tree-structured hierarchy of topics, as the one shown in Figure 5.1, to capture *issue-specific framing*. Thanks to the flexibility of the nCRP prior, SHLDA can learn tree structures with unbounded width and depth. To discover agenda-setting and framing as second-level agenda setting, we limit the number of levels in the tree to be *three*: one root node to capture shared background topic, first-level nodes correspond to agenda issues and second-level nodes

correspond to issue-specific frames. We also use this particular three-level hierarchy in designing the model to capture framing as second-level agenda setting more explicitly in the next chapter.

In addition, each node in a tree has at most one parent. This property allows us to capture issue-specific frames in which a low-level node corresponds to a frame that is a more specific subtopic of its parent node. For example, a node with the topic on *legalizing marijuana* might have children nodes which discuss subtopics such as the cost the the drug war, the benefits of medical marijuana, and the conflicts between state-level and federal-level legal systems when marijuana is legalized only in some states. These subtopics, under the assumptions of our models, are specifically about the *legalizing marijuana* topic. Conceptually, however, one could consider the above subtopics an *economic frame*, a *health frame*, and a *legal frame* respectively, which can be shared across different topics. One example of this approach is the *Policy Frames Codebook* which defines frame categories such as “Economic”, “Capacity & Resources”, “Morality & Ethics” and “Fairness & Equality” that are applicable across multiple policy issues like abortion, immigration, and marriage equality (Boydston et al., 2013c). We consider capturing this type of framing an interesting direction for future work, which can draw upon various existing relevant computational methods such as using a directed acyclic graph (DAG) instead of a tree with PAM-like topic models (Li and McCallum, 2006; Li et al., 2007; Mimno et al., 2007) or combining additively the effects of multiple topics in the log space with SAGE (Eisenstein et al., 2011), factorial LDA (Paul and Dredze, 2012), structural topic model (Roberts et al., 2014), and SPRITE (Paul and Dredze, 2015).

Chapter 6: Discovering Agendas and Frames from Roll Call Votes and Text

6.1 Introduction

In the previous chapter, we propose a hierarchical topic model to discover agendas and frames from political text. The model learns a hierarchy of topics where top-level topics correspond to more general agenda issues and bottom-level topics correspond to issue-specific frames. To discover frames used by legislators with different ideologies, the model jointly captures the text and the first dimension of DW-NOMINATE score—a commonly used quantity estimated from voting records to approximate the position, or the *ideal point*, of lawmakers on a single liberal-conservative dimension (Poole and Rosenthal, 2007). In reality, however, people might hold different positions on different issues, which motivates work on estimating lawmakers’ ideologies using multi-dimensional ideal points.

One major disadvantage of multi-dimensional ideal point models is that the estimated dimensions are often difficult to interpretable. To mitigate this problem, recent research has introduced methods to estimate multi-dimensional ideal points using both voting data and the associated text. One popular approach is to use

topic models to discover topics from bill text, each of which maps to a dimension of the ideal point space (Gerrish and Blei, 2012; Lauderdale and Clark, 2014; Gu et al., 2014; Sim et al., 2015).

Following this approach, in this chapter, we introduce HIPTM, a *Hierarchical Ideal Point Topic Model*, which estimates multi-dimensional ideal points of legislators using their votes and speeches as well as the bill text. We improve the interpretability of ideal points' dimensions by leveraging existing labeled data from the Congressional Bills Project. Using these data, our ideal point model contains 19 dimensions, each of which corresponds to a major topic in the Policy Agendas Topics Codebook. In addition, HIPTM discovers a hierarchy of topics, which allows us to analyze both agenda issues and issue-specific frames that legislators use on the congressional floor.

We first use HIPTM as a tool for exploratory data analysis. We apply HIPTM to qualitatively analyze how Republican legislators vote and talk in the 112th U.S. Congress with respect to the *Tea Party movement*, a recent American political movement which has attracted a great deal of attention from both the public and academic scholars. We analyze (1) the difference in ideological positions on different issues of members of the Tea Party Caucus—the first institutional organization for the Tea Party movement—in comparison with other legislators with no Tea Party Caucus membership and (2) what policy agenda issues that Republican legislators pay attention to and how these issues are framed by legislators with different positions on the Tea Party. Then, we quantitatively show the effectiveness of our HIPTM model in capturing the “Tea Partiness” of legislators from their votes and

text by conducting experiments on classifying Tea Party Caucus membership.

In the remainder of this chapter, we briefly introduce ideal point models, compare one-dimensional with multi-dimensional ideal points, discuss how recent work incorporates text to improve the interpretability of multi-dimensional ideal point models, and motivate the focus of our analysis on the Tea Party in the House of Representatives.

6.1.1 A Brief Overview of Ideal Point Models

In Chapter 5, we use the DW-NOMINATE score estimated by Lewis and Poole (2004) to approximate how liberal or conservative a legislator is. Estimating positions (or preferences) of political actors in the ideological space, like the DW-NOMINATE score on the liberal-conservative spectrum, has been a fundamental component of contemporary political science research. These positions are often called ideal points and methods to estimate them are ideal point models, an application of the item response theory.

Item response theory (IRT) is a popular approach used for describing probabilistic relationship between observed responses on a set of items by a set of responders who are characterized by some continuous latent traits (Fox, 2010). Pioneered by Lord (1953) and Rasch (1961), one of an IRT's early uses is to estimate participants' scores on standardized tests. Since then, IRT has been widely applied on many problems in a wide range of research disciplines, some recent examples of which include alcohol disorder (Feske et al., 2007; Beseler et al., 2010), psychiatric

epidemiology (Tsutsumi et al., 2009) and nicotine dependence symptoms (Rose and Dierker, 2010).

| Legislator | Roll Call Number | | | |
|-------------------------|------------------|-----|-----|-----|
| | 18 | 32 | 96 | 149 |
| David McKinley (R-WV) | Yea | Yea | Yea | Nay |
| Jeff Flake (R-AZ) | Yea | Yea | Yea | Yea |
| Justin Amash (R-MI) | Yea | Yea | Yea | Yea |
| Timothy Bishop (D-NY) | Yea | - | Yea | Yea |
| Robert Andrews (D-NJ) | Nay | Nay | Yea | Nay |
| Suzanne Bonamici (D-OR) | - | - | Yea | Nay |

Table 6.1: Example voting records of legislators in the 112th House of Representatives. A legislator might not vote on a bill, which is denoted by ‘-’ in this table.

In political science, *ideal point models* are IRT models which estimate political preferences, called *ideal points*, of lawmakers (i.e., the latent traits) from binary data such as legislative votes or judicial decisions (i.e., the observed responses). Figure 6.1 shows an example of a set of roll call votes in the 112th U.S. Congress. A popular formulation of *one-dimensional ideal point models*, based on which we develop our model in this chapter, posits an *ideal point* $u_a \in \mathbb{R}$ for each lawmaker u , a *polarity* $x_b \in \mathbb{R}$ and a *popularity* $y_b \in \mathbb{R}$ for each bill b (Martin and Quinn, 2002; Bafumi et al., 2005; Gerrish and Blei, 2011). The probability that lawmaker u votes “Yes” on bill b is

$$p(v_{a,b} = \text{Yes} \mid u_a, x_b, y_b) = \Phi(u_a x_b + y_b) \quad (6.1)$$

where $\Phi(\alpha) = \exp(\alpha)/(1 + \exp(\alpha))$ is the logistic (or inverse-logit) function.¹ Intuitively, most lawmakers will vote “Yes” on bills with high popularity y_b and vote “No” on bills with low y_b . When the popularity of a bill is near zero, the outcome of

¹A probit function is also often used where $\Phi(\alpha)$ is instead the cumulative distribution function of a Gaussian distribution (Martin and Quinn, 2002).

$v_{a,b}$ depends on the interaction between the lawmaker’s ideal point u_a and the bill’s polarity x_b . Under this setting, a legislator with ideal point u_a will more likely to vote “Yes” on a bill b if $u_a > -y_b/x_b$.

In one-dimensional ideal point models, if two legislators have similar ideal points, they will vote similarly on every bill. To capture the intuition that people might hold different positions on different policy issues, various works have been done to extend these models to multi-dimensional space. In *multi-dimensional ideal point models*, the ideal point of each legislator is no longer characterized by a scalar, but a multi-dimensional vector $\mathbf{u}_a \in \mathbb{R}^K$ (Heckman and Jr., 1997; Jackman, 2001; Clinton et al., 2004). Extending from Equation 6.1, the probability that lawmaker u votes “Yes” on bill b in multi-dimensional ideal point models is

$$p(v_{a,b} = \text{Yes} \mid \mathbf{u}_a, \mathbf{x}_b, y_b) = \Phi(\mathbf{u}_a^T \mathbf{x}_b + y_b) \equiv \Phi\left(\sum_{k=1}^K u_{a,k} x_{b,k} + y_b\right) \quad (6.2)$$

6.1.2 On the Dimensionality of Ideal Points

Arguably, one of the most influential methods for ideal point estimation in political science is the *nominal three-step estimation* procedure, or more commonly known as NOMINATE, by Poole and Rosenthal (1997). Poole and Rosenthal (1985, 1987) develop the original NOMINATE model which estimates one-dimensional ideal points of legislators in U.S. Congress. Subsequently, different variants of NOMINATE are developed: D-NOMINATE (Poole and Rosenthal, 1991), W-NOMINATE (Poole and Rosenthal, 1997), Common Space NOMINATE (Poole, 1998), and DW-NOMINATE—

the latest version which we use in developing SHLDA in Chapter 5 (McCarty et al., 1997). More details about the development of these models can be found in Poole and Rosenthal (1997), Poole and Rosenthal (2001) and Carroll et al. (2009).^{2,3}

Although all these extensions of NOMINATE are multi-dimensional by design, in practice, only one or two dimensions are often used. Analyzing roll call votes in the U.S. Congress from 1959–1980, Poole and Daniels (1985) report that “a single liberal-conservative dimension accounts for more than 80% of the variance in the ratings. A second dimension, associated with party unity, accounts for 7% of the variance.” Similarly, Grofman and Brazill (2002) observe a “fundamental unidimensionality in the data on Supreme Court voting patterns 1951–1993”. Hix et al. (2006) analyze two dimensions of politics in the European Parliament: the main dimension is “the classic left-right dimension found in domestic politics”, while the second dimension captures “government-opposition conflicts as well as national and European party positions on European integration”.

Despite the successes of ideal point models using one or two dimensions in fitting the observed roll call data statistically, many scholars have debated and argued for using higher dimensional models (Koford, 1989; Wilcox and Clausen, 1991; Poole et al., 1991; Snyder Jr, 1992; Carmines and D’Amico, 2015). For example, by splitting up the roll call data into “subsets of relatively homogeneous subject matter” to analyze ideal points, Crespín and Rohde (2010) find that “voting is multidimensional and members do not vote in a consistent ideological fashion across

²Descriptions of different NOMINATE models are also described at <http://www.voteview.com/page2a.htm>.

³Estimated ideal points with other related data can be found at <http://www.voteview.com/>.

all issue areas”. [Lauderdale and Clark \(2014\)](#) provide two main obstacles which have prevented scholars from moving beyond two dimensions. First, when moving to more than two dimensions, the dimensions are less politically interpretable. Second, binary data most commonly used such as roll call votes or judicial decisions are “insufficiently informative to support analyses beyond one or two dimensions using multidimensional scaling methods”. To overcome these difficulties, recent work has proposed multi-dimensional ideal point models to jointly capture both the binary votes and the associated text.

6.1.3 Scaling Multi-dimensional Ideal Points using Votes and Text

As discussed in previous chapters of this thesis, topic models such as LDA and many of its extensions provide us a useful set of tools to extract thematic structure from a large collection of documents. Applying topic models to political text, we can extract topics, each of which is a multinomial distribution over words and can map to a political agenda issue. By modeling both the votes and the text jointly, recent work discovers multi-dimensional ideal points, each of which is associated with a topic.

For example, [Gerrish and Blei \(2012\)](#) introduce the *issue-adjusted ideal point* model, which posits that each legislator a is characterized by a *base* ideal point $u_a \in \mathbb{R}$ and an *issue-adjusted* vector $z_a \in \mathbb{R}^K$. The model captures both the votes and the text associated with the bills and define the probability of lawmaker a voting

“Yes” on bill b as

$$p(v_{a,b} = \text{Yes} \mid \mathbf{u}_a, \mathbf{z}_a, x_b, y_b, \mathbf{w}_b) = \Phi \left(\left(\sum_{k=1}^K z_{a,k} \theta_{b,k} + u_a \right) x_b + y_b \right) \quad (6.3)$$

where θ_b denotes the topic proportion of bill b estimated from its text \mathbf{w}_b . Also extending the multi-dimensional model described in Equation 6.2, Gu et al. (2014) introduce the *topic-factorized ideal point* model (TF-IPM), which defines

$$p(v_{a,b} = \text{Yes} \mid \mathbf{u}_a, \mathbf{x}_b, y_b, \mathbf{w}_b) = \Phi \left(\sum_{k=1}^K \theta_{b,k} u_{a,k} x_{b,k} + y_b \right) \quad (6.4)$$

where again θ_b denotes the estimated topic proportion of bill b .

Following Jackman (2001) who uses a probit model instead, Lauderdale and Clark (2014) define

$$v_{a,b} = \begin{cases} \text{Yes,} & \text{if } v'_{a,b} \geq 0; \\ \text{No,} & \text{if } v'_{a,b} < 0. \end{cases} \quad \text{where } v'_{a,b} \sim \mathcal{N} \left(x_b \sum_{k=1}^K \theta_{b,k} u_{a,k} + y_b \right) \quad (6.5)$$

In this model, a bill b has a one-dimensional polarity x_b as in Equation 6.1. The quantity $\sum_{k=1}^K \theta_{b,k} u_{a,k}$ can be seen as a *vote-specific ideal point* of voter u_a on bill b , which essentially is the average of voter a 's K -dimensional ideal points \mathbf{u}_a weighted by bill b 's estimated topic proportions θ_b . Sim et al. (2015) use a similar framework studying judicial decisions of the U.S. Supreme Court to incorporate text authored by amici curiae (“friends of the court” separate from the litigants) who seek to weigh in on the decision.

6.1.4 Tea Party in the House

The recent rise of the *Tea Party* in U.S. politics, with its complex political views, provides an excellent case to study multi-dimensional ideologies. The *Tea Party movement* is an American political movement which has attracted much recent attention from both the media and social science scholars. The movement burst into the public attention after President Barack Obama passed the American Recovery and Reinvestment Act of 2009 (ARRA), more commonly known as the “Stimulus bill” in February 2009 to address the fiscal crisis by providing temporary relief programs in infrastructure, education, health, federal tax etc. Widely attributed as “amorphous, grassroots, bottom-up, anti-government” (Gervais and Morris, 2012), the Tea Party movement began as a series of small rallies centered on demands for limited government and lower taxes, and gradually formed a mix of grassroots networks assembled by local organizers and national organizations such as the Tea Party Express, the Tea Party Patriots and Freedom Works (Williamson et al., 2011).

Within a short period of time, the Tea Party movement had created a huge amount of energy vilifying President Obama and congressional Democrats, and protesting their agendas on various issues including economic, environmental, and health care. In July 2010, taking advantage of the energy, the *Tea Party Caucus*, the first institutional organization for the Tea Party movement, was founded by Representative Michele Bachmann (R-MN). The caucus immediately attracted several dozen legislators, all from the Republican Party, and has become increasingly prominent on the U.S. political scene.

Despite being the focus of much recent public attention, understanding and explaining the Tea Party and its ideologies are still extremely challenging, as [Carmines and D’Amico \(2015\)](#) observe: “Conventional views of ideology as a single-dimensional, left-right spectrum experience great difficulty in understanding or explaining the Tea Party.” Recent research argues that the Tea Party is just a rebranding of Republicanism ([Williamson et al., 2011](#); [Skocpol and Williamson, 2012](#)). They also show that, while Tea Partiers are notable for their opposition against the Affordable Care Act, or ObamaCare, they are supporters of long-standing federal social programs like Social Security, Medicare, and generous benefits for military veterans. Some have suggested that the Tea Party is primarily based on the opposition to President Obama and racial resentment ([Barreto et al., 2011](#); [Maxwell and Parent, 2012](#)), while others argue that the Tea Party is mainly a religious movement ([Clement and Green, 2011](#)).

6.1.5 Main Contributions

In this chapter, following the multi-dimensional ideal point approach, we introduce HIPTM—a *Hierarchical Ideal Point Topic Model*—to jointly capture the votes and the text associated with *both* legislators and bills. The main contributions of our work in this chapter includes

- By using the votes and the associated text from both legislators and bills, our model jointly discovers (1) a hierarchy of topics which captures agenda issues and issue-specific frames, and (2) ideal points in multiple dimensional space.

- By leveraging existing labeled data from the Congressional Bills Project, we discover ideal points in multiple interpretable dimensions, each of which maps to an issue in the widely used Policy Agendas Topics Codebook. While previous multi-dimensional ideal point models learn a topic—a distribution over words—for each dimension, we also associate each topic with a predefined label to improve the interpretability and reduce post-analysis cost.
- By using a two-level hierarchical structure to model legislators’ speeches, we go beyond discovering topics associated with agenda issues and also learn issue-specific frames, each of which has a position on the corresponding ideal point dimension. The discovered frames not only help analyze the framing behaviors of legislators on different issues, but also provide a way to estimate multi-dimensional ideal points of new legislators using their text only.

Figure 6.1 illustrates an overview of the outputs expected from our model HIPTM.

We present the model in detail in the next section and describe the inference algorithm that we use to infer the posteriors over latent variables from observed data in Section 6.3. In Section 6.4, we describe the data that we collect before presenting our analysis on the different positions on various policy agenda issues of members and non-members of the Tea Party Caucus using their voting records on the collection of key votes selected by Freedom Works. We analyze the topic hierarchy discovered by HIPTM to understand what legislators pay attentions to and how they frame different issues in Section 6.5. In Section 6.6, we quantitatively evaluate the effectiveness of our model in capturing the “Tea Partiness” of legislators

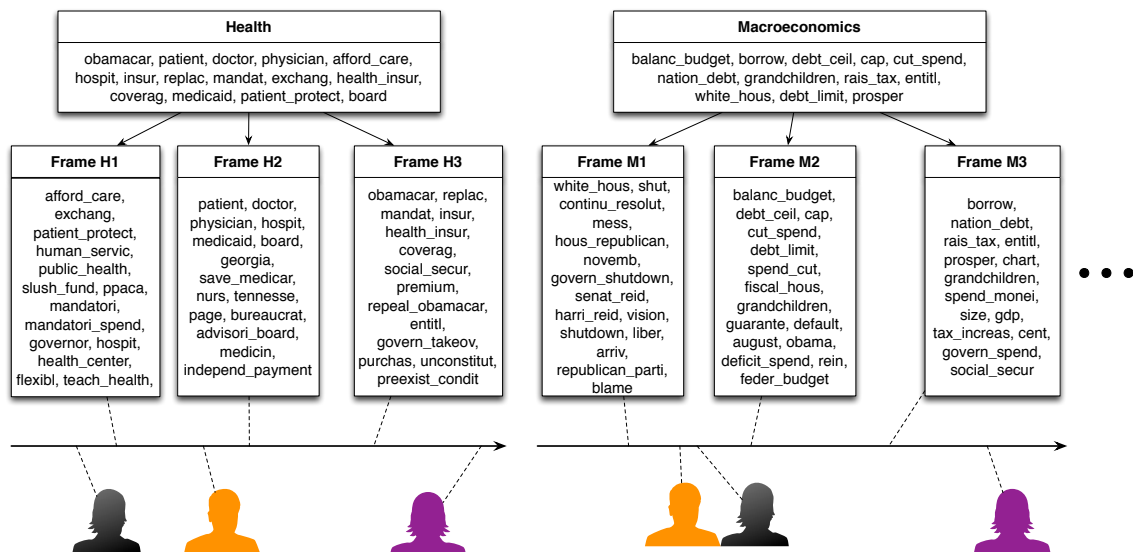


Figure 6.1: Overview of HIPTM’s outputs: (1) first-level nodes map to policy issues, each of which corresponds to a major topic in the Policy Agendas Topics codebook, (2) second-level nodes map to issue-specific frames, and (3) each frame node and each lawmaker are associated with an issue-specific ideological position.

via experiments on classifying whether they are members of the Tea Party Caucus—an institutional organization of the Tea Party movement. Section 6.7 concludes the chapter with a summary and some directions for future work.

6.2 Hierarchical Ideal Point Topic Model

Our model takes as input a collection of votes $\{v_{a,b}\}$, each of which is a binary response of voter $a \in [1, A]$ on item $b \in [1, B]$. In addition to the votes, the data also contain two different sets of text: (1) a collection of D documents $\{\mathbf{w}_d\}$, each of which is authored by a voter a_d , and (2) a collection of B documents $\{\mathbf{w}'_b\}$, each of which describes an item b . More specifically in the legislative context, there are A legislators voting on B bills. The text content of bill b is \mathbf{w}'_b and \mathbf{w}_d denotes a speech that legislator a_d gives on the congressional floor. Note that even

though congressional speeches are usually specific about a certain bill or a collection of related bills, in general we are not making any assumptions about the mapping between \mathbf{w}_d and \mathbf{w}'_b . This allows our model to be applicable to more general settings where \mathbf{w}_d can be text authored by legislator a_d that is obtained from various other sources such as blogs, social media, press releases etc. Figure 6.2 shows the plate notation diagram of the HIPTM, which has the following generative process:

1. For each issue $k \in [1, K]$
 - (a) Draw a global distribution over frames $\psi_k \sim \text{GEM}(\lambda_0)$
 - (b) Draw a topic $\phi_k \sim \text{Dirichlet}(\beta, \phi_k^*)$
 - (c) For each frame $j \in [1, \infty)$
 - i. Draw a topic $\phi_{k,j} \sim \text{Dirichlet}(\beta, \phi_k)$
 - ii. Draw $\eta_{k,j} \sim \mathcal{N}(0, \gamma)$
2. For each document $d \in [1, D]$
 - (a) Draw a topic proportion $\theta_d \sim \text{Dirichlet}(\alpha)$
 - (b) For each issue $k \in [1, K]$, draw a distribution over frames $\psi_{d,k} \sim \text{DP}(\lambda, \psi_k)$
 - (c) For each token $n \in [1, N_d]$
 - Draw an issue $z_{d,n} \sim \text{Multinomial}(\theta_d)$
 - Draw a frame given the issue $t_{d,n} \sim \text{Multinomial}(\psi_{d,z_{d,n}})$
 - Draw a word type $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}, t_{d,n}})$
3. For each voter $a \in [1, A]$ on each issue $k \in [1, K]$
 - Draw issue-specific ideal point $u_{a,k} \sim \mathcal{N}(\sum_{j=1}^{J_k} \hat{\psi}_{a,k,j} \eta_{k,j}, \rho)$

4. For each bill $b \in [1, B]$
 - Draw a polarity $x_b \sim \mathcal{N}(0, \sigma)$ and a popularity $y_b \sim \mathcal{N}(0, \sigma)$
 - Draw topic proportion $\vartheta_b \sim \text{Dirichlet}(\alpha)$
 - For each token $m \in [1, M_b]$
 - Draw an issue $z'_{b,m} \sim \text{Multinomial}(\vartheta_b)$
 - Draw a word type $w'_{b,m} \sim \text{Multinomial}(\phi_{z'_{b,m}})$
5. For each vote $v_{a,b}$ of voter a on bill b
 - $p(v_{a,b} = \text{Yes} \mid \mathbf{u}_a, x_b, y_b, \hat{\vartheta}_b) = \Phi \left(x_b \sum_{k=1}^K \hat{\vartheta}_{b,k} u_{a,k} + y_b \right)$

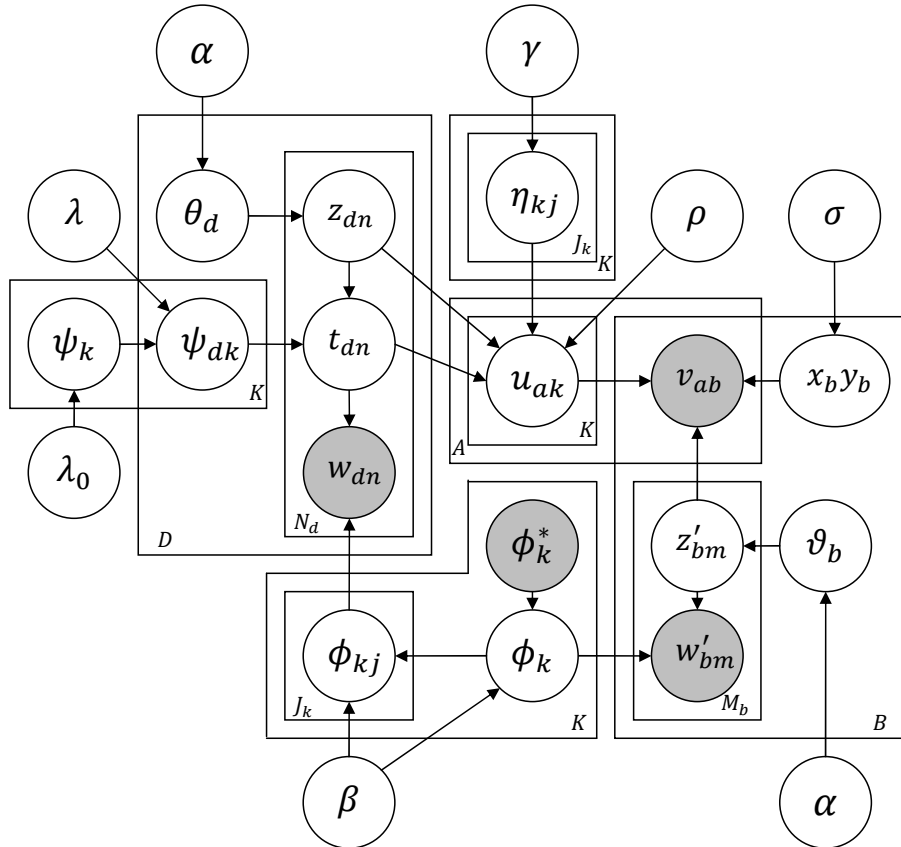


Figure 6.2: Plate notation diagram of our HIPTM model.

6.2.1 Defining the Topic Hierarchy

With the goal of analyzing agendas and frames in mind, we specifically design our topic hierarchy with two levels of nodes: (1) *issue nodes* at the first level to capture different agenda issues, and (2) *frame nodes* at the second level to capture issue-specific frames. More specifically, there are K issue nodes, each having a topic ϕ_k drawn from a Dirichlet distribution with concentration parameter α and a prior mean vector ϕ_k^* , i.e., $\phi_k \sim \text{Dirichlet}(\beta, \phi_k^*)$.

To improve the interpretability of the topics, we use $K = 19$ issue nodes, where each of which maps to a major topic in the Policy Agendas Topics Codebook. To ensure the mapping, we estimate prior distributions $\{\phi_k^*\}$ using labeled data from the Congressional Bills Project. As discussed in Chapter 4, the Congressional Bills Project provides a large collection of labeled congressional bill text, each tagged with a major topic from the Policy Agendas Topics Codebook. We obtain these labeled data and represent each labeled document as a term-frequency vector. Then, we obtain the prior distribution ϕ_k^* for an issue k by averaging the vectors of all bills labeled with k and normalizing the averaged vector.⁴ Table 6.2 shows words with highest weights for each prior vector.

Similar to the way L2H captures the relationship between topic of a node and its parent’s topic in Chapter 4, the topic $\phi_{k,j}$ at each frame node is drawn from a Dirichlet distribution whose mean is the topic ϕ_k at the corresponding issue node, i.e., $\phi_{k,j} \sim \text{Dirichlet}(\beta, \phi_k)$. In our topic hierarchy, the number of issue nodes is

⁴Labels are obtained from the Congressional Bills Project and the bill text are from the Library of Congress as in Chapter 4.

| |
|---|
| Agriculture: food; agriculture; loan; farm; crop; dairy; rural; conserve; commodity; eligible; farmer; margin; milk; contract; nutrition; livestock; plant |
| Banking, Finance, and Domestic Commerce: insure; bank; patent; mortgage; loan; commission; issuer; director; fee; application; contract; transaction; property; internet; flood_insurance; code; file |
| Civil Rights, Minority Issues, and Civil Liberties: vote; entity; abortion; violation; ballot; civil; attorney; employment; commission; discrimination; cybersecurity; disclosure; notice; privacy; breach; notification; data |
| Community Development and Housing Issues: mortgage; reside; loan; eligible; property; enterprise; qualification; income; resident; urban_development; foreclosure; neighborhood; rehablity; homeless; rental |
| Defense: unit; army_force; transfer; army; contract; acquisition; subsection; air_force; homeland_security; nuclear; personnel; public_law; nation_defense; navy; command |
| Education: student; local_education; teacher; eligible; academic; elementary; instruct; assess; literacy; parent; secondary_education; award; evaluation; grade; english; teach |
| Energy: oil; electricity; fuel; shelf; outer_continent; pipeline; facility; environment; qualification; re-new_energy; energy_efficiency; interior; energy_policy; exploration |
| Environment: chemical; substance; fish; conservatory; fishery; marine; coastal; ecosystem; habitat; species; discharge; environment_protection; region; council; gulf_coast; waste; pollution_control; treatment; environment |
| Foreign Trade: unit; duty; tariff; harmony; schedule; suspend; date; suspense; enter; consumption; effect_date; assembly; temporary; session; insert; chapter; trade |
| Government Operations: commission; unit; postal_service; code; execute; transfer; coin; candidate; domestic; official; contract; postal; expend; vote; salary; inspector; partner |
| Health: drug; medicine; coverage; disease; public_health; hospital; social_security; health_insurance; patient; application; treatment; payment; physician; nurse; clinic |
| International Affairs and Foreign Aid: internal; foreign; iran; sanction; human; syria; export; congression; bank; israel; democracy; freedom; diplomat; foreign_affair; financial_institution; army; violate; official |
| Labor, Employment, and Immigration: employment; immigration; labor; paragraph; eligible; status; compensation; application; wage; homeland_security; unemployment; board; violation; file; perform; mine |
| Law, Crime, and Family Issues: attorney; criminal; offense; child; sexual_assault; crime; domestic_violence; firearm; court; abuse; offend; law_enforcement; violate; traffick; prison; investigation; justice; gang |
| Macroeconomics: internal_revenu; income; property; qualify; corporation; treat; calendar; december; deductible; partnership; effect_date; excess; extension; income_tax |
| Public Lands and Water Management: indian; river; indian_tribe; tribe; tribal; acre; interior; map; federal_land; national_forest; boundary; property; country; native_hawaiian; bureau; recreation; trust; creek; park |
| Social Welfare: social_security; disable; eligible; payment; social; food; nutrition; insurance; employment; income; poverty; earn; calendar |
| Space, Science, Technology and Communications: spectrum; cybersecurity; director; public_safety; network; internet; broadband; critical_infrastructure; mobile; federal_agency; cyber; license; disclosure; band; computer |
| Transportation: transport; highway; motor_vehicle; metropolitan; airport; freight; rail; carrier; chapter; driver; motor; october; traffic; paragraph; surface_transport |

Table 6.2: Words with highest weights in the priors ϕ_k^* for 19 Policy Agendas Topics, estimated by using labeled data from the Congressional Bills Project.

fixed to leverage existing resources from research on policy agendas. Given each issue, the number of frames is unbounded. In particular, for each issue k , we draw a global distribution over an infinite number of frames using a stick breaking process $\psi_{0,k} \sim \text{GEM}(\lambda_0)$, which we reviewed in Chapter 2.

We also associate each frame node with an ideal point $\eta_{k,j} \sim \mathcal{N}(0, \gamma)$ which captures the position of the given frame on the issue-specific ideal point dimension. This is similar to the conventional way that supervised topic models such as

sLDA (Blei and McAuliffe, 2007) and SHLDA (Nguyen et al., 2013c, Chapter 5) discover topics that are polarized on the spectrum of the associated response variable. Unlike SHLDA, in which each node in the unbounded hierarchy has one regression parameter, in HIPTM we assume a two-level hierarchy to specifically capture the two-level model of agenda-setting described in Chapter 1. In this hierarchical structure, first-level nodes map to agenda issues, which we treat as non-polarized, and second-level nodes map to issue-specific frames, which we assume polarize on the issue-specific dimension.

6.2.2 Generating Congressional Speeches

As mentioned above, one of our model’s goals is to study how legislators *frame* various policy agenda issues on the congressional floor. To achieve that, we analyze congressional speeches $\{\mathbf{w}_d\}$, each of which is delivered by a legislator a_d . To generate each token $w_{d,n}$ of a speech d , legislator a_d will (1) first choose an issue $z_{d,n}$ from a document-specific multinomial distribution θ_d , (2) then choose a frame $t_{d,n}$ from the set of infinitely many possible frames of the given issue $z_{d,n}$ using the frame proportion $\psi_{d,k}$ drawn from a Dirichlet process, and (3) finally choose a word type from the chosen frame’s topic $\phi_{z_{d,n},t_{d,n}}$.

Like LDA, to define the prior distribution for the topic proportion θ_d of each speech d , we use a symmetric Dirichlet distribution $\theta_d \sim \text{Dirichlet}(\alpha)$. For each issue $k \in [1, K]$, the document-specific distribution over frames $\psi_{d,k}$ is distributed according to a Dirichlet process $\text{DP}(\lambda, \psi_k)$ with λ as the concentration parameter

and the issue-specific global distribution ψ_k as the base distribution. In other words, our model generates text in the speeches using a mixture of K HDPs (Teh et al., 2006). If we abandon the labeled data from the Congressional Bills Project to obtain the prior means ϕ_k^* for the 19 topics, it is relatively straightforward to extend to a fully nonparametric model where K is unbounded and can change to fit the data, like nCRF by Ahmed et al. (2013a) and nHDP by Paisley et al. (2014).

6.2.3 Generating Bill Text

The bill text provides information about the policy agenda issues that each bill is about. We use standard unsupervised topic model LDA to model the bill text $\{\mathbf{w}'_b\}$. Each bill b is a mixture ϑ_b over K issues, which is again drawn from a symmetric Dirichlet prior, i.e., $\vartheta_b \sim \text{Dirichlet}(\alpha)$. Each token $w'_{b,m}$ in bill b is generated by first choosing a topic $z'_{b,m} \sim \text{Multinomial}(\vartheta_b)$, and then choosing a word type $w'_{b,m} \sim \text{Multinomial}(\phi_{z'_{b,m}})$, just like LDA’s generative process.

6.2.4 Generating Roll Call Votes

Following recent work on multi-dimensional ideal points described in Section 6.1.3 (Lauderdale and Clark, 2014; Sim et al., 2015), we define the probability of legislator a voting “Yes” on bill b as

$$p(v_{a,b} = \text{Yes} \mid \mathbf{u}_a, x_b, y_b, \hat{\vartheta}_b) = \Phi \left(x_b \sum_{k=1}^K \hat{\vartheta}_{b,k} u_{a,k} + y_b \right) \quad (6.6)$$

where $\hat{\vartheta}_b$ is the empirical distribution of bill b over the K issues and is defined as $\hat{\vartheta}_{b,k} = \frac{M_{b,k}}{M_{b,\cdot}}$. Here, $M_{b,k}$ is the number of times in which tokens in b are assigned to issue k and $M_{b,\cdot}$ is the marginal count, i.e., the number of tokens in bill b .

We use $u_{a,k}$ to denote the ideal point of legislator a specifically on issue k . To capture the relationship between how legislator a talks about issue k and their issue-specific ideal point $u_{a,k}$, we define

$$u_{a,k} \sim \mathcal{N}(\hat{\psi}_{a,k}^T \boldsymbol{\eta}_k, \rho) \equiv \mathcal{N}\left(\sum_{j=1}^{J_k} \hat{\psi}_{a,k,j} \eta_{k,j}, \rho\right) \quad (6.7)$$

where J_k is the number of frames for topic k , which is unbounded. The mean of the Gaussian distribution is a linear combination of the ideal points $\{\eta_{k,j}\}$ of all issue k 's frames, weighted by how much time legislator a spends on each frame when talking about issue k , i.e., $\psi_{a,k,j} = \frac{N_{a,k,j}}{N_{a,k,\cdot}}$. Here, $N_{a,k,j}$ is the number of tokens authored by a that are assigned to frame j of issue k , and $N_{a,k,\cdot}$ is the marginal count. When $N_{a,k,\cdot} = 0$, which means that legislator a does not talk about issue k , we back off to an uninformed mean of 0.

Equation 6.7 represents a similar but more complex way than traditional supervised topic model (sLDA) to link the topics with the response, in that the response $u_{a,k}$ here is latent. It is similar to how [Gerrish and Blei \(2011\)](#) use the bill text to regress on the bill's latent polarity x_b and popularity y_b , which are then used for modeling the votes downstream. In this chapter, we only use text from congressional speeches for regression, since we mainly focus on studying agendas and frames. Incorporating the bill text into the regression as well is an interesting direction for

future work.

6.3 Posterior Inference

Given observed data which consist of (1) a set of legislative votes $\{v_{a,b}\}$ by A legislators on B bills, (2) a collection of congressional speeches $\{\mathbf{w}_d\}$, each of which is given by a legislator a_d , and (3) the bill text $\{\mathbf{w}'_b\}$, we estimate the posterior distributions over the latent variables in our model using a stochastic EM inference algorithm, similar to Chapter 5. We alternate between (1) sampling the issue assignments $\{z'_{b,m}\}$ for tokens in the bill text, (2) sampling the issue assignments $\{z_{d,n}\}$ and frame assignments $\{t_{d,n}\}$ for tokens in the speeches, (3) sampling the topics at first-level issue nodes $\{\phi_k\}$, (4) sampling the global frame proportion $\{\psi_k\}$ for all issues, (5) optimizing frames' regression parameters $\{\eta_{k,j}\}$ using L-BFGS (Liu and Nocedal, 1989), and (6) updating the legislators' multi-dimensional ideal points $\{u_{a,k}\}$ and the bills' polarity $\{x_b\}$ and popularity $\{y_b\}$ using gradient ascent.

6.3.1 Sampling Issue Assignments for Bill Tokens

The probability of assigning a token $w'_{b,m}$ in the bill text to an issue k is

$$p(z'_{b,m} = k \mid \text{rest}) \propto \frac{M_{b,k}^{-b,m} + \alpha}{M_{b,\cdot}^{-b,m} + K\alpha} \cdot \hat{\phi}_{k,w'_{b,m}} \quad (6.8)$$

where $M_{b,k}$ denotes the number of tokens in bill text b that are assigned to issue k . The current estimated probability of word type v given issue k is denoted by $\hat{\phi}_{k,v}$, which we update during the inference as described in Section 6.3.3. Marginal counts

are denoted by \cdot and the superscript $^{-b,m}$ denotes the exclusion of the assignment for token $w'_{b,m}$ from the corresponding count.

6.3.2 Sampling Frame Assignments for Speech Tokens

To sample the assignments for tokens in the speeches, we first sample an issue using the following sampling equation

$$p(z_{d,n} = k \mid \text{rest}) \propto \frac{N_{d,k}^{-d,n} + \alpha}{N_{d,\cdot}^{-d,n} + K\alpha} \cdot \hat{\phi}_{k,w_{d,n}} \quad (6.9)$$

where $N_{d,k}$ similarly denotes the number of times that tokens in d are assigned to issue k . Given the sampled issue k , we sample the frame as $p(t_{d,n} = j \mid z_{d,n} = k, a_d = a, \text{rest}) \propto$

$$\begin{cases} \mathcal{N}(u_{a,k}; \mu_{a,k,j}, \rho) \cdot \left(\frac{N_{d,k,j}^{-d,n}}{N_{d,k,j}^{-d,n} + \lambda} + \frac{\lambda}{N_{d,k,j}^{-d,n} + \lambda} \cdot \hat{\psi}_{k,j} \right), & \text{if } j \text{ exists;} \\ \mathcal{N}(u_{a,k}; \mu_{a,k,j^{\text{new}}}, \rho) \cdot \frac{\lambda}{N_{d,k,j}^{-d,n} + \lambda} \cdot \hat{\psi}_{k,j^{\text{new}}}, & \text{if } j^{\text{new}} \text{ is new.} \end{cases} \quad (6.10)$$

where $\mu_{a,k,j} = (\sum_{j'=1}^{J_k} \eta_{k,j'} N_{d,k,j'}^{-d,n} + \eta_{k,j}) / N_{d,k,\cdot}$ for an existing frame j , and for a newly created frame j^{new} , we have $\mu_{a,k,j^{\text{new}}} = (\sum_{j'=1}^{J_k} \eta_{k,j'} N_{d,k,j'}^{-d,n} + \eta_{k,j^{\text{new}}}) / N_{d,k,\cdot}$, where $\eta_{k,j^{\text{new}}}$ is drawn from the Gaussian prior $\mathcal{N}(0, \gamma)$. Here, the estimated global probability of choosing a frame j of issue k is $\hat{\psi}_{k,j}$. We describe how we update this probability during inference in Section 6.3.4.

6.3.3 Sampling Issue Topics

In the generative process of HIPTM, the topic ϕ_k of issue k is used both (1) for generating tokens in the bill text and (2) as the mean of the Dirichlet priors generating topics of this issue’s frames. Following [Ahmed et al. \(2013a\)](#), we sample this distribution using

$$\hat{\phi}_k \sim \text{Dir}(\mathbf{m}_k + \tilde{\mathbf{n}}_k + \beta\phi_k^*) \quad (6.11)$$

where $\mathbf{m}_k \equiv (M_{k,1}, M_{k,2}, \dots, M_{k,V})$ is the actual count vector of tokens from the bill text assigned to each issue. The vector $\tilde{\mathbf{n}}_k \equiv (\tilde{N}_{k,1}, \tilde{N}_{k,2}, \dots, \tilde{N}_{k,V})$ denotes the token counts propagated from words assigned to topics that are associated with frames of issue k , which can be approximated effectively using either the minimal or maximal path assumptions ([Cowans, 2006](#); [Wallach, 2008](#); [Ahmed et al., 2013a](#)).

6.3.4 Sampling Frame Proportions

Following the *direct assignment* method described in [Teh et al. \(2006\)](#), we sample the global frame proportion as

$$\hat{\psi}_k \equiv (\hat{\psi}_{k,1}, \hat{\psi}_{k,2}, \dots, \hat{\psi}_{k,j^{\text{new}}}) \sim \text{Dir}(\hat{N}_{\cdot,k,1}, \hat{N}_{\cdot,k,2}, \dots, \hat{N}_{\cdot,k,J_k}, \lambda_0) \quad (6.12)$$

where $\hat{N}_{\cdot,k,j} = \sum_{d=1}^D \hat{N}_{d,k,j}$ and $\hat{N}_{d,k,j}$ can be sampled effectively using the Antoniak distribution ([Antoniak, 1974](#)). More details can be found in [Teh et al. \(2006, page 1574\)](#) or [Ahmed et al. \(2013a, Appendix\)](#).

6.3.5 Optimizing Frame Regression Parameters

We update the regression parameters $\boldsymbol{\eta}_k$ of frames under issue k by using L-BFGS (Liu and Nocedal, 1989) to optimize the following log likelihood

$$\mathcal{L}(\boldsymbol{\eta}_k) = -\frac{1}{2\rho} \sum_{a=1}^A (u_{a,k} - \boldsymbol{\eta}_k^T \hat{\boldsymbol{\psi}}_{a,k}) - \frac{1}{2\gamma} \sum_{j=1}^{J_k} \eta_{k,j}^2 \quad (6.13)$$

6.3.6 Updating Ideal Points, Polarity and Popularity

We update the multi-dimensional ideal point \mathbf{u}_a of each legislator a and the polarity x_b and popularity y_b of each bill b by optimizing the following log likelihood using gradient ascent.

$$\begin{aligned} \mathcal{L}(\mathbf{u}, \mathbf{x}, \mathbf{y}) &= \sum_{a=1}^A \sum_{b=1}^B v_{a,b} \log p(v_{a,n} = 1) + (1 - v_{a,b}) \log p(v_{a,b} = 0) \\ &\quad - \frac{1}{2\rho} \sum_{a=1}^A \sum_{k=1}^K (u_{a,k} - \boldsymbol{\eta}_k^T \hat{\boldsymbol{\psi}}_{a,k}) - \frac{1}{2\sigma} \sum_{b=1}^B x_b^2 - \frac{1}{2\sigma} \sum_{b=1}^B y_b^2 \end{aligned} \quad (6.14)$$

6.4 Analyzing Tea Party Ideal Points

6.4.1 Data Collection

As motivated in Section 6.1.4, we are interested in applying our model to perform exploratory analysis of the Tea Party in the U.S. House of Representatives. To scale the multi-dimensional ideal points with respect to the Tea Party movement, we obtained the set of *key votes* identified by Freedom Works as the most impor-

tant votes on issues of economic freedom. Led by former House Majority Leader Dick Armey (R-TX), Freedom Works is a conservative non-profit organization which promotes “Lower Taxes, Less Government, More Freedom” and has been widely associated with the Tea Party movement.⁵ A recent study reports that, among the endorsements of various Tea Party organizations, Freedom Works endorsements are the most successful, associated with a statistically significant increase in votes for the Republican candidates in the 2010 midterm election (Karpowitz et al., 2011).

For the 112th Congress, Freedom Works selected 60 key votes, 40 in 2011 and 20 in 2012. Since in our study, we are interested in ideal points with respect to the Tea Party movement, i.e., on the anti-pro Tea Party dimension, we consider whether a legislator agrees with the position of Freedom Works on a bill the binary response used in scaling the ideal points. More specifically, we assign $v_{a,b}$ to be 1 if legislator a agrees with the position of Freedom Works on bill b , and 0 otherwise. In addition to the votes, we obtained the bill text with labels from the Congressional Bills Project as described in Chapter 4 and the congressional speeches as in Chapter 5. In total, we have 240 Republicans, 60 of which self-identify as a the member of the Tea Party Caucus, and 13,856 votes.

6.4.2 One-dimensional Ideal Points

First, as a baseline for comparison, we estimate the one-dimensional ideal points of each legislator in our dataset using Equation 6.1. We put a Gaussian prior $\mathcal{N}(0, \sigma)$ over u_a , x_b and y_b and use gradient ascent to optimize the following log

⁵<http://congress.freedomworks.org/>

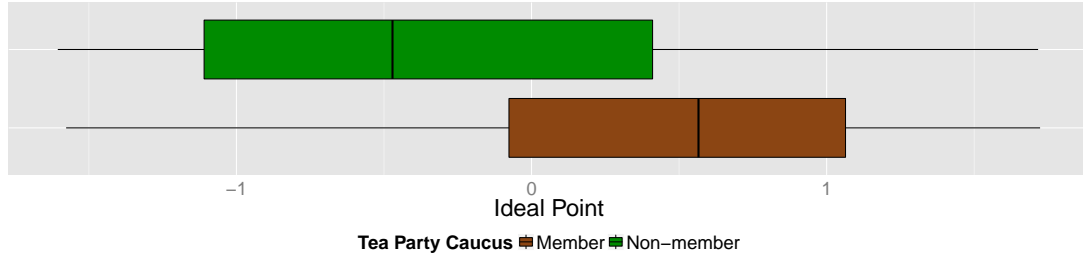
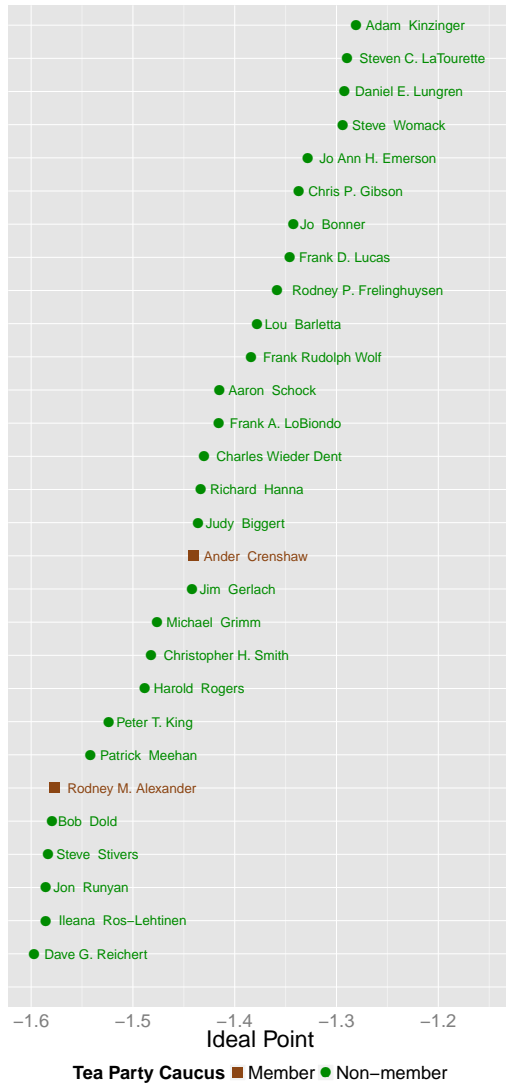


Figure 6.3: Box plots of the estimated one-dimensional Tea Party ideal points for members and non-members of the Tea Party Caucus among Republican Representatives in the 112th U.S. House. The median of members’ ideal points is significantly higher than that of non-members’ ideal points, though there are a lot of overlaps between the two distributions.

likelihood

$$\begin{aligned}
 \mathcal{L}(\mathbf{u}, \mathbf{x}, \mathbf{y}) &= \sum_{a=1}^A \sum_{b=1}^B v_{a,b} \log p(v_{a,n} = 1) + (1 - v_{a,b}) \log p(v_{a,b} = 0) \\
 &\quad - \frac{1}{2\sigma} \left(\sum_{a=1}^A u_a^2 + \sum_{b=1}^B x_b^2 + \sum_{b=1}^B y_b^2 \right)
 \end{aligned} \tag{6.15}$$

One problem with this type of ideal point model is that the signs of the estimated ideal points might be flipped. This is due to the fact that $u_a x_b = (-u_a)(-x_b)$ which makes no difference in Equation 6.1 if the sign of u_a and x_b are flipped. To avoid this problem and make sure our analysis is consistent in that higher ideal point values are associated with “pro-Tea Party”, we select a set of *anchor legislators* and initialize their ideal points with some predefined values (Gerrish and Blei, 2011). More specifically, we first sort the legislators according to the fraction of votes for which they agree with Freedom Works. Then, we initialize the ideal points of the top and bottom five legislators with $+3\sigma$ and -3σ respectively, where σ is the variance of the Gaussian prior we put on u_a .



(a)



(b)

Figure 6.4: Republican legislators having the (a) lowest and (b) highest estimated one-dimensional ideal points.

Figure 6.3 shows the box plots of estimated Tea Party ideal points for both members and non-members of the Tea Party Caucus among Republican Representatives in the 112th U.S. House. The estimated Tea Party ideal points are strongly correlated with the DW-NOMINATE scores, with a correlation coefficient of 0.908. As we can clearly see from the figure, the median ideal point of legislators with Tea Party Caucus membership is significantly higher than that of legislators who do not join the caucus. This observation confirms a widely accepted belief that Tea Partiers are generally more conservative than other Republicans (Williamson et al., 2011; Karpowitz et al., 2011; Gervais and Morris, 2012, 2014).

However, we can also see a great deal of overlap between the ideal points of the two groups. This shows that not all legislators with voting behaviors aligning with Freedom Works’s positions self-identify with the Tea Party Caucus. Figures 6.4a and 6.4b show the Republican Representatives who have the lowest and highest estimated ideal points respectively.

From our estimate, Jeff Flake (R-AZ) has the second highest ideal point, but is not a member of the Tea Party Caucus. Looking more closely into his voting record, out of 60 key votes selected by Freedom Works he only disagrees with Freedom Works’s position on one where he voted “Nay” on the bill “H.R.1: Full-Year Continuing Appropriations Act, 2011”. This bill includes the largest single discretionary spending cut in history, cutting \$106 billion from various programs and departments. Another example is Justin Amash (R-MI), who founded and is the Chairman the Liberty Caucus; its members are conservative and libertarian Republicans. Amash has agreed with Freedom Works on every single key votes selected

by Freedom Works since 2011.

Conversely, there are members of the Tea Party Caucus who do not often agree with Freedom Works, and thus have relatively low ideal points. For example, Rodney Alexander (R-LA), who agrees with Freedom Works only 48% of the time in the 112th Congress, was a member of the Democrat party before changing his party affiliation in 2004. Another example is Ander Crenshaw (R-FL) with 50% agreement with Freedom Works's positions on key votes in 2011 and 2012. Both Alexander and Crenshaw are categorized as “Green Tea” by [Gervais and Morris \(2014\)](#), which refers to Republican legislators who are strongly “associated with the Tea Party on their own initiative” but are not strongly supported by Tea Party organizations.

6.4.3 Multi-dimensional Ideal Points

In this section, we will analyze how the ideal points of the two groups of Republican Representatives are different from each other on different dimensions. Figure 6.5 shows the boxplots of the estimated ideal points for each policy agenda issue, sorted by the difference between the median of the two groups' ideal points. On most issues, the ideal point distributions of the two Republican groups overlap significantly. This is not surprising given that the one-dimensional ideal points of the two groups also overlap a great deal as we discuss in the previous section.

However, on several issues, the ideal point distributions of the two groups of legislators differ significantly. To understand why these issues polarize, we look at the set of key votes on each issue and how Republicans vote on them. Recall that

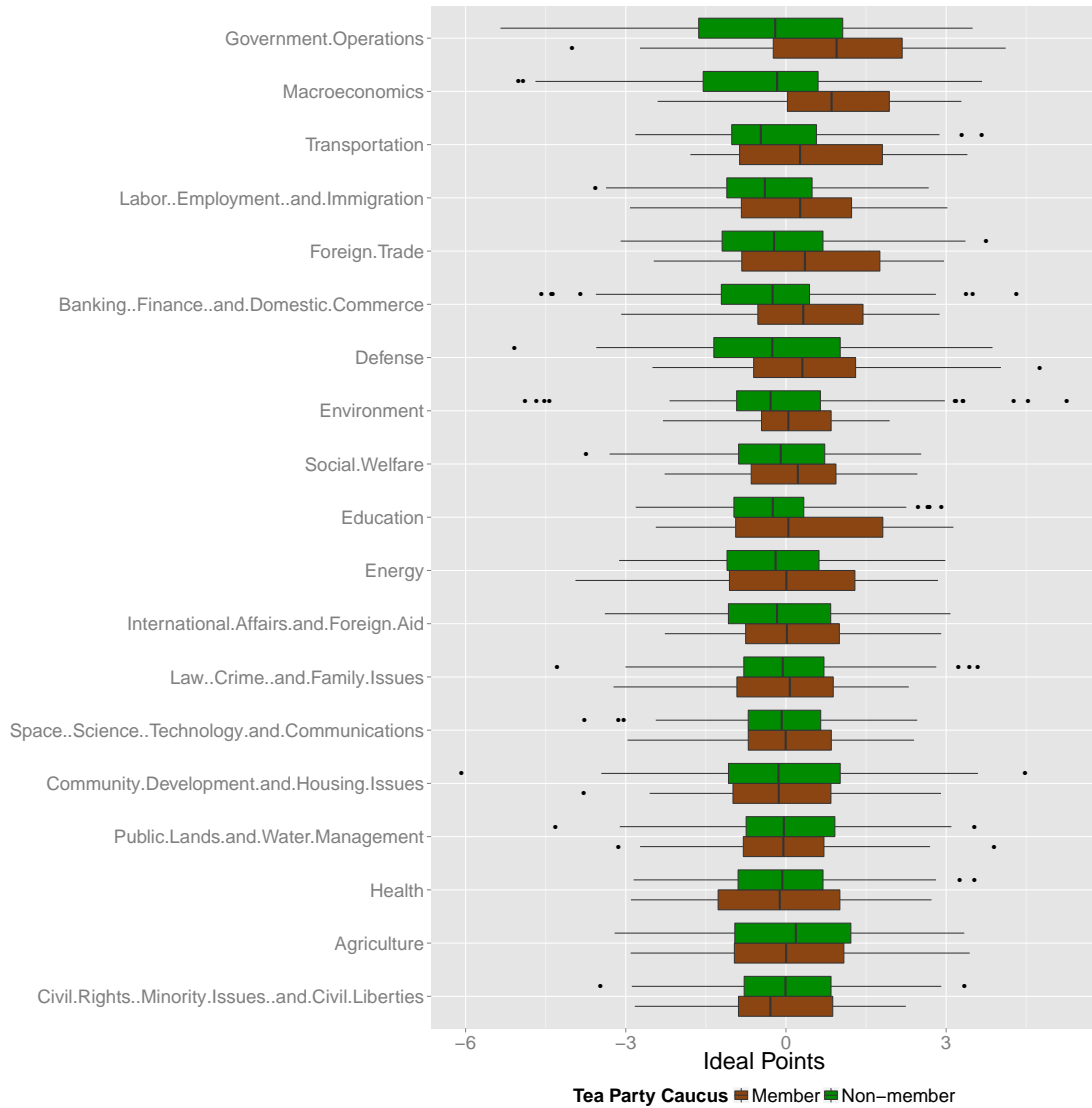


Figure 6.5: Boxplots of ideal points on 19 dimensions, each of which corresponds to a major topic in the Policy Agendas Codebook estimated by our model. On most issues the ideal point distributions over the two Republican groups (member vs. non-member of the Tea Party Caucus) overlap significantly. The most polarized issues are ‘Government Operations’ and ‘Macroeconomics’, which align well with the agenda of the Tea Party movement supporting small government and lower taxes.

in our model, each bill b has a distribution ϑ_b over K issues, capturing what the bill is about. For each key vote b , we choose the issue with the highest probability $\vartheta_{b,k}$ and use it to label. Although using a single issue for each key vote conflicts with our model’s admixture assumption in which each document is a mixture of topics, it provides a good estimation of what the key vote is primarily about and helps reduce the complexity of our analysis.

In the remainder of this section, we analyze the voting records of Republicans on the key votes that our model assigns to the “Government Operation”, “Macroeconomics”, and “Transportation” issues, which helps explain why our model estimates these issues as the most polarized.

Government operations Table 6.3 shows the list of key votes that are assigned to the “Government Operation” issue, with details about how the two groups of legislators vote on each key vote. As shown, Republicans do not unanimously agree on any vote. The majority of the two groups, members vs. non-members of the Tea Party Caucus, vote differently on eight out of eleven key votes on this issue. Most of these key votes are to reduce the government spending on various federal programs including the Economic Development Administration (key vote 2012-207), the Energy Efficiency and Renewable Energy Program (key vote 2012-311) and the Fossil Fuel Research and Development programs (key vote 2012-317). More specifically, for example, on the key vote *to eliminate the Energy Efficiency and Renewable Energy Program* (2012-311), nearly 80% (41 out of 53) of the Tea Party Caucus members vote “Yea” agreeing with the Freedom Works, while only about

43% of non-Tea Party Caucus members vote similarly. This difference in voting behaviors explains why this issue is estimated as the most polarized issue by our model, which aligns well with the agenda of the Tea Party movement fighting for less government and more federal spending cuts.

Macroeconomics Table 6.4 shows the list of key votes on ‘Macroeconomics’ estimated by HIPTM. Among these ten key votes, the majority of Republican legislators agree with each other on six of them including, for example, key vote to eliminate the requirement to submit an IRS form 1099 for all goods and services purchased over \$600 starting in 2012 (key vote 2011-162) or key vote to reform the way that the Congressional Budget Office (CBO) calculates the spending baseline each year (key vote 2012-32). However, the remaining four key votes are quite divisive among the Republicans.

Among these four, there are two key votes in which the majority of the two groups vote differently. These are key vote 2011-275 (*To replace the Paul Ryan budget with the RSCs budget*) and key vote 2012-149 (*Substitute amendment containing the Republican Study Committee budget for FY 2013*). Both of these key votes, one in 2011 and the other in 2012, are to replace Paul Ryan’s budget plan with the Republican Study Committee’s (RSC) alternate proposal “Back to Basics”, which would cut the government spending more aggressively to balance the federal budget in about a decade, instead of about three decades as in Paul Ryan budget. In 2011, the key vote 2011-275 split the Republican Representatives in the House right in the middle with 118 Yea’s and 119 Nay’s. However, while the majority (104 out

| ID | Key vote title | FW | # Agrees | | | # Disagrees | | | $\vartheta_{b,k}$ |
|-----------------|--|----|----------|----|-----|-------------|----|-----|-------------------|
| | | | All | TP | NTP | All | TP | NTP | |
| 2012-221 | To cut Commerce, Science, & Justice appropriations by 1% | Y | 156 | 48 | 108 | 77 | 10 | 67 | .47 |
| 2012-207 | To eliminate the Economic Development Administration | Y | 128 | 45 | 83 | 104 | 14 | 90 | .46 |
| 2012-222 | To cut \$2.7 billion from selected portions of CJS appropriations | Y | 105 | 35 | 70 | 128 | 23 | 105 | .46 |
| 2012-311 | To eliminate the Energy Efficiency and Renewable Energy Program | Y | 114 | 41 | 73 | 106 | 12 | 94 | .39 |
| 2012-336 | To cut \$3.1 billion from Energy and Water appropriations | Y | 125 | 43 | 82 | 110 | 15 | 95 | .36 |
| 2012-317 | To defund the Fossil Fuel Research and Development programs | Y | 102 | 39 | 63 | 123 | 19 | 104 | .35 |
| 2012-513 | To require a full audit of the Federal Reserve System and the Federal reserve banks | Y | 237 | 60 | 177 | 1 | 0 | 1 | .35 |
| 2011-538 | To cut spending 9.93% (\$3.04 billion) from Energy & Water Appropriations Act of 2012 | Y | 95 | 34 | 61 | 135 | 22 | 113 | .32 |
| 2012-450 | Making appropriations for the Departments of Transportation, and Housing and Urban Development | N | 54 | 22 | 32 | 182 | 36 | 146 | .31 |
| 2011-434 | To cut \$900 million in waste and apply to a spending reduction account | Y | 108 | 38 | 70 | 128 | 22 | 106 | .30 |
| 2011-424 | To cut \$700 million dollars in waste to pay off the debt | Y | 82 | 32 | 50 | 151 | 28 | 123 | .30 |

Table 6.3: Key votes having “Government operations” as the most probable issue, estimated by our model. The last column shows the estimated probability $\vartheta_{b,k}$. Each key vote is shown with a short description, the preferred voting position of Freedom Works (Y for Yea, N for Nay), the number of Republicans whose votes agree and disagree with Freedom Works (‘All’ denotes all voting Republican legislators, ‘TP’ denotes Tea Party Caucus members, and ‘NTP’ denotes non-Tea Party Caucus members). Bolded key votes are the ones on which the majority of the two groups vote differently.

| ID | Key vote title | FW | # Agrees | | | # Disagrees | | | Est |
|-----------------|---|----|----------|----|-----|-------------|----|-----|------|
| | | | All | TP | NTP | All | TP | NTP | |
| 2011-275 | To replace the Paul Ryan budget with the RSC’s budget | Y | 118 | 45 | 73 | 119 | 15 | 104 | 0.85 |
| 2011-277 | Congressman Paul Ryan’s Budget for Fiscal Year 2012 | Y | 234 | 58 | 176 | 3 | 2 | 1 | 0.84 |
| 2012-149 | Substitute amendment containing the Republican Study Committee budget for FY 2013 | Y | 135 | 50 | 85 | 104 | 10 | 94 | 0.78 |
| 2011-690 | The Budget Control Act of 2011 | N | 65 | 27 | 38 | 173 | 33 | 140 | 0.74 |
| 2011-162 | Small Business Paperwork Mandate Elimination Act of 2011 | Y | 236 | 60 | 176 | 0 | 0 | 0 | 0.71 |
| 2011-606 | Cut, Cap, and Balance Act | Y | 229 | 58 | 171 | 8 | 2 | 6 | 0.69 |
| 2012-32 | To amend the Balanced Budget and Emergency Deficit Control Act of 1985 to reform the budget base-line | Y | 231 | 57 | 174 | 0 | 0 | 0 | 0.64 |
| 2012-659 | On Concurring with the Senate Amendments: H.R. 8 - Taxpayer Relief Act of 2012 | N | 150 | 50 | 100 | 84 | 9 | 75 | 0.57 |
| 2011-14 | Repealing the Job-Killing Health Care Law Act | Y | 238 | 60 | 178 | 0 | 0 | 0 | 0.51 |
| 2011-901 | Making major executive regulations subject to Congressional vote (REINS Act) | Y | 236 | 59 | 177 | 0 | 0 | 0 | 0.49 |

Table 6.4: Key votes having “Macroeconomics” as the most probable issue, estimated by our model. The last column shows the estimated probability $\vartheta_{b,k}$. Each key vote is shown with a short description, the preferred voting position of Freedom Works (Y for Yea, N for Nay), the number of Republicans whose votes agree and disagree with Freedom Works (‘All’ denotes all voting Republican legislators, ‘TP’ denotes Tea Party Caucus members, and ‘NTP’ denotes non-Tea Party Caucus members). Bolded key votes are the ones on which the majority of the two groups vote differently.

| ID | Key vote title | FW | # Agrees | | | # Disagrees | | | Est |
|-----------------|--|----|----------|----|-----|-------------|----|-----|------|
| | | | All | TP | NTP | All | TP | NTP | |
| 2012-378 | To require that transportation spending be capped | Y | 82 | 32 | 50 | 145 | 23 | 122 | 0.58 |
| 2012-451 | To provide an extension of Federal-aid highway ... transit, and other programs | N | 51 | 20 | 31 | 186 | 38 | 148 | 0.57 |

Table 6.5: Key votes having “Transportation” as the most probable issue, estimated by our model. The last column shows the estimated probability $\vartheta_{b,k}$. Each key vote is shown with a short description, the preferred voting position of Freedom Works (Y for Yea, N for Nay), the number of Republicans whose votes agree and disagree with Freedom Works (‘All’ denotes all voting Republican legislators, ‘TP’ denotes Tea Party Caucus members, and ‘NTP’ denotes non-Tea Party Caucus members). Bolded key votes are the ones on which the majority of the two groups vote differently. Both of these votes focus on the federal spending

of 119) of non-Tea Party Caucus members vote against this amendment, 45 out of 60 members of the Tea Party Caucus vote for it. In 2012, even more Tea Party Caucus members vote for the key vote 2012-149, while there are still more than half of non-Tea Party Caucus members vote against it.

Although not as polarized as the two key votes above in which the majority of both groups vote similarly, the two remaining key votes still see a lot of disagreements among the Republicans. The first key vote is to the *Budge Control Act of 2011* (key vote 2011-690) which allows President Obama to raise the debt ceiling to over \$16 trillion, while the second key vote is about the *Taxpayer Relief Act of 2012* (key vote 2012-659) to avert the “fiscal cliff”.

Transportation The third most polarized issue estimated by our model is “Transportation”, which includes two key votes focusing on the federal spending on transportation. The first key vote (2012-378) is about the motion to insist upon capping

highway spending at the amount taken in by the gas tax. More than half of Tea Party Caucus members (32 out of 55) vote for this motion, while the majority of non-members vote against it. Conversely, the second key vote (2012-451) is to authorize federal highway spending at a level that far exceeds its revenue from the gas tax, which is opposed by Freedom Works and the majority of the two Republican groups.

6.5 Agendas and Frames: Analyzing Topic Hierarchy

In this section, we qualitatively analyze the topic hierarchy discovered by HIPTM. We first focus our analysis on the topics learned by HIPTM at each first-level node, i.e., an agenda issue which corresponds to a major topic in the Policy Agendas Topics Codebook. These topics in general capture the issues that Republican legislators focus on during the 112th Congress. We then take a closer look at some subtrees which contain most polarized second-level nodes.

6.5.1 Analyzing Agenda Issues

Table 6.6 shows the list of words with highest probabilities for each issue. As we can observe, in general, the topics learned at all first-level nodes coherently describe the corresponding agenda issues. This is assuring since our model learns this set of topics leveraging the prior distributions over words from labeled data as shown in Table 6.2.

More interestingly, the learned topics capture some key debates happened on

| |
|--|
| Agriculture: farmer agriculture food farm usda fda farm_bill brazil rancher art radio crop rural |
| Banking, Finance, and Domestic Commerce: patent internet fcc file inventor nfip flood_insurance innov fee pto fema application invent financial_service |
| Civil Rights, Minority Issues, and Civil Liberties: abort plan_parenthood baby child taxpay_dollar taxpay_fund federal_fund human clinic unborn conscience mother ohio protect_life |
| Community Development and Housing Issues: loan bank mortgage homeowner treasury property failure country borrow foreclosure payment lender sell terminate |
| Defense: afghanistan troop mission iraq air_forc armi intellig_commun soldier nation_defens uniform intellig navi command pakistan laden |
| Education: student freedom parent charter_school principl liberti indiana kid colleg columbia countih |
| Energy: oil drill gulf mexico leas permit gasolin pipelin pump american_energi fuel moratorium explor gallon |
| Environment: epa clean_air permit plant environment_protect emiss rule_xxi mercuri pollut clean_water florida cement compli coal_ash environment |
| Foreign Trade: export trade trade_agreement manufactur colombia panama south_korea tariff free_trade textil china custom duti job_bill intern |
| Government Operations: motion recommit revis union georgia accordingli insert commiss legisl_dai florida tempor michigan february short_titl |
| Health: obamacar patient doctor physician afford_care hospit insur replac mandat exchang health_insur coverag medicaid patient_protect board |
| International Affairs and Foreign Aid: libya human israel peac war_power violat commiss regim democraci freedom march hostile unit_nation alli articl |
| Labor, Employment, and Immigration: employ hire job_creator south_carolina union busi_owner nlrh uncertainti boe labor mandat manufactur econom_growth |
| Law, Crime, and Family Issues: border patriot_act judg court law_enforc enforc terrorist investig homeland_secur crimin crime extens citi alabama attorney |
| Macroeconomics: balanc_budget borrow debt_ceil cap cut_spend nation_debt grandchildren social_secur rais_tax debt_limit white_hous spend_monei gdp chart |
| Public Lands and Water Management: water river flood arizona engin mine fish west corp lake endang_speci copper idaho dam interior |
| Social Welfare: underli_bill continu_resolut revis homeland_secur legisl_dai transpar underli_legisl spend_reduct regular earmark |
| Space, Science, Technology and Communications: victim gabbi arizona prai tragedi prayer father wife saturdai mother violenc wound event duti medal |
| Transportation: transport extens faa airport reauthor jurisdict data sincer aviat flight confer aircraft titl air |

Table 6.6: Words with highest probabilities for each first-level issue nodes learned by HIPTM.

the congressional floor during the 112th Congress. For example, one major event during this Congress is the *debt-ceiling crisis of 2011*, in which major debate between the Republican Party, which had taken control of the House the prior year, and the President centered around the raising of the debt ceiling. This debate dominates the discussions on “Macroeconomics”, whose learned topic focuses on “balanc_budget”, “borrow”, “debt_ceiling”, “cap”, “cut_spending”, “nation_debt”, etc. Another interesting event during this period of time is the *international military intervention* of the U.S. in the Libyan Civil War, which is the focus of debates on “International Affairs and Foreign Aid”. Debates in this Congress on the “Defense” issue center around the *withdrawal of troops* from Iraq in December 2011, which formally ends the Iraq War. There are also a lot of discussions on *repealing* the Affordable Care Act, more commonly known as ObamaCare, from Republican legislators.

6.5.2 Analyzing Issue-specific Frames

We now turn our focus on the second-level nodes of the hierarchy, which are designed to capture issue-specific frames. In our model, each second-level frame node is associated with a regression parameter $\eta_{k,j}$, which is essentially the ideal point of that frame on the dimension corresponding to the issue that the frame belongs to. To analyze polarized issues, we first compute, for each issue k , the span of the ideal points of the frames associated with k .⁶ This span is defined as the difference between the maximum ideal point and the minimum ideal point of any

⁶The Dirichlet process prior we put on the frame proportions ψ_k has the “rich-get-richer” effect, in which a few frames get used a lot and there are always a set of unstable frame nodes which gets created and destroyed frequently during the sampling process. In this analysis, we focus on the more stable frames and ignore those with posterior probability $\psi_{k,j} < 0.1$.

frames under that issue. We then sort all issues by this difference. In the remainder of this section, we will analyze in more detail the issues that are most polarized according to our model.

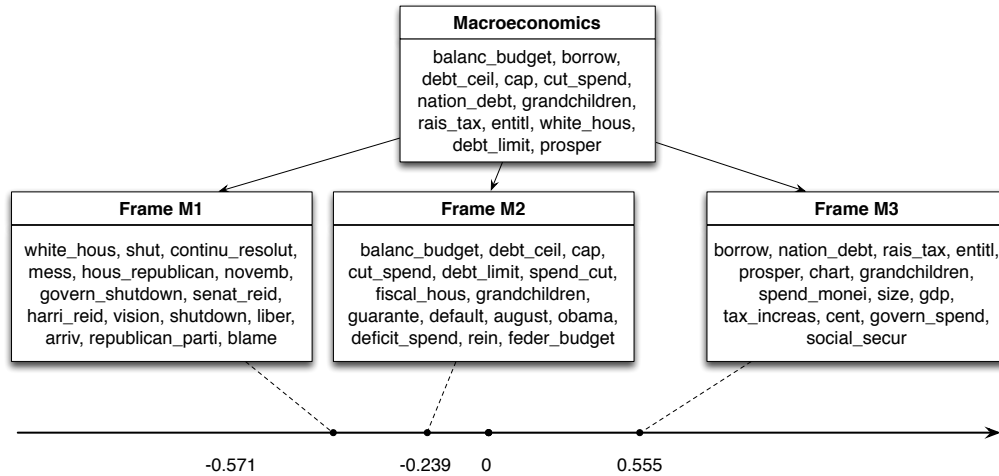


Figure 6.6: Subtree on “Macroeconomics” learned by our model.

Macroeconomics Figure 6.6 shows the subtree on “Macroeconomics” in the topic hierarchy learned by our model. The most positive frame node, Frame M3, focuses on criticizing *government overspending*. Our model reveals that many members of the Tea Party Caucus center their speeches on this issue including Todd Akin (R-MO), Steven E. Pearce (R-NM) and Lamar S. Smith (R-TX). Also about budget balancing, both Frame M1 and Frame M2 discuss problems surrounding the *debt-ceiling crisis of 2011*. While Frame M1 is about the potential government shutdown of the crisis, Frame M2 mainly focuses on the “Cut, Cap, and Balance Act of 2011”, which includes a cut in federal government spending, a cap on future spending, and an increase in the national debt ceiling on certain conditions. As shown in Table 6.4, the vast majority of Republican Representatives vote for this bill.

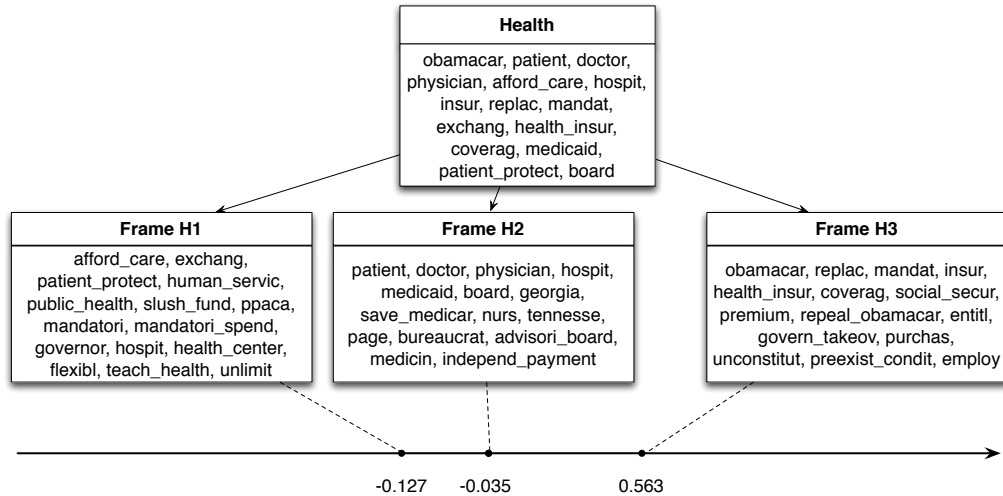


Figure 6.7: Subtree on “Health” issue in the topic hierarchy learned by our model.

Health Health care is a central issue during the 112th Congress with debates around the Affordable Care Acts. Even though all Republicans vote against the health-care reform bill, Figure 6.7 shows some distinctions in the languages that Republicans talk about this issue. The first two frame nodes, Frames H1 and H2, talk about various aspects of the health care system including Medicare and Medicaid. For example, Glenn Thompson (R-PA) argues “The only prescription to save Medicare is a Republican prescription. I have to tell you, on the Democratic side, they’re just willing to pull the plug and let it die, because if you don’t make changes to the Medicare program, that’s exactly what happens.” On the other hand, Frame H3, being highly positive, emphasizes strongly on *repealing ObamaCare*. Talking about the issue, Michele Bachmann (R-MN) argues

“This Chamber already passed a bill to repeal ObamaCare, which the American people have asked. This is now an effort to defund ObamaCare. Because as we have seen from the Congressional Research Service, the inge-

nious nature of the ObamaCare bill was to already put the funding in place so that if the majority lost the gavel, which they did, the new majority would be unable to defund this bill.”

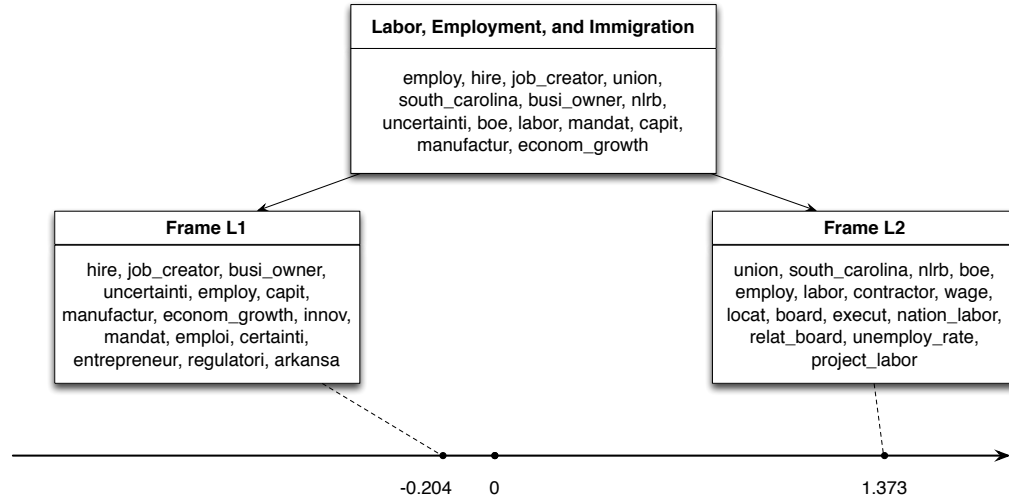


Figure 6.8: Subtree on the “Labor, Employment and Immigration” issue in the topic hierarchy learned by our model.

Labor, Employment and Immigration Figure 6.8 shows the subtree corresponding to the “Labor, Employment and Immigration” issue in the topic hierarchy discovered by our HIPTM model. Analyzing the text assigned to each frame node by our model together with the topic at each node reveals that Frame L1 is mainly about bill H.R. 4 which is to eliminate the paperwork mandate for small businesses. This bill is the focus of the key vote 2011-162 (*Small Business Paperwork Mandate Elimination Act of 2011*) in Table 6.4. As shown in the voting records in Table 6.4, all Republican legislators vote ‘Yea’ unanimously on this bill, which explains the negative ideal point of this frame.

On the other hand, Frame L2 is highly positive, which mainly focuses on the

controversial case between the National Labor Relations Board (NLRB) and the airline manufacturer Boeing. In 2011, Boeing built a new 787 assembly plant in South Carolina and was accused by the NLRB for violating “federal labor law deciding to transfer a second airplane production line from a union facility in the state of Washington to a non-union facility in South Carolina for discriminatory reasons”.⁷ This complaint from the NLRB was strongly opposed by Representatives from South Carolina, which is captured by our model. The Republican Representatives who talk about this the most, revealed by our model, include Trey Gowdy (R-SC), Addison G. Wilson (R-SC), Mick Mulvaney (R-SC) and Jeff Duncan (R-SC). Among these Representatives, all but Trey Gowdy are members of the Tea Party Caucus and all four legislators have high ideal points (Figure 6.4b), which explains the highly positive ideal point of this frame node.

Although the polarization in the two frames under this issue might not be directly related to the Tea Party movement, it shows an interesting example of the effectiveness of our model as an exploratory data analysis tool. Since our speech data include all of the congressional floor debates in the House, many discussions are not particularly about issues that the Tea Party focuses on. Discovering the polarization between Tea Partiers and non-Tea Partiers on unconventional issues provides interesting insights in the differences between them. As an example, the polarization in this policy issue discovered by our model, is mainly due to a geographic reason: the conflict between NLRB and Boeing happened in South Carolina

⁷<http://www.nlr.gov/news-outreach/fact-sheets/fact-sheet-archives/boeing-complaint-fact-sheet>

where many Representatives are members of the Tea Party Caucus. This might provide evidence on the influence of the geographic factors on a member’s decision to join (or not) the Tea Party Caucus (Gervais and Morris, 2012).

6.6 Predicting Tea Party Caucus Membership

To quantitatively evaluate the effectiveness of our proposed HIPTM model in capturing the “Tea Partiness” of legislators, we conduct experiments on a binary classification task to predict the Tea Party Caucus membership of legislators given their votes and text. The goals of our experiments are to examine (1) how effective the baseline features extracted from the votes and text are in predicting the Caucus membership, and (2) how much improvement in prediction performance, if any, we can gain using the features extracted from our model. For the baselines, we consider the following sets of features:

- **Normalized term frequency (TF)**: each legislator is represented by a vector of term frequency of all word types in the vocabulary, normalized to unit length.
- **TF-IDF**: each legislator is represented by a TF-IDF vector.
- **Vote**: each legislator is represented by a binary vector containing their voting record on the set of key votes selected by Freedom Works. If a vote is not recorded, we treat it as a missing value.

In our dataset from the 112th U.S. Congress, there are totally 240 Republican Representatives, out of which 60 self-identify as Tea Party Caucus members.

We perform 5-fold cross-validation using stratified sampling, which preserves the ratio of the two classes in both the training and test sets. We use AUC-ROC, which measures the area under the Receiver-Operating-Characteristic (ROC) curve, as the evaluation metric. We use SVM as the classifier and use the $\text{SVM}^{\text{light}}$ implementation (Joachims, 1999).⁸ We preprocess the data using similar pipeline as described in Chapter 2. After preprocessing, our vocabulary contains 5,349 unique word types.

6.6.1 Membership Prediction given Votes and Text

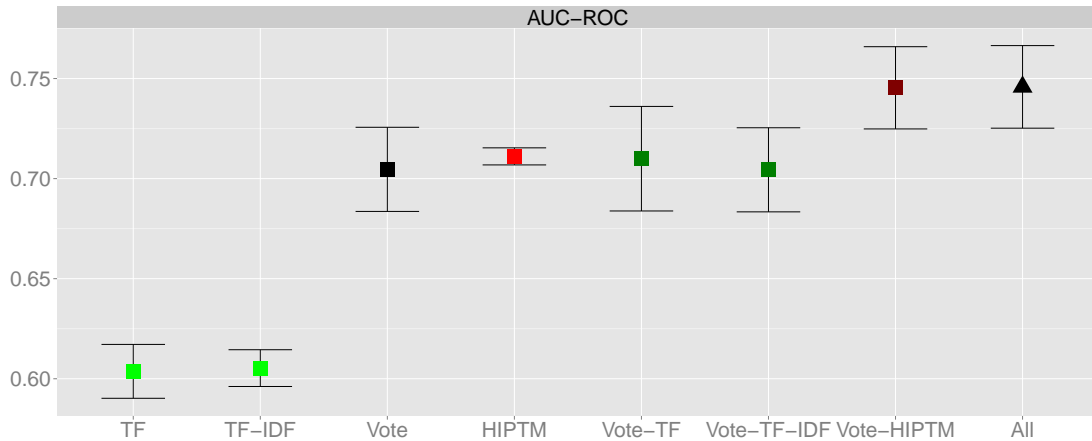


Figure 6.9: Tea Party Caucus membership prediction results over five folds using AUC-ROC (higher is better, random baseline achieves 0.5). The features extracted from our model are estimated using both the votes and the text.

First, given the votes and text of all the legislators, we run HIPTM for 1,000 iterations with a burn-in period of 500 iterations. After burning in, we keep the sampled state of the model after every 50 iterations. The feature values are obtained by averaging over the 10 stored models. Each legislator a is represented by a vector

⁸We use the default settings of $\text{SVM}^{\text{light}}$, except that we set the cost-factor equal to the ratio between the number of negative examples (i.e., number of non-Tea Party Caucus members) and the number of positive examples (i.e., number of Tea Party Caucus members).

concatenating the following features:

- K dimensional ideal point vector estimated from both votes and text $u_{a,k}$
- K dimensional vector, estimating the ideal point using only text $\boldsymbol{\eta}_k^T \hat{\boldsymbol{\psi}}_{a,k}$
- B probabilities estimating a 's votes on B bills $\Phi(x_b \sum_{k=1}^K \hat{v}_{b,k} u_{a,k} + y_b)$

Figure 6.9 shows the AUC-ROC results for different sets of features. The results show that all feature sets perform better than the random predictor, which always achieves an AUC-ROC of 0.5. VOTE-based features clearly outperform significantly text-based features like TF and TF-IDF. Combining VOTE with either TF or TF-IDF does not improve the prediction performance much (i.e., VOTE-TF and VOTE-TF-IDF). The set of features extract from our model, HIPTM, also outperforms TF and TF-IDF significantly, but only slightly better than VOTE. However, when combining HIPTM and VOTE, we can achieve relatively large gain compared with using VOTE alone.

6.6.2 Membership Prediction given Text Only

In the previous section, we experiment with features from both the votes and the text from legislators to predict their Tea Party Caucus memberships. However, its applicability is limited since we need to have both the votes and text to be able to make predictions. In this section, we look at a more difficult, yet more practical problem, which predicts the Tea Party Caucus membership using only the text of new lawmakers.

We first run our inference algorithm on the training data, which does include

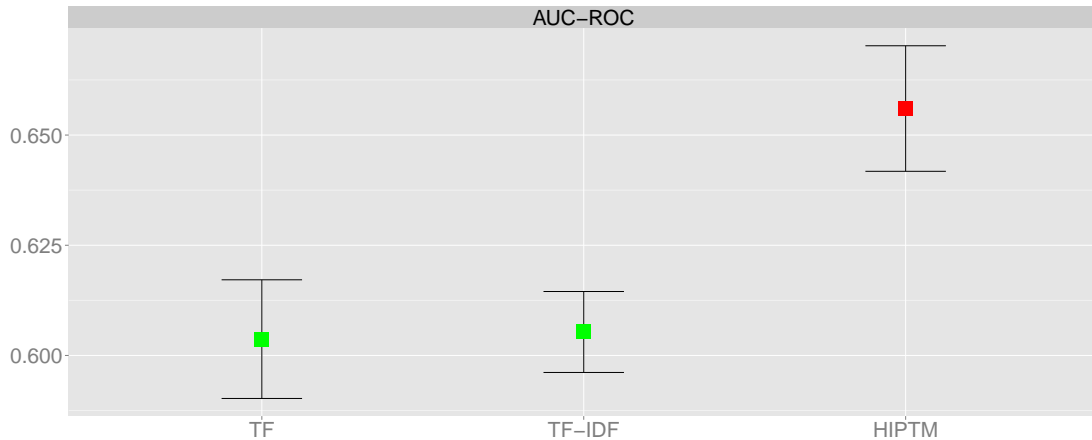


Figure 6.10: Tea Party Caucus membership prediction results over five folds using AUC-ROC (higher is better, random baseline achieves 0.5). The features extracted from our model for unseen legislators are estimated using their text only.

both votes and text. After training, using multiple models stored, we sample the issue and frame assignments for each token of the text authored by test lawmakers.⁹

Since the votes are not available, in this section, HIPTM’s extracted features only consist of (1) the K dimensional vector estimating legislators’ ideal point using text only $\boldsymbol{\eta}_k^T \hat{\boldsymbol{\psi}}_{a,k}$, and (2) the B probabilities estimating the votes $\Phi(x_b \sum_{k=1}^K \hat{v}_{b,k} u_{a,k} + y_b)$.

Figure 6.10 shows the results of our features in comparison with the two text-based baselines TF and TF-IDF. As we can see, since HIPTM can no longer access the votes in the test data, its performance drops significantly compared with VOTE. However, HIPTM still outperforms significantly the two text-based baselines TF and TF-IDF, which shows that our model provides an effective set of features, compared with other commonly used text-based baselines, to capture the “Tea Partiness” of legislators.

⁹The MCMC configuration is the same as in Section 6.6.1.

6.7 Conclusion and Future Directions

6.7.1 Summary

In this chapter, we continue the approach we took in Chapter 5 to develop a hierarchical topic model to discover and analyze agendas and frames in political text. We present HIPTM, a Hierarchical Ideal Point Topic Model, which jointly captures (1) the votes of legislators on congressional bills and (2) both the text associated with the legislators (e.g., congressional speeches) and the bills (e.g., bill text). By leveraging existing labeled data from the Congressional Bills Project, HIPTM estimates ideal points for each legislators on multiple interpretable dimensions, each of which corresponds to a major topic in the Policy Agendas Topics Codebook. Moreover, the model is also able to discover a hierarchy of topics, in which first-level nodes maps to agenda issues and second-level nodes maps to issue-specific frames.

We apply HIPTM to perform exploratory data analysis on how Republicans votes and debates in the 112th U.S. House of Representatives. Our analysis using one-dimensional ideal points shows that, among the Republican Representatives, the single ideological positions of members and non-members of the Tea Party Caucus overlap a great deal. The results of our model on multi-dimensional ideal points help disentangle this overlapping by showing on which issues the two groups of Republicans hold similar positions and on which issues they do not. We also analyze the topic hierarchy learned by HIPTM. Our analysis shows that topics learned at first-level nodes indeed can reveal the focus of congressional floor debates on each

issue. Topics at second-level nodes, even though not as coherent, can capture the polarization in certain issue.

To quantitatively evaluate the effectiveness of HIPTM, we conduct experiments on predicting the Tea Party Caucus membership of legislators. Empirical results show that vote-based features are much more effective than text-based features in predicting the caucus membership. However, voting information is not always available, especially for those who are not in the Congress. We show that using only text, HIPTM’s features outperform significantly two traditional text-based sets of features TF and TF-IDF.

6.7.2 Discussion and Future Directions

In this chapter, we mainly focus on studying and analyzing votes and text of legislators in the U.S. Congress. However, as shown in Section 6.6.2, our HIPTM model provides an effective way to study multi-dimensional ideal points of new legislators given only the text that they author. This gives our method the flexibility to study not only legislative text but also text from potentially many other sources such as press releases, debates, and social media. Applying our model to study multi-dimensional ideal points of people, given their text in different settings than the U.S. Congress, is one interesting future direction that we plan to pursue. This research direction also falls under a broader research area on scaling ideal points using new sources of data such as social media like Twitter (Barberá, 2015) or campaign contributions (Bonica, 2014).

In addition, as a case study for our model, we focus our analysis on Tea Party in the U.S. House by using the set of key votes selected by Freedom Works. However, the analysis can be naturally extended to different sets of legislative votes to study other political phenomena of interest. For example, one might be interested in the *Immigration Reform* and selects a set of legislative votes that are related to immigration, which using our model can analyze how legislators with different perspectives on the Immigration Reform talk about different issues. Another interesting direction is to apply the model to political data from local government entities such as state, county, district, and city. This might be of particular interest and usefulness because manual reading and human coding at this level are especially challenging due to the large number of lawmakers and legislative documents.

Studying the changes in ideological positions of lawmakers over time is also another application that might benefit from using our model. One straightforward way is to apply HIPTM to longitudinal data to estimate the ideologies at different points in time. Recent advances in topic modeling such as the dynamic topic model (Blei and Lafferty, 2006) or the recursive Chinese restaurant process (Ahmed and Xing, 2008) also give us powerful computational tools to jointly model the timestamped data and capture how agenda issues and issue-specific frames change over time.

Chapter 7: Conclusion and Future Work

Extracting from text *what* topics people talk about and *how* they talk about them is an important, but very challenging problem. Traditional approaches to this problem rely on manual coding and close reading which are labor-intensive and not scalable to big data. In this thesis, we have presented a series of *automated content analysis* methods using the probabilistic topic modeling approach to discover and analyze agendas and frames in political text *at lower cost*. In this last chapter, we summarize the contributions of the methods introduced in the thesis and discuss some directions for future work.

7.1 Summary of Contributions

In Chapter 3, we present the *Speaker Identity for Topic Segmentation* (SITS) model to study agendas and agenda control behaviors of individual in political debates and other multi-party conversations. The model uses Bayesian nonparametrics to improve existing methods and is able to discover (1) the topics used in a set of conversations, (2) how these topics are shared across conversations, (3) when these topics changes, and (4) a speaker-specific measure of agenda control. Using SITS we analyze the agenda control behaviors of candidates in the 2008 U.S. election debates

and the 2012 Republican primary debates. We also apply SITS on a large-scale set of political debate transcripts from CNN’s TV show *Crossfire*. To make the analysis process more effective, we build *Argviz*, an interactive visualization which leverages SITS’s outputs to allow users to quickly grasp the topical dynamics of the conversation, discover when the topic changes and by whom, and interactively visualize the conversation’s details on demand. In addition to providing insights on agendas and agenda control in conversations, through extensive empirical experiments, we also show that SITS can effectively improve the performance of two quantitative tasks: topic segmentation and influencer detection.

In Chapter 4, we study agendas in legislative text in the U.S. Congress. We introduce the *Label-to-Hierarchy* (L2H) model to learn a hierarchy of topics from multi-labeled data, in which each congressional bill is tagged with multiple policy agenda issues from a flexible list of labels. We discuss the advantages of using this type of labeled data over traditional single-labeled data using a fixed coding system: (1) it captures the multi-faceted nature of many congressional bills, and (2) it helps reduce the pre-analysis cost of creating and maintaining the well-defined coding system. The introduced model L2H captures the dependencies among labels using an interpretable tree-structured hierarchy. Applying L2H on congressional bill text from four U.S. Congresses (109th–112th), qualitative analysis shows that L2H is able to learn interpretable hierarchies, which helps provide insights about the political attentions that policymakers focus on, and how these policy issues relate with each other. Quantitative experiments also show the effectiveness of L2H on two computational tasks: predicting held-out words in test documents and predicting

multiple labels for unseen text.

In Chapter 5, we go beyond agenda-setting (i.e., what topics people talk about) and expand our focus to framing (i.e., how they talk about different issues). We describe the *Supervised Hierarchical Latent Dirichlet Allocation* (SHLDA) model, which can discover a hierarchy of topics from a collection of documents, each is associated with the ideological position of the author on a liberal-conservative spectrum, generally called the response variable. In the topic hierarchy discovered by SHLDA, higher-level nodes map to more general agenda issues while lower-level nodes map to issue-specific frames. Applying SHLDA on a collection of congressional floor debates, we show qualitatively that the topic hierarchies learned by SHLDA indeed capture the topic structure in line with the theory that motivates the work. Quantitative experiments on predicting the response variable show that SHLDA can improve the performance over commonly used baselines. Without using any topic labeled data, SHLDA enjoys a low pre-analysis cost but suffers from a moderately high post-analysis cost due to the complex and abstract nature of framing. Improving the interpretability of the hierarchy motivates the work in the next chapter.

In Chapter 6, we continue the approach in Chapter 5 to develop hierarchical topic model to discover and analyze agendas and frames in political text. We introduce the *Hierarchical Ideal Point Topic Model* (HIPTM) which jointly captures (1) the votes of legislators on congressional bills and (2) both the text associated with the legislators (e.g., congressional speeches) and the bills (e.g., bill text). More customized specifically for capturing the two-level views of agenda-setting and framing

than in SHLDA, HIPTM discovers a two-level hierarchy of topics in which first-level nodes map to policy agenda issues and second-level nodes map to issue-specific frames. To improve the interpretability of issue nodes, we leverage existing labeled data from the Congressional Bills Project to build topic priors, each corresponds to one of the 19 major topics in the Policy Agendas Topics Codebook. In addition, instead of using pre-computed ideal point like in SHLDA, HIPTM jointly estimate multi-dimensional ideal points of legislators, in which each dimension maps to one of the 19 interpretable topics. We show the effectiveness of HIPTM as an exploratory data analysis tool by applying on data from the 112th Congress to study the language uses (via discovered agendas and frames) and voting behaviors (via estimated ideal points) of members of the Tea Party Caucus, the institutional organization of the recent Tea Party movement in American politics.

7.2 Directions for Future Work

Measuring analysis cost of automated content analysis Even though the analysis costs described in Chapter 1, based on the work by [Quinn et al. \(2010\)](#), has played a central role of a guiding framework for developing models in this thesis, actually measuring the cost quantitatively is not the focus of this research. We focus mainly on developing, implementing and applying our models to study agendas and frames in political text and other related settings. However, we believe that quantifying the cost of automated content analysis methods for analyzing agendas and frames in particular and for political text in general is challenging yet very useful, especially

in guiding the development of future models.

There are various components of the cost whose estimation can benefit greatly from various prior work. For example, measuring the labeling cost, in general, has been studied extensively in active learning, a subfield of machine learning in which the learning algorithm is allowed to select the data, ask for label and then incorporate into the training set (Settles, 2012). The post-analysis cost of interpreting discovered topics for topic models is also related to work on quantifying the topic coherence (Chang et al., 2009b; Lau et al., 2014b), which we briefly discuss in Chapter 2 and expand the discussion for hierarchy of topics next.

Evaluating topic hierarchy Three out of the four models we introduce in this thesis discover a hierarchy of topics from text data. L2H in Chapter 4 learns a tree-structured hierarchy to capture the dependency among labels in multi-labeled data, in which each node consists of a predefined label and a topic. To capture the hierarchical view of agendas and frames, both SHLDA in Chapter 5 and HIPTM in Chapter 6 discover topic hierarchies in which higher-level nodes map to agenda issues and lower-level nodes map to issue-specific frames. For all of these models, we perform qualitative analysis of the discovered hierarchies and show that they capture intuitively the hierarchical structures that motivate the work. However, a more formal evaluation of the hierarchy is needed.

As briefly mentioned in Chapter 2, evaluating topics learned by topic models in general is an active research area. Various approaches have been proposed in the literature including topic coherence using automatic measurements (Newman et al.,

2010; Mimno et al., 2011; Aletras and Stevenson, 2013a; Lau et al., 2014b) and human judgement (Chang et al., 2009b). However, when evaluating topic hierarchy, not only we need to judge the quality of each individual topic, we also need to evaluate the hierarchical structure such as the topics at parent nodes should be more general than those of their child nodes. We believe evaluating the quality of discovered topic hierarchy is still an open research question.

Applying to new applications Although all are motivated by specific problems in political science to study agendas and frames, many models introduced in this thesis can be applied to much broader settings.

First, L2H is applicable to any multi-labeled data in which each document is tagged with multiple labels. It is particularly useful when the number of unique labels is large and there are potentially dependencies among them, for which L2H captures using a tree-structured hierarchy. The learned hierarchy can be used for various purposes including (1) letting the users search and browse the label space efficiently, (2) improving predictions of labels for unseen text as shown in Chapter 4, and (3) updating the existing label hierarchy (in the case that such hierarchy exists and the labeled data change over time (Ramage et al., 2010b)).

Second, SHLDA is broadly applicable to any setting in which there is a set of documents, each of which is associated with a continuous response variable. As shown already in the experiments in Chapter 5, SHLDA improves the prediction performance over commonly used baselines when applied to two domain: sentiment analysis to predict review ratings given the review text, and congressional floor

debates to predict legislators' ideological positions based on their speeches.

Using different loss functions In both models introduced in Chapter 5 and Chapter 6, we use a Gaussian linear regression model to link the empirical distribution over topics with the continuous responses, either observed or latent. Prior research has shown that using the max-margin principle yields models that learn more discriminative topics and thus achieve better predictive performance (Zhu et al., 2012). Max-margin learning falls under the umbrella of using different loss functions to model the metadata, which we plan to explore in future research. This approach has become particularly attractive due to recent advances in machine learning which makes max-margin supervised topic models more scalable (Zhu et al., 2013, 2014b).

Collaborating more closely with social scientists All of the models introduced in this thesis are motivated by different questions and problems in political science in particular and social science in general. These models join other work in the emerging field of computational social science (Lazer et al., 2009) to provide social scientists new computational tools to perform studies and analysis of large-scale data that are impractical otherwise. On the other hand, these models also represent recent advances in topic modeling and machine learning, whose complexity might prevent people who do not have the relevant technical training and experience from using them. As Hanna Wallach observed (personal communication), for researchers who actually use this type of models as part of their research workflow, they need enough understanding to talk confidently about the output of the models and what

should and should not be trusted.

We acknowledge that this is one major issue for the type of work done in this thesis. While we do not have an absolute solution for it, we would like to discuss a few directions that might help mitigate the problem. First, although users of these models might not need to understand every technical detail, basic modeling structures, especially in the outputs of the models, should be made approachable. One effective way which has benefited us greatly is to develop simple *visualizations* to display the models' outputs. For example, we built an interactive visualization called *Argviz* to show the outputs of SITS in Chapter 3 and generated simple HTML pages to display the topic hierarchies learned by SHLDA in Chapter 5 and HIPTM in Chapter 6. In developing these visualizations, we had to iterate with social scientists multiple times to get feedbacks and make modifications to how the information should be displayed. It is through these interactions that help us—the computer scientists—refine the visualizations to make the models' outputs more accessible, and help social scientists understand them better.

Second, despite the fact that various parts of this thesis are the result of collaborations with social scientists in communication and political science, one major goal of the introduced models is *predictive*, which is a common practice in machine learning research. However, using the models and their outputs to provide explanations for the observed data as well as pre-registered questions and hypotheses in social science is also important, especially for social scientists. This contrast in research interests between computer scientists and social scientists has been captured nicely in Hopkins and King (2010)'s argument: “[C]omputer scientists may be interested

in finding the needle in the haystack [...], but social scientists are more commonly interested in characterizing the haystack.” Therefore, one important direction for this dissertation’s future work is to focus more on the explanation tasks by strengthening the collaboration with social scientists which, as argued by many before, is crucial for interdisciplinary research that involves social science (O’Connor, 2014; Wallach, 2014; Grimmer, 2015).

Bibliography

- Abbott, R., Walker, M., Anand, P., Fox Tree, J. E., Bowmani, R., and King, J. (2011). How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Language in Social Media (LSM)*.
- Adams, R., Ghahramani, Z., and Jordan, M. (2010). Tree-structured stick breaking for hierarchical data. In *Proceedings of Advances in Neural Information Processing Systems*, pages 19–27.
- Adler, E. S. and Wilkerson, J. (2006). Congressional Bills Project. NSF 00880066 and 00880061.
- Adler, E. S. and Wilkerson, J. D. (2013). *Congress and the Politics of Problem Solving*. Cambridge University Press.
- Ahlberg, C. (1996). Spotfire: an information exploration environment. *ACM SIGMOD Record*, 25(4):25–29.
- Ahmed, A., Hong, L., and Smola, A. (2013a). Nested Chinese restaurant franchise process: Applications to user tracking and document modeling. In *Proceedings of the International Conference of Machine Learning*, pages 1426–1434.
- Ahmed, A., Hong, L., and Smola, A. J. (2013b). Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of World Wide Web Conference*, pages 25–36.
- Ahmed, A. and Xing, E. P. (2008). Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: with applications to evolutionary clustering. In *Proceedings of SIAM International Conference on Data Mining*, pages 219–230.
- Ahmed, A. and Xing, E. P. (2010a). Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 1140–1150.

- Ahmed, A. and Xing, E. P. (2010b). Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of Uncertainty in Artificial Intelligence*.
- Alarcon-del Amo, M., Lorenzo-Romero, C., and Gomez-Borja, M. (2011). Classifying and profiling social networking site users: a latent segmentation approach. *Cyberpsychology, Behavior, and Social Networking*, 14(9).
- Aletras, N. and Stevenson, M. (2013a). Evaluating topic coherence using distributional semantics. In *Proceedings of the International Conference on Computational Semantics*, pages 13–22.
- Aletras, N. and Stevenson, M. (2013b). Representing topics using images. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 158–167.
- Aletras, N. and Stevenson, M. (2014). Labelling topics using unsupervised graph-based methods. In *Proceedings of the Association for Computational Linguistics*, pages 631–636.
- Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43.
- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174.
- Aran, O. and Gatica-Perez, D. (2010). Fusing audio-visual nonverbal cues to detect dominant people in group conversations. In *Proceedings of the International Conference on Pattern Recognition*, pages 3687–3690.
- Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4).
- Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of Uncertainty in Artificial Intelligence*.
- Bafumi, J., Gelman, A., Park, D. K., and Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13(2):171–187.
- Bakalov, A., McCallum, A., Wallach, H., and Mimno, D. (2012). Topic models for taxonomies. In *Proceedings of Joint Conference on Digital Libraries*, pages 237–240.
- Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, Proceedings of World Wide Web Conference, pages 519–528.

- Bales, R. F. (1970). *Personality and interpersonal behavior*. Holt, Rinehart, and Winston.
- Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91.
- Barreto, M. A., Cooper, B. L., Gonzalez, B., Parker, C. S., and Towler, C. (2011). The Tea Party in the age of Obama: Mainstream conservatism or out-group anxiety? *Political Power and Social Theory*, 22(1):105–137.
- Basu, S., Choudhury, T., Clarkson, B., and Pentland, A. S. (2001). Learning human interactions with the influence model. Technical Report 539, MIT Media Laboratory.
- Baumgartner, F. and Jones, B. (1993a). Policy Agendas project. NSF 9320922 and 0111611.
- Baumgartner, F. R. (2001). Political agendas. In Smelser, N. J. and Baltes, P. B., editors, *International Encyclopedia of Social and Behavioral Sciences: Political Science*, pages 288–90. New York: Elsevier Science and Oxford: Pergamon.
- Baumgartner, F. R., De Boef, S. L., and Boydston, A. E. (2008). *The decline of the death penalty and the discovery of innocence*. Cambridge University Press.
- Baumgartner, F. R., Green-Pedersen, C., and Jones, B. D. (2006). Comparative studies of policy agendas. *Journal of European Public Policy*, 13(7):959–974.
- Baumgartner, F. R. and Jones, B. D. (1993b). *Agendas and instability in American politics*. University of Chicago Press.
- Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. *Machine Learning*, 34.
- Bender, E. M., Morgan, J. T., Oxley, M., Zachry, M., Hutchinson, B., Marin, A., Zhang, B., and Ostendorf, M. (2011). Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media*, pages 48–57.
- Benoit, W. L., Hansen, G. J., and Verser, R. M. (2003). A meta-analysis of the effects of viewing U.S. presidential debates. *Communication Monographs*, 70(4):335–350.
- Benoit, W. L., McKinney, M. S., and Stephenson, M. T. (2002). Effects of watching primary debates in the 2000 U.S. presidential campaign. *Journal of Communication*, 52(2):316–331.
- Beseler, C. L., Taylor, L. A., and Leeman, R. F. (2010). An item-response theory analysis of dsm-iv alcohol-use disorder criteria and binge drinking in undergraduates. *Journal of Studies on Alcohol and Drugs*, 71(3):418.

- Biran, O., Rosenthal, S., Andreas, J., McKeown, K., and Rambow, O. (2012). Detecting influencers in written online conversations. In *Proceedings of the Workshop on Language in Social Media (LSM)*.
- Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The annals of statistics*, pages 353–355.
- Blais, A. and Perrella, A. M. (2008). Systemic effects of televised candidates’ debates. *The International Journal of Press/Politics*, 13(4):451–464.
- Blau, P. (1964). *Exchange and power in social life*. Sociology political science. Transaction Books.
- Blei, D., Carin, L., and Dunson, D. (2010a). Probabilistic topic models. *Signal Processing Magazine, IEEE*, 27(6):55–65.
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.
- Blei, D. M. (2014). Build, compute, critique, repeat: data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232.
- Blei, D. M. and Frazier, P. I. (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461–2488.
- Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010b). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7.
- Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2003a). Hierarchical topic models and the nested Chinese restaurant process. In *Proceedings of Advances in Neural Information Processing Systems*, pages 17–24.
- Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International Conference of Machine Learning*, pages 113–120.
- Blei, D. M. and McAuliffe, J. D. (2007). Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 121–128.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003b). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298.
- Bonica, A. (2014). Mapping the ideological marketplace. *American Journal of Political Science*, 58(2):367–386.

- Booth, N. and Matic, A. (2011). Mapping and leveraging influencers in social media to shape corporate brand perceptions. *Corporate Communications*, 16(3).
- Boyd-Graber, J. and Blei, D. M. (2008a). Syntactic topic models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 185–192.
- Boyd-Graber, J. and Blei, D. M. (2008b). Syntactic topic models. In *Proceedings of Advances in Neural Information Processing Systems*.
- Boyd-Graber, J. and Blei, D. M. (2009). Multilingual topic models for unaligned text. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 75–82.
- Boyd-Graber, J. and Resnik, P. (2010). Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 45–55.
- Boydston, A. and Gross, J. (2014). Policy Frames Codebook. Work in progress.
- Boydston, A. E., Glazier, R. A., and Phillips, C. (2013a). Agenda control in the 2008 presidential debates. *American Politics Research*, 41(5):863–899.
- Boydston, A. E., Glazier, R. A., and Pietryka, M. T. (2013b). Playing to the crowd: Agenda control in presidential debates. *Political Communication*, 30(2):254–277.
- Boydston, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2013c). Identifying media frames and frame dynamics within and across policy issues. In *New Directions in Analyzing Text as Data Workshop*.
- Bragg, J., Mausam, and Weld, D. S. (2013). Crowdsourcing multi-label classification for taxonomy creation. In *Proceedings of Conference on Human Computation and Crowdsourcing*.
- Brooke, M. E. and Ng, S. H. (1986). Language and Social Influence in Small Conversational Groups. *Journal of Language and Social Psychology*, 5(3).
- Burrell, N. A. and Koper, R. J. (1998). The efficacy of powerful/powerless language on attitudes and source credibility. In *Persuasion: Advances through meta-analysis*. Hampton Press Cresskill, NJ.
- Butler, B., Joyce, E., and Pike, J. (2008). Don’t look now, but we’ve created a bureaucracy: the nature and roles of policies and rules in Wikipedia. In *International Conference on Human Factors in Computing Systems*, pages 1101–1110.
- Cardie, C. and Wilkerson, J. (2008). Text annotation for political science research. *Journal of Information Technology & Politics*, 5(1):1–6.
- Carmines, E. G. and D’Amico, N. J. (2015). The new look in political ideology research. *Annual Review of Political Science*, 18(4).

- Carroll, R., Lewis, J. B., Lo, J., Poole, K. T., and Rosenthal, H. (2009). Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap. *Political Analysis*, 17(3):261–275.
- Chambers, A., Smyth, P., and Steyvers, M. (2010). Learning concept graphs from text with stick-breaking priors. In *Proceedings of Advances in Neural Information Processing Systems*, pages 334–342.
- Chaney, A. J.-B. and Blei, D. M. (2012). Visualizing topic models. In *International Conference on Weblogs and Social Media*, pages 419–422.
- Chang, J. (2012). lda: Collapsed Gibbs sampling methods for topic models. <http://cran.r-project.org/web/packages/lda/index.html>. [Online; accessed 02-June-2014].
- Chang, J., Blei, D. M., et al. (2010). Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150.
- Chang, J., Boyd-Graber, J., and Blei, D. M. (2009a). Connections between the lines: augmenting social networks with text. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 169–178.
- Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. (2009b). Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 288–296.
- Chen, H., Branavan, S., Barzilay, R., and Karger, D. R. (2009). Global models of document structure using latent permutations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 371–379.
- Chilton, L. B., Little, G., Edge, D., Weld, D. S., and Landay, J. A. (2013). Cascade: Crowdsourcing taxonomy creation. In *International Conference on Human Factors in Computing Systems*, pages 1999–2008.
- Choi, F. Y. Y., Wiemer-Hastings, P., and Moore, J. (2001). Latent semantic analysis for text segmentation. In *Proceedings of Empirical Methods in Natural Language Processing*.
- Chong, D. and Druckman, J. N. (2007). Framing theory. *Annual Review of Political Science*, 10:103–126.
- Chu, Y.-J. and Liu, T.-H. (1965). On shortest arborescence of a directed graph. *Scientia Sinica*, 14(10):1396–1400.
- Chuang, J., Manning, C. D., and Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77.

- Clement, S. and Green, J. C. (2011). The Tea Party, religion, and social issues. <http://pewresearch.org/pubs/1903/tea-party-movement-religion-social-issues-conservative-christian>. Pew Forum on Religion and Public Life.
- Clinton, J., Jackman, S., and Rivers, D. (2004). The statistical analysis of roll call data. *American Political Science Review*, 98(02):355–370.
- Cohen, B. C. (1963). *The press and foreign policy*. Princeton University Press.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3).
- Cowans, P. J. (2006). *Probabilistic Document Modelling*. PhD thesis, University of Cambridge.
- Crespin, M. H. and Rohde, D. W. (2010). Dimensions, issues, and bills: Appropriations voting on the House floor. *The Journal of Politics*, 72(4):976–989.
- Daley, J. A., McCroskey, J. C., and Richmond, V. P. (1977). Relationships between vocal activity and perception of communicators in small group interaction. *Western Journal of Speech Communication*, 41(3).
- Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., and Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of World Wide Web Conference*, pages 699–708.
- Dardis, F. E., Baumgartner, F. R., Boydston, A. E., De Boef, S., and Shen, F. (2008). Media framing of capital punishment and its impact on individuals’ cognitive responses. *Mass Communication & Society*, 11(2):115–140.
- Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the International Conference of Machine Learning*, pages 233–240.
- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255.
- Deng, J., Russakovsky, O., Krause, J., Bernstein, M. S., Berg, A., and Fei-Fei, L. (2014). Scalable multi-label annotation. In *International Conference on Human Factors in Computing Systems*, pages 3099–3102.
- Dou, W., Yu, L., Wang, X., Ma, Z., and Ribarsky, W. (2013). Hierarchical topics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2002–2011.

- Dowman, M., Savova, V., Griffiths, T. L., Kording, K., Tenenbaum, J. B., and Purver, M. (2008). A probabilistic model of meetings that combines words and discourse features. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1238–1248.
- Drake, B. H. and Moberg, D. J. (1986). Communicating influence attempts in dyads: Linguistic sedatives and palliatives. *The Academy of Management Review*, 11(3).
- Du, L., Buntine, W., and Jin, H. (2012). Modelling sequential text with an adaptive topic model. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 535–545.
- Du, L., Buntine, W. L., and Jin, H. (2010). Sequential latent Dirichlet allocation: Discover underlying topic structures within a document. In *Proceedings of International Conference on Data Mining*, pages 148–157.
- Du, L., Buntine, W. L., and Johnson, M. (2013). Topic segmentation with a structured topic model. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 190–200.
- Eagle, N., Pentland, A. S., and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278.
- Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71(4):233–240.
- Ehlen, P., Purver, M., and Niekrasz, J. (2007). A meeting browser that learns. In *AAAI Spring Symposium: Interaction Challenges for Intelligent Assistants*, pages 33–40.
- Eisenstein, J., Ahmed, A., and Xing, E. P. (2011). Sparse additive generative models of text. In *Proceedings of the International Conference of Machine Learning*, pages 1041–1048.
- Eisenstein, J. and Barzilay, R. (2008). Bayesian unsupervised topic segmentation. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 334–343.
- Eisenstein, J., Chau, D. H., Kittur, A., and Xing, E. (2012). TopicViz: Interactive topic exploration in document collections. In *International Conference on Human Factors in Computing Systems*, pages 2177–2182.
- Eisenstein, J., O’Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 1277–1287.
- Emerson, R. M. (1981). Social exchange theory. In Rosenberg, M. and Turner, R. H., editors, *Social Psychology: Sociological Perspectives*, pages 30–65. Basic Books, New York.

- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.
- Erosheva, E., Fienberg, S., and Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5220–5227.
- Fallows, J. (2008). Your VP debate wrapup in four bullet points. *The Atlantic*. <http://www.theatlantic.com/technology/archive/2008/10/your-vp-debate-wrapup-in-four-bullet-points/8887/>.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2).
- Feske, U., Kirisci, L., Tarter, R. E., and Pilkonis, P. A. (2007). An application of item response theory to the DSM-III-R criteria for borderline personality disorder. *Journal of Personality Disorders*, 21(4):418–433.
- Fleischmann, K. R., Templeton, T. C., and Boyd-Graber, J. (2011). Modeling diverse standpoints in text classification: Learning to be human by modeling human values. In *Proceedings of the iConference*, pages 672–673.
- Foa, U. G. and Foa, E. B. (1972). Resource exchange: Toward a structural theory of interpersonal communication. In Siegman, A. W. and Pope, B., editors, *Studies in dyadic communication*, pages 291–325. Pergamon Press.
- Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2008). An HDP-HMM for systems with state persistence. In *Proceedings of the International Conference of Machine Learning*, pages 312–319.
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer.
- Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the Association for Computational Linguistics*, pages 562–569.
- Gardner, M. J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., and Seppi, K. (2010). The topic browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization*.
- Geer, J. G. (1988). The effects of presidential debates on the electorate’s preferences for candidates. *American Politics Research*, 16(4):486–501.
- Gerrish, S. and Blei, D. M. (2011). Predicting legislative roll calls from text. In *Proceedings of the International Conference of Machine Learning*, pages 489–496.
- Gerrish, S. and Blei, D. M. (2012). How they vote: Issue-adjusted models of legislative behavior. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2753–2761.

- Gershman, S. J. and Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1).
- Gervais, B. T. and Morris, I. L. (2012). Reading the tea leaves: Understanding Tea Party Caucus membership in the US House of Representatives. *PS: Political Science & Politics*, 45(02):245–250.
- Gervais, B. T. and Morris, I. L. (2014). Black Tea, Green Tea, White Tea, and Coffee: Understanding the variation in attachment to the tea party among members of Congress. In *Annual Meeting of the American Political Science Association*.
- Ghanem, S. (1997). Filling in the tapestry: The second level of agenda setting. *Communication and democracy: Exploring the intellectual frontiers in agenda-setting theory*, pages 3–14.
- Giles, H., Coupland, N., and Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In Giles, H., Coupland, N., and Coupland, J., editors, *Contexts of accommodation: Developments in applied socio-linguistics*, pages 1–68. Cambridge University Press.
- Giles, J. (2012). Computational social science: Making the links. *Nature*, 488(7412):448–450.
- Gretarsson, B., Odonovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., and Smyth, P. (2012). Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology*, 3(2):23.
- Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. (2004). Integrating topics and syntax. In *Proceedings of Advances in Neural Information Processing Systems*, pages 537–544.
- Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1):1–35.
- Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48(01):80–83.
- Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.
- Grofman, B. and Brazill, T. J. (2002). Identifying the median justice on the Supreme Court through multidimensional scaling: Analysis of “natural courts” 1953–1991. *Public Choice*, 112(1-2):55–79.
- Gu, Y., Sun, Y., Jiang, N., Wang, B., and Chen, T. (2014). Topic-factorized ideal point estimation model for legislative voting network. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 183–192.

- Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman, New York.
- Hamilton, M. A. and Hunter, J. E. (1998). The effect of language intensity on receiver evaluations of message, source, and topic. In *Persuasion: Advances through meta-analysis*. Hampton Press Cresskill, NJ.
- Hardisty, E., Boyd-Graber, J., and Resnik, P. (2010). Modeling perspective using adaptor grammars. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 284–292.
- Hariharan, B., Vishwanathan, S. V., and Varma, M. (2012). Efficient max-margin multi-label classification with applications to zero-shot learning. *Machine Learning*, 88(1-2):127–155.
- Hawes, T., Lin, J., and Resnik, P. (2009). Elements of a computational model for multi-party discourse: The turn-taking behavior of Supreme Court justices. *Journal of the American Society for Information Science and Technology*, 60(8).
- Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1).
- Heckman, J. J. and Jr., J. M. S. (1997). Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *The RAND Journal of Economics*, 28:142–189.
- Heymann, P. and Garcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford InfoLab.
- Hillard, D., Purpura, S., and Wilkerson, J. (2007). An active learning framework for classifying political text. In *Annual meeting of the Midwest Political Science Association*.
- Hillard, D., Purpura, S., and Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46.
- Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3).
- Hix, S., Noury, A., and Roland, G. (2006). Dimensions of politics in the European Parliament. *American Journal of Political Science*, 50(2):494–520.
- Ho, Q., Eisenstein, J., and Xing, E. P. (2012). Document hierarchies from text and links. In *Proceedings of World Wide Web Conference*, pages 739–748.
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

- Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 289–296.
- Holbrook, T. M. (1999). Political learning from presidential debates. *Political Behavior*, 21(1):67–89.
- Hopkins, D. J. and King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.
- Horton, S. (2008). The Ifill factor. *Harpers Magazine*.
- Hsueh, P.-Y., Moore, J. D., and Renals, S. (2006). Automatic segmentation of multiparty dialogue. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.
- Huffaker, D. (2010). Dimensions of leadership and social influence in online communities. *Human Communication Research*, 36(4):593–617.
- Hung, H., Huang, Y., Friedland, G., and Gatica-Perez, D. (2011). Estimating dominance in multi-party meetings using speaker diarization. *Transactions on Audio, Speech, and Language Processing*, 19(4).
- Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., and Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1).
- Jackman, S. (2001). Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis*, 9(3):227–241.
- Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The ICSI meeting corpus. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 364–367.
- Jayagopi, D. B., Hung, H., Yeo, C., and Gatica-Perez, D. (2009). Modeling dominance in group conversations using nonverbal activity cues. *Transactions on Audio, Speech, and Language Processing*, 17(3).
- Jiang, Q., Zhu, J., Sun, M., and Xing, E. P. (2012). Monte Carlo methods for maximum margin supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1592–1600.
- Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of ACM International Conference on Web Search and Data Mining*, pages 219–230.
- Jo, Y. and Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of ACM International Conference on Web Search and Data Mining*, pages 815–824.

- Joachims, T. (1999). Making large-scale SVM learning practical. In *Advances in Kernel Methods - SVM*. Universität Dortmund.
- John, P. (2006). The Policy Agendas Project: A review. *Journal of European Public Policy*, 13(7):975–986.
- Jones, B. D. and Baumgartner, F. R. (2005). *The politics of attention: How government prioritizes problems*. University of Chicago Press.
- Joty, S., Carenini, G., Murray, G., and Ng, R. T. (2010). Exploiting conversation structure in unsupervised topic segmentation for emails. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 388–398. Association for Computational Linguistics.
- Joty, S. R., Carenini, G., Murray, G., and Ng, R. T. (2011). Supervised topic segmentation of email conversations. In *International Conference on Weblogs and Social Media*, pages 530–533.
- Joty, S. R., Carenini, G., and Ng, R. T. (2013). Topic segmentation and labeling in asynchronous conversations. *Journal of Artificial Intelligence Research*, 47:521–573.
- Karpowitz, C. F., Monson, J. Q., Patterson, K. D., and Pope, J. C. (2011). Tea time in America? The impact of the Tea Party movement on the 2010 midterm elections. *PS: Political Science & Politics*, 44(02):303–309.
- Katz, E. and Lazarsfeld, P. F. (1955). *Personal influence: the part played by people in the flow of mass communications*. Foundations of communications research. Free Press.
- Kim, J. H., Kim, D., Kim, S., and Oh, A. (2012). Modeling topic hierarchies with the recursive Chinese restaurant process. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 783–792.
- Koford, K. (1989). Dimensions in congressional voting. *The American Political Science Review*, pages 949–962.
- Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*.
- Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage, 3rd edition.
- Kwon, N., Zhou, L., Hovy, E., and Shulman, S. W. (2007). Identifying and classifying subjective claims. In *Proceedings of International Conference on Digital Government Research: Bridging Disciplines & Domains*, pages 76–81. Digital Government Society of North America.

- Lacoste-Julien, S., Sha, F., and Jordan, M. I. (2008). DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Proceedings of Advances in Neural Information Processing Systems*, pages 897–904.
- Lakoff, G. (2010). Why it matters how we frame the environment. *Environmental Communication*, 4(1):70–81.
- Lau, J. H., Grieser, K., Newman, D., and Baldwin, T. (2011). Automatic labelling of topic models. In *Proceedings of the Association for Computational Linguistics*, pages 1536–1545.
- Lau, J. H., Newman, D., and Baldwin, T. (2014a). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, page 530539.
- Lau, J. H., Newman, D., and Baldwin, T. (2014b). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the Association for Computational Linguistics*, pages 530–539.
- Lau, J. H., Newman, D., Karimi, S., and Baldwin, T. (2010). Best topic word selection for topic labelling. In *Proceedings of International Conference on Computational Linguistics*, pages 605–613.
- Lauderdale, B. E. and Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, 58(3):754–771.
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915):721–723.
- Ledgerwood, A. and Boydstun, A. E. (2014). Sticky prospects: Loss frames are cognitively stickier than gain frames. *Journal of Experimental Psychology: General*, 143(1):376.
- Leskovec, J. (2008). *Dynamics of large networks*. PhD thesis, Carnegie Mellon University.
- Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.
- Lewis, J. B. and Poole, K. T. (2004). Measuring bias and uncertainty in ideal point estimates via the parametric bootstrap. *Political Analysis*, 12(2):105–127.
- Li, A. Q., Ahmed, A., Ravi, S., and Smola, A. J. (2014). Reducing the sampling complexity of topic models. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 891–900.

- Li, W., Blei, D. M., and McCallum, A. (2007). Nonparametric Bayes Pachinko allocation. In *Proceedings of Uncertainty in Artificial Intelligence*.
- Li, W. and McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the International Conference of Machine Learning*, pages 577–584.
- Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., and Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 939–948.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.
- Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.
- Liu, X., Song, Y., Liu, S., and Wang, H. (2012). Automatic taxonomy construction from keywords. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 1433–1441.
- Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*.
- Mackay, D. J. C. and Peto, L. C. B. (1995). A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):289–308.
- Madjarov, G., Kocev, D., Gjorgjevikj, D., and Deroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104.
- Magatti, D., Calegari, S., Ciucci, D., and Stella, F. (2009). Automatic labeling of topics. In *International Conference on Intelligent Systems Design and Applications*, pages 1227–1232.
- Mao, X.-L., Ming, Z.-Y., Chua, T.-S., Li, S., Yan, H., and Li, X. (2012a). SSHLDA: a semi-supervised hierarchical topic model. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 800–809.
- Mao, X.-L., Ming, Z.-Y., Zha, Z.-J., Chua, T.-S., Yan, H., and Li, X. (2012b). Automatic labeling hierarchical topics. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 2383–2386.

- Martin, A. D. and Quinn, K. M. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999. *Political Analysis*, 10(2):134–153.
- Mast, M. S. (2002). Dominance as expressed and inferred through speaking time. *Human Communication Research*, 28(3):420–450.
- Matthes, J. and Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, 58(2):258–279.
- Maxwell, A. and Parent, T. W. (2012). The Obama trigger: Presidential approval and Tea Party membership. *Social Science Quarterly*, 93(5):1384–1401.
- McCarty, N. M., Poole, K. T., and Rosenthal, H. (1997). *Income redistribution and the realignment of American politics*. AEI Press publisher for the American Enterprise Institute.
- McCombs, M. (2004). *Setting the agenda: The mass media and public opinion*. John Wiley & Sons.
- McCombs, M. (2005). A look at agenda-setting: Past, present and future. *Journalism Studies*, 6(4):543–557.
- McCombs, M. and Ghanem, S. I. (2001). The convergence of agenda setting and framing. *Framing public life: Perspectives on media and our understanding of the social world*, pages 67–81.
- McCombs, M., Llamas, J. P., Lopez-Escobar, E., and Rey, F. (1997). Candidate images in Spanish elections: Second-level agenda-setting effects. *Journalism & Mass Communication Quarterly*, 74(4):703–717.
- McCombs, M. E. and Shaw, D. L. (1972). The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.
- McCombs, M. E. and Shaw, D. L. (1993). The evolution of agenda-setting research: Twenty-five years in the marketplace of ideas. *Journal of Communication*, 43(2):58–67.
- McKinney, M. S. and Carlin, D. B. (2004). Political campaign debates. *Handbook of Political Communication Research*, pages 203–234.
- McKinney, M. S. and Warner, B. R. (2013). Do presidential debates matter? Examining a decade of campaign debate effects. *Argumentation and Advocacy*, 49(4).
- Mei, Q., Shen, X., and Zhai, C. (2007). Automatic labeling of multinomial topic models. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 490–499.

- Mimno, D., Li, W., and McCallum, A. (2007). Mixtures of hierarchical topics with Pachinko allocation. In *Proceedings of the International Conference of Machine Learning*, pages 633–640.
- Mimno, D. and McCallum, A. (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 500–509.
- Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 880–889.
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 262–272.
- Mimno, D. M. and McCallum, A. (2008). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of Uncertainty in Artificial Intelligence*.
- Monroe, B. L. and Schrodtt, P. A. (2008). Introduction to the special issue: The statistical analysis of political text. *Political Analysis*, 16(4):351–355.
- Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.
- Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, 19(1).
- Murray, G., Renals, S., and Carletta, J. (2005). Extractive summarization of meeting recordings. In *European Conference on Speech Communication and Technology*.
- Nallapati, R. M., Ahmed, A., Xing, E. P., and Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 542–550.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.
- Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 31:705–767.
- Nelson, T. E., Clawson, R. A., and Oxley, Z. M. (1997). Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, 91(3):567–583.
- Newman, D., Chemudugunta, C., and Smyth, P. (2006). Statistical entity-topic models. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 680–686.

- Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.
- Ng, S. H., Bell, D., and Brooke, M. (1993). Gaining turns and achieving high influence ranking in small conversational groups. *British Journal of Social Psychology*, 32(3):265–275.
- Ng, S. H. and Bradac, J. J. (1993). *Power in language: verbal communication and social influence*. Language and language behaviors. Sage Publications.
- Nguyen, V.-A., Boyd-Graber, J., and Altschul, S. F. (2013a). Dirichlet mixtures, the Dirichlet process, and the structure of protein space. *Journal of Computational Biology*, 20(1):1–18.
- Nguyen, V.-A., Boyd-Graber, J., Chang, J., and Resnik, P. (2013b). Tree-based label dependency topic models. In *NIPS Workshop on Topic Models Computation, Application, and Evaluation*.
- Nguyen, V.-A., Boyd-Graber, J., and Resnik, P. (2012). SITS: A hierarchical non-parametric model using speaker identity for topic segmentation in multiparty conversations. In *Proceedings of the Association for Computational Linguistics*, pages 78–87.
- Nguyen, V.-A., Boyd-Graber, J., and Resnik, P. (2013c). Lexical and hierarchical topic regression. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1106–1114.
- Nguyen, V.-A., Boyd-Graber, J., and Resnik, P. (2014a). Sometimes average is best: The importance of averaging for prediction using MCMC inference in topic modeling. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 1752–1757.
- Nguyen, V.-A., Boyd-Graber, J., Resnik, P., Cai, D. A., Midberry, J. E., and Wang, Y. (2014b). Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95(3):381–421.
- Nguyen, V.-A., Boyd-Graber, J., Resnik, P., and Chang, J. (2014c). Learning a concept hierarchy from multi-labeled documents. In *Proceedings of Advances in Neural Information Processing Systems*, pages 3671–3679.
- Nguyen, V.-A., Hu, Y., Boyd-Graber, J., and Resnik, P. (2013d). Argviz: Interactive visualization of topic dynamics in multi-party conversations. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 36–39.

- Nielsen (2012). Final presidential debate draws 59.2 million viewers. <http://www.nielsen.com/us/en/insights/news/2012/final-presidential-debate-draws-59-2-million-viewers.html>. [Online; accessed 31-December-2014].
- Nikolova, S., Boyd-Graber, J., and Fellbaum, C. (2012). Collecting semantic similarity ratings to connect concepts in assistive communication tools. In *Modeling, Learning, and Processing of Text Technological Data Structures*, pages 81–93. Springer.
- North, C. and Shneiderman, B. (2000). Snap-together visualization: a user interface for coordinating visualizations via relational schemata. In *Proceedings of the working conference on Advanced visual interfaces*, pages 128–135.
- O’Connor, B. (2014). *Statistical Text Analysis for Social Science*. PhD thesis, Carnegie Mellon University.
- Olney, A. and Cai, Z. (2005). An orthonormal basis for topic segmentation in tutorial dialogue. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 971–978.
- Otsuka, K., Yamato, J., Takemae, Y., and Murase, H. (2006). Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *International Conference on Human Factors in Computing Systems*, pages 1175–1180.
- Paisley, J., Wang, C., Blei, D. M., and Jordan, M. I. (2014). Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Palmer, M. T. (1989). Controlling conversations: Turns, topics and interpersonal control. *Communication Monographs*, 56(1):1–18.
- Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Association for Computational Linguistics*, pages 115–124.
- Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.
- Papadimitriou, C. H., Tamaki, H., Raghavan, P., and Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS, pages 159–168.
- Passonneau, R. J. and Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1).

- Paul, M. and Dredze, M. (2012). Factorial LDA: Sparse multi-dimensional text models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2582–2590.
- Paul, M. J. and Dredze, M. (2015). SPRITE: Generalizing topic models with structured priors. *Transactions of the Association for Computational Linguistics*, 3:43–57.
- Paul, M. J. and Girju, R. (2010). A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Association for the Advancement of Artificial Intelligence*, pages 545–550.
- Paul, R. (2007). Political power and the rule of law. *Texas Straight Talk*.
- Pele, O. and Werman, M. (2008). A linear time histogram metric for improved sift matching. In *Proceedings of the European Conference on Computer Vision*, pages 495–508.
- Pele, O. and Werman, M. (2009). Fast and robust earth mover’s distances. In *International Conference on Computer Vision*, pages 460–467.
- Perer, A. and Shneiderman, B. (2006). Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):693–700.
- Perotte, A. J., Wood, F., Elhadad, N., and Bartlett, N. (2011). Hierarchically supervised latent Dirichlet allocation. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2609–2617.
- Petinot, Y., McKeown, K., and Thadani, K. (2011). A hierarchical model of web summaries. In *Proceedings of the Association for Computational Linguistics*, pages 670–675.
- Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28.
- Planalp, S. and Tracy, K. (1980). Not to change the topic but ...: A cognitive approach to the management of conversations. In *Communication yearbook 4*, pages 237–258. New Brunswick.
- Plangprasopchok, A. and Lerman, K. (2009). Constructing folksonomies from user-specified relations on Flickr. In *Proceedings of World Wide Web Conference*, pages 781–790.
- Podesta, J., Pritzker, P., Moniz, E., Holdren, J., and Zients, J. (2014). Big data: seizing opportunities, preserving values. *Executive Office of the President, The White House, Washington*.

- Poole, K. T. (1998). Recovering a basic space from a set of issue scales. *American Journal of Political Science*, 42(3):954–993.
- Poole, K. T. and Daniels, R. S. (1985). Ideology, party, and voting in the U.S. Congress, 1959-1980. *The American Political Science Review*, pages 373–399.
- Poole, K. T. and Rosenthal, H. (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384.
- Poole, K. T. and Rosenthal, H. (1987). Analysis of congressional coalition patterns: A unidimensional spatial model. *Legislative Studies Quarterly*, pages 55–75.
- Poole, K. T. and Rosenthal, H. (1991). Patterns of congressional voting. *American Journal of Political Science*, pages 228–278.
- Poole, K. T. and Rosenthal, H. (1997). *Congress: A political-economic history of roll call voting*. Oxford University Press.
- Poole, K. T. and Rosenthal, H. (2001). D-NOMINATE after 10 years: A comparative update to Congress: A political-economic history of roll-call voting. *Legislative Studies Quarterly*, pages 5–29.
- Poole, K. T., Rosenthal, H., and Koford, K. (1991). On dimensionalizing roll call votes in the U.S. Congress. *The American Political Science Review*, 85(3):955–976.
- Poole, K. T. and Rosenthal, H. L. (2007). *Ideology and Congress*. New Brunswick, NJ: Transaction Publishers.
- Purpura, S., Cardie, C., and Simons, J. (2008). Active learning for e-rulemaking: Public comment categorization. In *Proceedings of International Conference on Digital Government Research*, pages 234–243.
- Purpura, S. and Hillard, D. (2006). Automated classification of congressional legislation. In *Proceedings of the 2006 international conference on Digital government research*, pages 219–225. Digital Government Society of North America.
- Purver, M. (2011). Topic segmentation. In *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley & Sons.
- Purver, M., Griffiths, T. L., Körding, K. P., and Tenenbaum, J. B. (2006). Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 17–24.
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

- Racine Group (2002). White paper on televised political campaign debates. *Argumentation and Advocacy*, 38(4):199–218.
- Ramage, D., Dumais, S. T., and Liebling, D. J. (2010a). Characterizing microblogs with topic models. In *International Conference on Weblogs and Social Media*, pages 130–137.
- Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 248–256.
- Ramage, D., Manning, C. D., and Dumais, S. (2011). Partially labeled topic models for interpretable text mining. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 457–465.
- Ramage, D., Manning, C. D., and McFarland, D. A. (2010b). Which universities lead and lag? toward university rankings based on scholarly output. In *In NIPS Workshop on Computational Social Science and the Wisdom of the Crowds*.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 321–333. University of California Press Berkeley, CA.
- Regula, R. and Julian, W. (1973). The impact of quality and frequency of task contributions on perceived ability. *Journal of Social Psychology*, 89(1):115–122.
- Reid, S. A. and Ng, S. H. (2000). Conversation as a resource for influence: evidence for prototypical arguments and social identification processes. *European Journal of Social Psychology*, 30(1):83–100.
- Ren, L., Dunson, D. B., and Carin, L. (2008). The dynamic hierarchical Dirichlet process. In *Proceedings of the International Conference of Machine Learning*, pages 824–831.
- Resnik, P., Garron, A., and Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 1348–1353.
- Resnik, P. and Hardisty, E. (2010). Gibbs sampling for the uninitiated. Technical Report UMIACS-TR-2010-04, University of Maryland. <http://drum.lib.umd.edu/handle/1903/10058>.
- Rienks, R. and Heylen, D. (2006). Dominance detection in meetings using easily obtainable features. In Renals, S. and Bengio, S., editors, *Machine Learning for Multimodal Interaction*, volume 3869, pages 76–86. Springer.
- Rienks, R., Zhang, D., Gatica-Perez, D., and Post, W. (2006). Detection and application of influence rankings in small group meetings. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 257–264.

- Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 1524–1534.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*.
- Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of ACM International Conference on Web Search and Data Mining*.
- Rogers, E. M. and Dearing, J. W. (1988). Agenda-setting research: Where has it been, where is it going? In Anderson, J. A., editor, *Communication Yearbook(11)*, pages 555–594. Sage.
- Rogers, E. M., Dearing, J. W., and Bregman, D. (1993). The anatomy of agenda-setting research. *Journal of Communication*, 43(2):68–84.
- Rogers, T. and Norton, M. I. (2011). The artful dodger: Answering the wrong question the right way. *Journal of Experimental Psychology: Applied*, 17(2):139–147.
- Rose, J. S. and Dierker, L. C. (2010). An item response theory analysis of nicotine dependence symptoms in recent onset adolescent smokers. *Drug and Alcohol Dependence*, 110(1):70–79.
- Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 487–494.
- Rubin, T. N., Chambers, A., Smyth, P., and Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208.
- Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover’s distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2).
- Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using german online news. *Quality & Quantity*, 47(2):761–773.
- Schattschneider, E. E. (1960). *The Semi-Sovereign People: A Realist’s View of Democracy in America*. Holt, Rinehart and Winston.
- Scheer, L. K. and Stern, L. W. (1992). The effect of influence type and performance outcomes on attitude toward the influencer. *Journal of Marketing Research*, 29(1):128–142.

- Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122.
- Schlenker, B. R., Nacci, P., Helm, B., and Tedeschi, J. T. (1976). Reactions to coercive and reward power: The effects of switching influence modes on target compliance. *Sociometry*, 39(4):316–323.
- Schmitz, P. (2006). Inducing ontology from Flickr tags. In *WWW Collaborative Web Tagging Workshop*.
- Schroeder, A. (2008). *Presidential debates: Fifty years of high-risk TV*. Columbia University Press.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.
- Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.
- Sim, Y., Routledge, B., and Smith, N. A. (2015). The utility of text: The case of Amicus briefs and the Supreme Court. In *Association for the Advancement of Artificial Intelligence*.
- Singh, S., Wick, M., and McCallum, A. (2012). Monte Carlo MCMC: Efficient inference by approximate sampling. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 1104–1113.
- Skocpol, T. and Williamson, V. (2012). *The Tea Party and the remaking of Republican conservatism*. Oxford University Press.
- Slutsky, A., Hu, X., and An, Y. (2013a). Tree Labeled LDA: A hierarchical model for web summaries. In *Proceedings of International Conference on Big Data*, pages 134–140.
- Slutsky, A., Hu, X., and An, Y. (2013b). Tree labeled LDA: A hierarchical model for web summaries. In *Proceedings of International Conference on Big Data*, pages 134–140.
- Smola, A. and Narayanamurthy, S. (2010). An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710.
- Snyder Jr, J. M. (1992). Committee power, structure-induced equilibria, and roll call votes. *American Journal of Political Science*, 36(1):1–30.
- Sopan, A., Freier, M., Taieb-Maimon, M., Plaisant, C., Golbeck, J., and Shneiderman, B. (2013). Exploring data distributions: Visual design and evaluation. *International Journal of Human-Computer Interaction*, 29(2):77–95.

- Sorrentino, R. M. and Boutillier, R. G. (1975). The effect of quantity and quality of verbal interaction on ratings of leadership ability. *Journal of Experimental Social Psychology*, 11(5):403–411.
- Stang, D. J. (1973). Effect of interaction rate on ratings of leadership and liking. *Journal of Personality and Social Psychology*, 27(3):405–408.
- Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961.
- Steyvers, M. and Griffiths, T. (2006). Probabilistic topic models. In Landauer, T., Mcnamara, D., Dennis, S., and Kintsch, W., editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.
- Sudderth, E. B. (2006). *Graphical models for visual object recognition and tracking*. PhD thesis, Massachusetts Institute of Technology.
- Teh, Y. W. (2006). A hierarchical bayesian language model based on Pitman-Yor processes. In *Proceedings of the Association for Computational Linguistics*, pages 985–992.
- Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In Hjort, N., Holmes, C., Müller, P., and Walker, S., editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476).
- Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 327–335.
- Tibely, G., Pollner, P., Vicsek, T., and Palla, G. (2013). Extracting tag hierarchies. *PLoS ONE*, 8(12):e84133.
- Trammell, K. D. and Keshelashvili, A. (2005). Examining the new influencers: A self-presentation study of a-list blogs. *Journalism & Mass Communication Quarterly*, 82(4):968–982.
- Tsoumakas, G., Katakis, I., and Vlahavas, I. P. (2010). Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer.
- Tsutsumi, A., Iwata, N., Watanabe, N., De Jonge, J., Pikhart, H., Fernández-lópez, J. A., Xu, L., Peter, R., Knutsson, A., Niedhammer, I., et al. (2009). Application of item response theory to achieve cross-cultural comparability of occupational stress measurement. *International Journal of Methods in Psychiatric Research*, 18(1):58–67.

- Tur, G., Stolcke, A., Voss, L., Peters, S., Hakkani-Tur, D., Dowding, J., Favre, B., Fernández, R., Frampton, M., Frandsen, M., et al. (2010). The CALO meeting assistant system. *Transactions on Audio, Speech, and Language Processing*, 18(6):1601–1611.
- Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.
- Vavreck, L. (2009). *The message matters: the economy and presidential campaigns*. Princeton University Press.
- Verberne, S., Dhondt, E., van den Bosch, A., and Marx, M. (2014). Automatic thematic classification of election manifestos. *Information Processing & Management*, 50(4):554–567.
- Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428.
- Wallach, H. (2014). Big data, machine learning, and the social sciences: Fairness, accountability, and transparency. In *NIPS Workshop on Fairness, Accountability, and Transparency in Machine Learning*.
- Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the International Conference of Machine Learning*, pages 977–984.
- Wallach, H. M. (2008). *Structured Topic Models for Language*. PhD thesis, University of Cambridge.
- Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the International Conference of Machine Learning*, pages 1105–1112.
- Wang, C., Blei, D., and Li, F.-F. (2009). Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition*, pages 1903–1910.
- Wang, C., Blei, D. M., and Heckerman, D. (2008). Continuous time dynamic topic models. In *Proceedings of Uncertainty in Artificial Intelligence*.
- Wang, H., Lu, Y., and Zhai, C. (2010). Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 783–792.
- Wang, X. and McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 424–433.
- Wang Baldonado, M. Q., Woodruff, A., and Kuchinsky, A. (2000). Guidelines for using multiple views in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pages 110–119.

- Watts, D. J. (2007). A twenty-first century science. *Nature*, 445(7127):489.
- Weaver, C. (2004). Building highly-coordinated visualizations in improvise. In *IEEE Symposium on Information Visualization*, pages 159–166.
- Weimann, G. (1994). *The Influentials: People Who Influence People*. Suny Series in Human Communication Processes. State University of New York Press.
- Weng, L. (2014). *Information Diffusion on Online Social Networks*. PhD thesis, Indiana University.
- Wilcox, C. and Clausen, A. (1991). The dimensionality of roll-call voting reconsidered. *Legislative Studies Quarterly*, 16(3):393–406.
- Williamson, V., Skocpol, T., and Coggin, J. (2011). The Tea Party and the remaking of Republican conservatism. *Perspectives on Politics*, 9(01):25–43.
- Wolfe, M., Jones, B. D., and Baumgartner, F. R. (2013). A failure to communicate: Agenda setting in media and policy studies. *Political Communication*, 30(2):175–192.
- Yang, S.-H., Zha, H., and Hu, B.-G. (2009). Dirichlet-Bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2143–2150.
- Yu, B., Kaufmann, S., and Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48.
- Zhai, K., Boyd-Graber, J., Asadi, N., and Alkhouja, M. L. (2012). Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of World Wide Web Conference*, pages 879–888.
- Zhang, D., Gatica-Perez, D., Bengio, S., and Roy, D. (2005). Learning influence among interacting Markov chains. In *Proceedings of Advances in Neural Information Processing Systems*.
- Zhang, J. (2012). *Explore objects and categories in unexplored environments based on multimodal data*. PhD thesis, University of Hamburg.
- Zhang, M.-L. and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *Transactions on Knowledge and Data Engineering*, 26(8).
- Zhu, J., Ahmed, A., and Xing, E. P. (2012). MedLDA: Maximum margin supervised topic models. *Journal of Machine Learning Research*, 13(1):2237–2278.
- Zhu, J., Chen, N., Perkins, H., and Zhang, B. (2014a). Gibbs max-margin topic models with data augmentation. *Journal of Machine Learning Research*, 15:1073–1110.

- Zhu, J., Chen, N., Perkins, H., and Zhang, B. (2014b). Gibbs max-margin topic models with data augmentation. *Journal of Machine Learning Research*, 15(1):1073–1110.
- Zhu, J., Zheng, X., Zhou, L., and Zhang, B. (2013). Scalable inference in max-margin topic models. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 964–972.