Unexpected Partisan Unity Among Congressional Leaders and Legislators Using New Latent Variable Estimation Techniques and Frameworks

Thesis by
Daniel Ebanks

In Partial Fulfillment of the Requirements for the
Degree of
Doctorate of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2024
Defended September 5, 2023

ACKNOWLEDGEMENTS

I am extremely grateful to have the support of so many close friends, loving family, and constructive colleagues in my academic pursuits. Firstly, I am grateful for the advice, effort, and understanding of my advisors, Mike Alvarez and Jonathan N. Katz, for their unwavering belief in my ability to reach this important milestone. It is thanks to their support and encouragement that I have developed into the researcher I am today.

I would also like to thank my former colleagues at the Federal Reserve, especially Anton Badev, who encouraged me to engage in technical training, which then inspired my graduate studies and set me on my current path.

Thank you to my senior colleagues and coauthors, including Betsy Sinclair and Gary King, who have acted as continuous sources of inspiration, professional advice, and resources for me to achieve my research goals and to succeed in the field. They helped me learn how to develop new frameworks for methods research and push boundaries in the field. Additionally, I would like to thank my colleagues at Caltech CMS and Nvidia, including Bojan Tonguz, Jean Kossafi and Anima Anandkumar, who helped me engage in interdisciplinary research and learn how to develop methods on the frontiers of machine learning. In addition, I would like to thank Bob Sherman for his feedback and guidance as a member of my Committee.

I would also like to thank all of my colleagues at Caltech and coauthors, who have pushed me intellectually and taught me the value of collaborative research: Claudia Kann, Ramit Debnath, Shunto Kobayashi, Alena Buinskaya, Jacob Morrier, Melina Much, Sara Kangslanthi, Sreemanti Dey, Sarah Hashash, Joon Young Park.

To my life partner, Alec, who has always believed in me, even when I doubted my own abilities. Your consistent support as I try to realize my dreams and ambitions makes all the difference. To Matt, Nancy, Marcy, and Jayne, thank you for always cheerleading my work, providing a home away from home when I needed it, and always showing up to support me.

Finally, a special thank you to my family, Mom, Dad, and Alexa. You encouraged my curiosity, supported my education, and most importantly, you always let me be myself. Without your support, I could never have even dreamed of reaching for my goals. You keep me grounded, but you also taught me there is nothing I cannot achieve if I put my mind to it.

ABSTRACT

This dissertation is intended as a collection of essays which explore innovations in the development and estimation of latent variable models. These methods have many applications, including Natural Language Processing and latent correlation structures, which this dissertation explores. In addition to the statistical challenge of innovating on this class of model, latent variable methods are computationally demanding, requiring research insights related to how to render such methods feasible, both in terms of memory constraints and in terms of achieving rates of convergence in realistic time frames. The overall substantive angle of this dissertation is related to political representation, in particular to U.S. Congress. This substantive focus allows me to study the quality of our democratic institutions and their responsiveness to and interactions with the public. This dissertations harnesses novel, large datasets, which demand the innovative methods developed throughout this dissertation to answer pressing questions related to these issues of political representation. The dissertation focuses on three main data sources: social media data from members of the U.S. House on Twitter, public speech data derived from Congressional Record from 1877-2016 and elections data from the U.S. House from 1956 to 2022. All of these data relate how politicians relate to their constituents: by communicating with them through social media or in public speeches in the first and second cases; and further by trying to earn their votes in the third case. Thus, this dissertation aims to answer questions relating to the use of innovative statistical methods to recover latent features of the data. It explores these questions through the lens of their applications to questions of American legislature. A key finding across domains is the relative unity and stability between legislative leaders and members of their respective parties. In fact, this stability is apparent both in contemporaneous studies of social media, electoral representation in the post-war era, and over historical speeches on the floor of the U.S. House.

Methodologically, this dissertation argues for new frameworks for thinking about large data in political science contexts. It emphasizes the importance of descriptive statistical approaches that consider the full distribution of the data generation process, including higher-order moments beyond the mean. In Chapter 2, it shows how calibrating statistical models for accurate generative descriptions can significant implications for how researchers interpret their statistical results and can accurately uncover important quantities of interest. In Chapter 3, this dissertation proposes new ways to think about external validity when using unsupervised methods for textual analysis. Finally, all three chapters contend with approaches to latent features in the data: topical structure in chapters 1 and 3, and contemporaneous correlations in Chapter 2. All three chapters employ these latent variable methods while proposing solutions to contend with the estimation and computational obstacles imposed by using such methods on large-scale data. In doing so, these papers find unexpected stability and unity among congressional leaders and legislators, with important implications for legislative representation in the United States.

ATTRIBUTION AND ACKNOWLEDGEMENTS

At the time of submission, all of the work presented in this dissertation is in working paper format. However, portions of the chapters presented here the products of research collaborations.

The contents of Chapter 1 include work done with Besty Sinclair, Hao Yan, Sanmay Das and R. Michael Alvarez. A version of the chapter is currently under review. In this project, DE. contributed by developing the methodology and research design, generating the hypotheses from formal theory, processing the data, analysis of the results and writing.

Chapter 2 is based on work produced as a collaboration with Gary King and Jonathan N. Katz. DE wrote the software and implemented the models that produced the main results. With GK and JK, DE contributed to the development of substantive theory, modeling decisions, development of the statistical model, and development of the main methodological frameworks in the paper.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Number *Page*

Chapter 1

LEGISLATIVE COMMUNICATION AND POWER: MEASURING LEADERSHIP IN THE U.S. HOUSE OF REPRESENTATIVES FROM SOCIAL MEDIA DATA

1.1 Introduction

In January 2019, the U.S. Congress was on the brink of crisis and a shutdown. Due to a legislative impasse and political infighting, the legislature could not agree on a compromise to fund the government. Legislative leaders in both parties had to reconcile an uncertain political environment, high policy stakes, and potentially long-lasting electoral consequences. Legislators needed then to balance both their desire to coordinate on a unified message with their desire to actually espouse the right message (with respect to politics, policy, and electoral concerns). The ability of party leaders to set the messaging agenda during this crisis rested on their capacity to balance these concerns. A failure to coordinate on the message or the costs of choosing the wrong message could have resulted in dire political and electoral consequences for the party, as well as harm to the country from unsound policy. This recent example shows that understanding how the party settles on a message and when members choose to follow their party leaders is crucial for understanding how party leadership functions in a democracy.

Existing theories of Congress suggest party leadership power in modern American parties is best explained by national polarization and increased party cohesion that give rise to top-down party leaders. These existing studies focus on polarization as an explanation for agenda setting power, and they often compare power relations across parties. In contrast, we study within-party power relations, and analyze how leadership arises within a party. Our analysis considers a formal theoretical framework that considers the predictions from a signalling and coordination model of Congress due to Dewan and Myatt, 2007. Extending this theory to how party leaders influence member communications, the model suggests that parties balance tensions between coordination and information problems. Parties would like to coordinate around a unified message, but that is difficult because the underlying political, economic and social conditions of the world are uncertain. Leadership's role in this setting is to help facilitate coordination in the face of this uncertainty.

The recognition of this tension in simultaneously resolving these two problems guides our empirical research. Drawing on this theoretical insight, we develop and test a key hypothesis about party leadership in the contemporary U.S. House of Representatives. We test this hypothesis using social media data and unsupervised learning methods. Testing formal political theory with these data and methods is an important contribution of our research.

We focus on a hypothesis which illuminates this informational problem and connects the party members' need for policy direction with House leaders' willingness to initiate discussion. We show structural stability in the findings across a single presidential term, even when the party in power changes.

These expectations contrast with previous studies of congressional party leadership which are conditioned on ideology and legislative institutions. In fact, we believe our results confound expectations because we are focused on the domain of influence over communication on social media. For example, Aldrich and Rohde, 2001 present a theory of conditional government, whereby strong party leaders emerge when parties are internally homogeneous, but are polarized with respect to other parties. As the parties polarize, members delegate more authority to their partisan leaders. Additionally, Aldrich and Rohde, 1998 used DW-Nominate scores to quantify how parties have grown more polarized and ideologically homogeneous. Similarly, Gamm and Smith, 2020 argue that modern parties are top-down institutions, with party leaders exerting control over legislation and committees, especially in the U.S. House of Representatives. Others have argued that modern congressional leadership is powerful: various authors have noted that leaders are empowered with the capacity to bypass committees (Bendix, 2016; Howard and Owens, 2020), to directly negotiate policy (Curry, 2015; Wallner, 2013), set the agenda (Harbridge, 2015), and to limit floor debate (Tiefer, 2016).

We note two key distinguishing features of our analysis relative to earlier studies. First, we avoid the selection problems inherent in using roll call data to identify leadership influence. As party leaders are strategic and have agenda power, they control which bills reach the floor. Since they are unlikely to bring bills to the floor which divide their own party, the fact that leadership-supported bills obtain majorities could signal strength within the party (if leaders persuaded the rank-and-file to support a bill close to the leader's preferred stance), or weakness (if the rank-and-file overrules the leader in the party conference vote). Social media communications are not subject to the same level of leadership control – members of

Congress often cultivate their own online home styles. Second, the high frequency nature of social media data allow us to capture changes in legislative behavior at a much more granular level than roll call data. In particular, social media offers rich data concerning party leadership's ability to direct legislative communication and public engagement around specific topics *among* their members, and in real time.

Our paper contributes to the literature in four ways. First, because we define House leadership influence as the ability of leaders to persuade rank-and-file members to adopt communication strategies similar to their own, we can exploit social media data to measure policy positions (Yan et al., 2019). Specifically, we quantify House leadership influence in terms of leaders' ability to pull rank-and-file public stances on Twitter closer to the leadership's messaging on those same policy positions. Second, we use high-frequency data that shows that the dynamics of leadership can change daily. This suggests that leaders' influence over the party's policy positions varies based on the issues dominating discussion at a particular time. Third, our data let us study the influence of House rank-and-file members on their party leaders. We find that House rank-and-file members exert influence on their leaders' policy position messaging under certain conditions. Our results demonstrate that polarization alone is not sufficient to explain patterns of party leadership in the House. Finally we show that NLP methods and social media data provide insight into online home styles. Thus our work neatly dovetails with Fenno, 2003, as it offers a quantitative approach to understanding how members of Congress communicate with their constituencies and one another.

We argue that understanding the role of communication in shaping institutional structures in the House is central to theoretical understandings of leadership, especially within political parties. In particular, parties balance coordinating around a unified policy position while trying to communicate the best policy position in an uncertain world. We show that political communications data from Twitter illuminates understudied aspects of institutions in the House. Twitter is now a key platform that political leaders use to communicate with their constituents and with other politicians, yielding data on their revealed preferences like roll call votes or newsletters to constituents.[1] We use data from the official Twitter accounts of U.S. House members, collected for the 115th and 116th Congresses, between January

[1] Twitter provides a public forum for members of Congress to interact with each other and the public (Hall and Sinclair, 2018). Past research suggests that congressional Twitter activity is part of a legislator's strategic public communication plan that researchers can use to study legislative behavior (e.g., (Barbera et al., 2019; Kang et al., 2018)).

1st, 2017 and January 3, 2021. After pre-processing these data, we use weakly supervised machine learning methods to show that intra-party variation in our data is associated with observed member behavior, namely House of Representatives messaging mechanisms and the institutional structure within each party's conference. We next discuss the primary hypothesis which guides our analysis, detailing the tension between the coordination and information problems.

1.2 Theory of Leadership Communication and Power

Our empirical analysis is framed around theoretical insights from the Dewan and Myatt (2007) signalling and coordination game of party leadership and communication – where leadership facilitates coordination on a position in response to uncertain issues. In the context of this framework, uncertainty could be the political or electoral popularity of taking a position, or uncertainty about the policy outcome of a stance. For example, the government shutdown of 2019 presented uncertainty of all three types: there were reasons to believe the electoral impact of a shutdown could be either strong or mild and reasons to believe a shutdown could either favor or disfavor the Democratic House Caucus. Further, the policy outcome of the shutdown was uncertain, as the stalemate occurred over border wall policy. The correct position for Democratic and Republican House members to communicate publicly and in real time on social media was not immediately clear. The theoretical framework notes that leaders help resolve this tension between the information and coordination problems faced by party leaders and rank-and-file by acting as a coordination device around a position in light of this uncertainty. In the context of the model, party leaders issue a public speech and then party members try to coordinate on a public position in an uncertain state of the world.

To clarify the theory, we return to 2019 government shutdown debate. House Speaker Pelosi attempted to coordinate her party around a single stance and unite the moderate and progressive wings of her party. The government shut down when President Trump and House Democrats failed to agree on a government funding bill due to disagreements over financing the president's border wall with Mexico. The moderate wing had political incentives to break the impasse by appropriating funds for President Trump's border wall, while Democratic progressives desired a harder line of negotiation. In the meantime, House rank-and-file Democrats were privately discussing their sense of the party's mood around the most politically advantageous messaging strategy as they negotiated with a Republican president to resolve the crisis. These discussions occurred online, in person, and over conference calls.

The private signals in this legislative coordination game represent these online and offline discussions.

We explain the terms of our hypothesis in the context of our illustrative example; the precision of the private signals represents the variation over the moderate and progressive's internal discussions related to the messaging surrounding the border wall and government funding negotiations. As these signals are private, we do not measure this quantity directly. In the model, the party selects one position whose number of supporters exceeds a threshold. In our example, this is Speaker Pelosi's sense of the level of party support she needs to pursue a particular messaging strategy. In the case where neither position has sufficient support, the party fails to coordinate. In the government funding example, Speaker Pelosi initially struck a hardline messaging strategy, and her members followed her lead. She gauged internal support as sufficiently high for this strategy. This illustrates the concept of the *need of direction*. This concept represents the responsiveness of the messaging strategy to the fundamental political environment, and the gravity of choosing incorrectly. In our illustrative example, the *need for direction* is high, as failure to coordinate could result in prolonged national suffering and a calamitous electoral performance for the party assigned blame for the shutdown by the public.

To conclude the 2019 government shutdown example, some Democratic members publicly indicated they did not support the strategy pursued by their congressional leaders during the crisis, and feared political backlash for little electoral gain. We have no reason to believe that they privately supported this strategy, as they actively advocated for countervailing messaging on social media. Nor is it likely that Democratic legislators adopted their leadership's messaging strategy if they thought it was doomed politically. Thus, the public signals reflected internal dissent and internal support for Speaker Pelosi's and her leadership team's proposed messaging strategy regarding the shutdown. This ultimately resulted in Speaker Pelosi making concessions to ideologically diverse factions within her party to ensure they coordinated around her stance on a critical issue. Ultimately, President Trump relented after 35 days and the House and Senate passed a funding bill by voice vote.

In our setting, the public position for each party member is communicated on Twitter. To evaluate the ability of the party to coordinate around the leaders' preferred messages, we construct a measure for the concept of *need for direction* that is discussed in detail in Sections 3 and 4.[2] Specifically, need for direction captures the

[2] Readers interested in details of the theory can refer to (Dewan and Myatt, 2007).

importance of the party coordinating around the “correct” position. When need for direction is high, the information problem tends to dominate. This is because the merits of the position are especially responsive to underlying fundamentals which are uncertain.

We analyze our data at the individual sentiment-topic level. On issues where the party’s need for direction is low, we expect House rank-and-file to adopt the positions of their leaders. Here, the stakes for choosing the wrong position are relatively low, and members prefer to coordinate around a unified policy – even if it is “incorrect” – rather than fail to coordinate at all. For issues where need for direction is high, we expect House leaders to adopt the communication style of their rank-and-file. We define issues with low need for direction as those which explain the variation in the propensity to discuss sentiment-topics, such as the construction of a border wall – which Democrats generally oppose and Republicans generally favor. The “correct” stance on this type of issue for each party is clear. There is little electoral payoff or cost in taking these stances. Conversely, need for direction is high when coordinating on the “correct” stance has out-sized electoral and policy effects, such as a government shutdown. Government shutdowns have resulted in policy and electoral consequences. Here, we expect House leadership influence to be weaker, as the theory suggests that rank-and-file members will hedge against the leaders and adopt their private stance publicly, as the consequences for coordinating on the “wrong” message are high.

Table 1.1 presents the key theoretical concepts and their empirical measures. The first column describes the theoretical concepts as we have defined them in the preceding section, while the second column provides the theoretical meaning of each concept. The third column previews the empirical measures we derive from social media data, which we discuss in Section 3 of the paper. Then in Section 4, we discuss the methods we use to translate theoretical concepts into their empirical analogues, with results in Section 5, and the discussion and conclusion in Section 6.

1.3 Data and Methodology

Data

In order to study the dynamics of communication, we examine legislators’ Twitter posts. Using this high-frequency, individual-level data, we examine whether the House party rank and file discuss topics that are similar to their leaders’ communications on social media or vice versa. We collect the Twitter handles of 511

Concept	Revealed By	Empirical Analogue
Need for Direction	Sentiment-topics with outsized benefit or cost of coordinating	Classify top twenty topics for each party driving separation in sentiment-topic space as uncovered by PCA analysis as needing direction
Leadership Influence	Leaders' ability to convince rank-and-file members to follow their topics	Leaders have statistically significant IRFs on rank-and-file members

Table 1.1: Terminology

representatives from January 3rd, 2017 to January 3rd, 2021, covering exactly the 115th and 116th sessions of Congress. We used the official Twitter handles list collected by C-SPAN[3], following Barbera et al., 2019 who used the NYT Congress API to identify a list of handles for Members of Congress.

We do not include election, personal, or private accounts in our dataset. While many members have additional personal or campaign social media presences, in order to have a consistent method to collect Twitter data from members of Congress, we focus on their official Twitter accounts. It is precisely these accounts that best represent strategic interactions around substantive policy positions. Personal and electoral Twitter accounts often focus on non-policy issues, like personal family matters, sporting events or scheduling of specific campaign events (such as local town halls or rallies). We focus our study on social media posts that are most likely to discuss policy. Our dataset includes 738,066 tweets, including only original posts. Table A.2.1 shows that on average House members tweeted 727.17 times, with notable inter-party variation. Democratic Party members tweeted on average 894.45 times, while Republican Party members tweeted on average 528.31 times.[4]

Methodology

In summary, our analysis proceeds in three steps. First we analyze the original tweets using a Joint Sentiment Topic (JST) model, which we believe is new to legislative studies. We use this model to produce estimates of the daily propensity to discuss a sentiment-topic for each legislator. Second, to uncover the topics in need of direction, we use principal components analysis (PCA) on the member-level average of the topic weights to identify which topics best explain the variation

[3] `https://twitter.com/cspan/lists/members-of-congress/members`

[4] See Figure A.2.1 for the overall distribution of tweets by House members for this period.

between members' preferred discussion topics. Finally we use a daily average of the topical weights for House rank-and-file and for the House leaders to test whether House leaders exert influence and lead on the messaging regarding a policy position or whether House party rank-and-file exert influence and lead discussion.

Joint Sentiment Topic Analysis

We estimate a topic mixture and sentiment mixture, the Joint Sentiment Topic (JST) model, which we believe is new to the study of legislative communication and behavior. It is based on Latent Dirichlet Allocation (LDA), though it estimates an additional latent layer. However, unlike LDA (which estimates two latent layers, topic classification and words alone), the JST estimates three latent layers (sentiment orientation, then topic classification, then word mixtures). Importantly, the JST model estimates the unconditional probability of each sentiment. Note that this model is weakly supervised, as we place a weak prior over the sentiments orientations for a selection of common words.

In order to measure the structure of communication, we use the JST method to classify all tweets for all House members over both sessions of Congress at once. Previous work in political science has used topic analysis to classify open-ended survey responses (Roberts et al., 2014a), while Kim, Londregan, and Ratkovic (2018) have used text to augment an ideological spatial model. Our strategy is an amalgamation of these two approaches. Our work captures the discussion space, without relying on assumptions regarding exogenous covariates to uncover the latent topics.

By accounting for both topic and sentiment, a key feature of the communication structure uncovered by JST is the clear variation in how Democrats and Republicans communicate on social media. By uncovering this inter- and intra-party variation, we are able to analyze behavior within and across parties. Moreover, this method uncovers partisan separation in party communication, evidence that the unsupervised method has external validity. We strongly expect there is a partisan element to discussion on social media from the patterns of communication, which should be especially strong for our sample of members of Congress.

For all tweets in the dataset, we estimate a probability distribution for every word and every tweet which can be decomposed as:

$$\begin{aligned}\Pr(\text{ Word } = w, \text{ Sentiment } = j, \text{ Topic } = k) = \Pr(\text{ Word } = w \vert \text{ Sentiment } = j, \text{ Topic } = k) \\ \Pr(\text{ Topic } = k \vert \text{ Sentiment } = j) \Pr(\text{ Sentiment } = j)\end{aligned}$$

This produces a vector of kj independant sentiment-topic probabilities and j sentiment probabilities for each tweet, which are analgous to the estimates one derives from mix-membership topic models, such as Latent Dirchilet Allocation.

As with many standard topic models approaches, as we connect the JST model to political contexts, the model relies on exchangeability and is a bag-of-words approach to speech, which allows for feasible, tractable estimation. We provide a full technical overview in Appendix Section A.4.[5]

To calibrate the model, we optimize the coherence score of the model. Appendix Figure A.4.2 suggests that the optimal number of topics is 60 topics, the local maximum in the coherence score metric we employ – normalized pointwise mutual information. This is a measure of the extent to which, on average, words we say are likely to be in a topic to be associated in the same topic are actually associated based on what we see in the data. This measure is among the most accurate for determining quantitative coherence for uncovered topics Röder, Both, and Hinneburg, 2015. For the number of sentiments, we fix the number at 3, following the paradigmatic prior in Lin and He, 2009. This results in 84 conditional sentiment-topic probabilities, and three unconditional sentiment probabilities for each tweet.

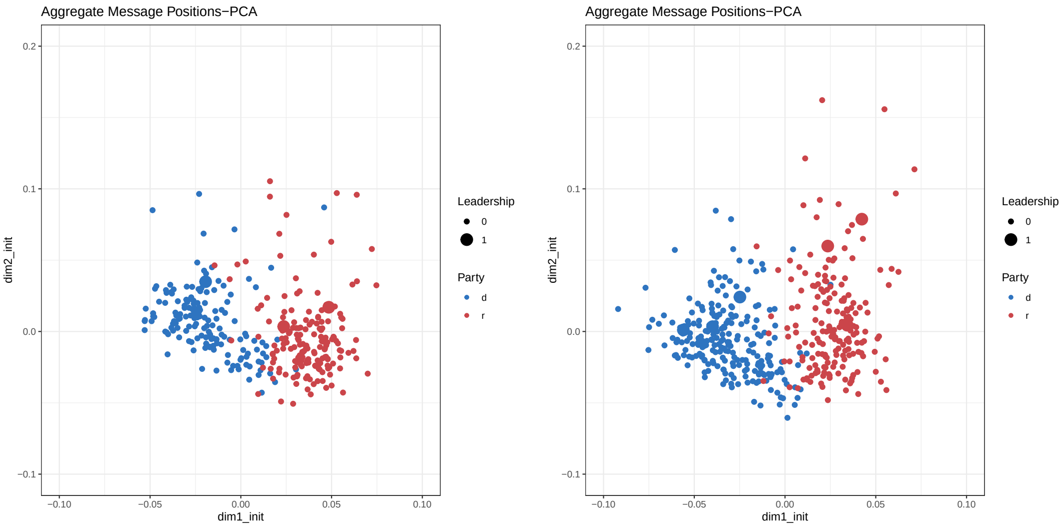
Appendix Table A.4.2 highlights the tweets with the highest probability of belonging to their sentiment-topic label. We report the pre-processed tweet and the associated author-generated labels. The tweets in Table A.4.2 highlight that the JST model produces coherent topic structure, in addition to mathematical coherence.[6]

Measuring Need For Direction

In order to measure need for direction on a policy, we examine structural notions of leadership derived from a PCA analysis of the sentiment-topic space. This is distinct from the topic-by-topic analysis in the preceding section as here we look at measures of party behavior at the party level.

[5] This is also reviewed in Lin and He, 2009 and Lin et al., 2012.

[6] For additional details, see Appendix Section A.4.



(a) 115th Congress

(b) 116th Congress

Figure 1.1: Aggregated legislator policy positioning in the two-dimensional topic space derived from the PCA analysis of the sentiment-topic propensities for the 115th (left) and 116th (right) Congresses. Red indicates a Republican member's policy position, blue indicates a Democratic member's policy position.

Communications decisions among House members are likely guided by exogenous events, party and peer effects, and personal preferences of legislators, which are not immediately obvious from looking at the raw mixtures at the document level. So to understand the individual-level data, we aggregate document-level data by averaging the topical weights for each member. By using PCA as a dimension reduction technique on this aggregate individual-level data, we can identify topics which explain the variation in what members in Congress discuss relative to one another. Figure 1.1 illustrates the sentiment-topic space for all members in our data, summarized by member for the entire period covered by the dataset. We call the coordinate pairs in this figure the policy position for each legislator.[7]

We employ PCA in the following fashion to uncover which topics are in need of direction and which are not. After computing the JST mixtures for each tweet, we find the average probability a House member tweeted about a particular sentiment-topic for the 115th and 116th Congresses.

We emphasize that these PCA results measure a position in sentiment-topic space

[7] We also estimate this measure restricted to only the respective Congress. Figure A.5.4 shows the contrast of rank-and-file members' position in the PCA-derived Twitter communcation space when we estimate it separately. We show the main result is robust to changes in this estimation routine.

over popular debates taking place on social media in real time. PCA analysis allows us to analyze messages espoused by legislators on social media. PCA is useful when taking our JST model as input, as JST accounts for both sentiment orientation and topic content. This allows the latent partisan structure of the data to be detected, without imposing additional structure from potentially endogenous variables to induce this structure. The output of this mapping is a two-dimensional coordinate for each legislator in Twitter communication space for each Congress. From these individual-level measures of communication, we can identify topics which need policy direction or not. These topics form the basis of our empirical tests of the hypothesis regarding party leaders' ability to coordinate.

Dynamic Analysis

Finally, we exploit the micro-level data to examine whether House leaders exert influence and lead discussion on Twitter within their party coalition (and thus exert influence and lead discussion over their rank-and-file), or whether they adopt their members' consensus. As we have stationary data (see Appendix Figures A.6.7 and A.6.8), we follow the time series strategy employed in Barbera et al. 2019. We measure daily propensity to discuss a sentiment-topic in precisely the same way – except using the posterior probability estimates of sentiment-topic JST mixture weights. This is the daily average probability of a House member discussing a particular topic with a particular sentiment orientation. Here, influence is measured by the impulse response functions (IRF) from a vector autoregression (VAR), and we say members or party leaders exert influence and lead when these IRF estimates are statistically and substantively significant.

As our data are stationary, but censored between 0 and 1, as in Barbera et al., 2019, we follow Wallis, 1987's logit specification for VAR. However, our specification contains only two endogenous variables: the average propensity to discuss a sentiment-topic by leader and rank-and-file within each party. We make this choice for two reason: first, because the theory makes predictions over which types of topics should facilitate the emergence of leadership within individual parties, we estimate VAR's separately for each topic and party to evaluate the extent that party leaders emerge as theory predicts. Second, the parameter space is large. Thus, the system of equations may not be identified for a reasonable number of lags. Assuming the topics allows us to identify more lags and improves computational tractability. It also avoids introducing spurious correlations, given the highly interrelated nature of the data. Finally, in cases where the nature of the structural relationships are not known to the researcher, interpreting the results from a VAR regression is difficult. Our parsimonious specification allows for a more direct analysis.

For our specification, we fix a sentiment-topic label k where k can take on one of three possible values: positive, negative, and neutral. Let $x^k_{mem,t}$ and $x^k_{lead,t}$ denote the probability of the average member and average leader respectively discussing a sentiment-topic label k. Let $X^k_t = \left(x^k_{lead,t}, x^k_{mem,t}\right)$. Then let

$$Z = \log\left(\frac{X}{1 - X}\right)$$

Our specification thus is:

$$Z^k_t = c^k + \sum_{p=1}^{7} \beta_{t-p} Z^k_{t-p} + \epsilon^k_t$$

Here c is a constant accounting for the fact the time series are stationary around a non-zero mean after taking logs. Appendix Figures A.6.7 and A.6.8 show for selected series that the times series in log-odds of daily propensity to discuss sentiment-topics are stationary over our period of analysis. Furthermore, Appendix Figures A.6.6 and A.6.5 show that we reject at the 1 percent level a null of unit roots for the vast majority of our time series for the Democratic and Republican Parties across both the 115th and 116th Congresses. These are key assumptions of VAR analysis, and these results indicate that our data are consistent with the the key assumptions of VAR. Finally, we choose a lag of 2 days, which captures the length of the news cycle on Twitter.[8]

Finally, to capture the extent that House leaders or followers exert influence and lead discussion, we estimate generalized impulse response functions for each specification following Koop, Pesaran, and Potter, 1996.[9] That is, we measure the effect of

[8] We also tried a method where we selected the optimum lags based on an AIC criterion, but we found the optimal number was always around 2 days, so we chose to fix the number of lags, given that this fixed number induces a consistent number lags across the specifications and did not substantively alter the results. In fact, choosing lags of 1, 5, and 7 days did not significantly alter the results.

[9] Generalized impulse-response functions IRFs are invariant to variable ordering, unlike orthogonalized IRFs, while still allowing the researcher to study relationships with non-zero entries in the variance-covariance matrix, unlike the forecast error IRF. The magnitude of this IRF is how we derive our second notion of leadership, as noted in Table 1.1. That is, for an n step-ahead response, we compute $\Theta^k_i(n) = \frac{\delta_j}{\sigma^2_j}\Sigma_\epsilon\beta$ where δ is two standard deviations of our data, approximately 10 percent.

a two standard deviation increase in a party leader's log-odds of discussing a given sentiment-topic on the average members' log-odds of discussing that topic and vice versa. Using the median daily propensity to discuss a sentiment-topic as a base rate, we convert the log-odds to relative risk. Using the relative risk, we estimate the change in daily propensity as a percentage point increase over the base rate in the contemporaneous period of the shock. We report 95-percent bootstrapped confidence intervals with 500 draws.

1.4 Operationalizing the Hypothesis

The theoretical framework from Dewan and Myatt, 2007 suggests a clear hypothesis regarding how House party leadership influence relates to party communication. In this section, we connect the theoretical framework to our empirical setting. See Table 1.1 for a road map to our analyses.

Need for Direction

To test the hypothesis that House leaders exert influence and lead discussion when the need for policy direction is low (and the coordination problem dominates) and high (when the information problem dominates), we first need to uncover when leaders exert influence and lead discussion and when rank-and-file members influence discussion.

Coordination Problem

In Tables 1.2 (115th Congress) and 1.3 (116th Congress), we show the sentiment-topics that define issues where the coordination problem dominates.[10]

Our criterion for determining whether each topic needs direction is based on this percent contribution to the variation of the top two components derived from the PCA. We take the top twenty topics that contribute to variation in the member-level propensity to discuss sentiment-topics for each Congress, and classify those topics as being low in need for direction. Sentiment-topics with low contribution to the variation in the sentiment-topic propensities do not drive legislators toward the extremes of sentiment-topic space, while large contributions drive them to the extreme portion of the space. As Figure 1.1 shows, policy positions for House members on these sentiment-topics often delineate membership in a particular party. Thus, for sentiment-topics that drive separation in this space (for example, immigration), we expect little coordination from party leadership, regardless of party, precisely

[10] In the supplementary information we provide the PCA topic contributions for member-driven topics: Appendix Table SI 3 for the 115th Congress and Table SI 4 for the 116th Congress.

Table 1.2: PCA Topic Contributions - Leader Driven 115th

Topic	Contribution
Tax Policy Benefits-Positive	13.79
Tax Cuts-Positive	4.92
Enjoyable Visit - Positive	4.44
Protect Health Insurance -Neutral	3.91
Tune In/Watch Cable News-Positive	2.83
Family Seperations-Negative	2.55
NDAA Passage-Negative	2.28
Middle Class Tax Cut -Positive	2.28
Opioid Task Force-Negative	2.12
Enroll in ACA-Positive	1.96
Pro Trump Mobilization- Positive	1.80
Jobs/Economy - Positive	1.76
Agriculture - Positive	1.73
Signed Legislation-Negative	1.72
Trump Asuylum Policy	1.66
Prevent Gun Violence-Negative	1.57
Abortion Rights-Negative	1.53
Manfacturing Jobs - Neutral	1.51
DACA Policy - Positive	1.49
Trump/Russia Investigation -Negative	1.39

because these are policy positions which delineate belonging to a particular party. In theory, it is on these types of partisan topics that leaders have the most influence over the rank-and-file, since the outsized costs or benefits of coordinating on the wrong messaging are low.

Information Aggregation Problem

We classify the topics not in the top twenty as sentiment-topics as in high need of policy direction. These topics do not contribute to variation in the propensity to discuss topics rank-and-file members of the House. We argue these remaining sentiment-topics, many of which explain less than 1 percent of the variation in the individual propensities to discuss sentiment-topics, represent sentiment-topics where the underlying political fundamentals of the topics are more uncertain, so the information aggregation problem dominates. In this case, failure to coordinate would be preferable to coalescing around the wrong message. For example, on arcane matters of budgetary politics, the optimal message is not immediately clear. The parties may not coordinate on any message, but that might be preferable to coordinating on a message that would be bad for the party. (For the Democrats,

Table 1.3: PCA Topic Contributions - 116th Leader Driven

Topic	Contribution
Tune In/Watch Cable News-Positive	12.53
Impeachment-Negative	11.74
USMCA/Trade Deals-Positive	5.96
GOP attack Democrats as Socialists- Negative	5.03
Humanitarian Aid at Border-Negative	3.09
Trump/Russia Investigation -Negative	2.94
Tune in/Watch Interview-Negative	2.86
COVID economic Relief-Positive	2.52
Lowest Unemployment Rate - Positive	2.44
Census Encouragement - Positive	1.90
Wear a Mask-Negative	1.53
Religious Freedom-Negative	1.46
Climate Change-Positive	1.38
Partisan Attacks on Trump/Biden-Negative	1.36
Border Crimes - Negative	1.36
Criminal Justive Reform-Negative	1.31
Jobs/Economy - Positive	1.28
Racial Inequality in Health Care - Positive	1.26
Public Health and Safety - Neutral	1.26
Snap Benefits-Positive	1.18

they might coordinate on raising taxes, or for Republicans, they might coordinate on cutting Social Security. Neither position would be particularly popular.)

House Leadership Influence

To test the hypothesis that party leaders exert influence and lead when the need for direction is high, for each party, we measure the autoregressive correlations between the average propensity to discuss a topics leaders with the the average propensity of the rank-and-file. To quantify influence, we employ IRF analyses from a vector-autoregression, similar to Barbera et al., 2019. The IRFs enable us to quantify the ability of House leaders to exert influence and lead discussion. We regress the average daily propensity to discuss a sentiment-topic by party leadership on by party rank-and-file, and vice versa. The IRF analysis then supposes a hypothetical shock to the leadership's propensity to discuss a sentiment-topic and estimates the increase in the propensity of rank-and-file member's to discuss. If this shock is statistically significant, we say House leadership influences rank-and-file members' propensity to discuss a sentiment-topic. We also test the reverse – the influence of rank-and-file members on leadership's propensity to discuss a topic.

1.5 Results

Need for Direction by Leadership - Coordination Problem

We find evidence consistent with the theory outlined in the previous sections. The IRF analysis suggests leaders can increase the rank-and-file's propensity to discuss the most partisan topics by between 0.1 and 1 percent for each standard deviation increase in the leadership's daily propensity to discuss a topic. These are substantively large – shocks of 3 or 4 standard deviations (40 to 60 percent) on the daily propensity to discuss a topic are common, so finding discernible effects at the more conservative level of 1 standard deviation suggests the result is would be stronger under conditions that are normal for social media. This reflects the nature of conversation on Twitter, which reacts sensitively to the news cycle. This result is consistent across parties and time, even when the party in power changes. This consistency is evidence that the result is robust across these same dimensions, during the period of 2017 to 2021.

In Figure 1.2, we show the impulse response functions in the first period for the Democrats in the 115th Congress for topic-sentiments that are low in needing direction. Democratic leaders in this period exert statistically significant levels of influence for messaging around preventing gun violence, protecting health insurance, abortion rights, and DACA policy. These topics make sense as having low need for direction – in these cases, the Democrats desired retaining the status quo (preserving Obamacare, DACA) or were discussing topics that are central to Democratic Party ideology, such as abortion and gun violence. In both cases, the party needs little direction in terms of their stances on these issues, so the party would rather coordinate on some message than no message at all.

For Republicans in the 115th Congress, Figure 1.3 shows that economic sentiment-topics are statistically significant. Given the overall strength of the economy from 2017 to 2018, the GOP benefited politically from raising the salience of the economy. We interpret this result as evidence that mis-calibrating the message on the positive economy was less costly than not coordinating on any message at all.

In Figure 1.4, we show the impulse response functions for the Democrats in the 116th Congress for topic-sentiments that are low in needing direction. Democratic leaders in this period exert statistically significant levels of influence for messaging around public health topics, COVID economic relief, climate change, and impeachment. Similar to the 115th Congress, these topics are consistent with being in low need for direction. In these cases, the Democrats discussed two types of such issues. In

Figure 1.2: Democratic Topics: Need for Direction
Predicted Leader Driven 115th Congress

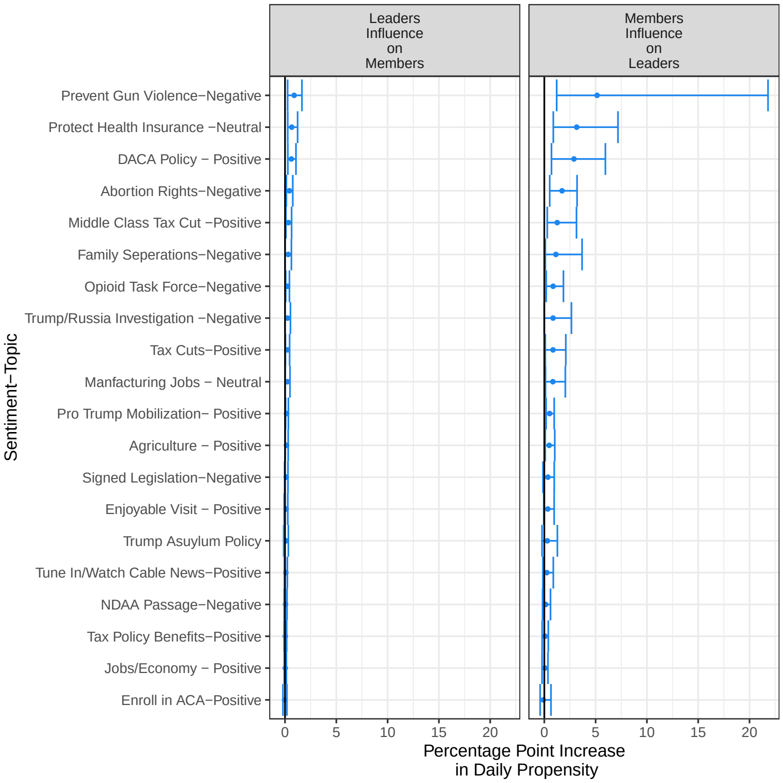


Figure 1.2 : Impulse response functions for sentiment-topics predicted to be leader driven for the Democratic Party. Bootstrapped 95-percent confidence intervals are shown.

the first type, they raised the salience of issues where Republicans faced political downside risk (for example impeachment). Second, they discussed topics that are central to the Democratic Party's ideology, such as racial equality and public health.

Republicans in the 116th Congress exhibit similar behavior to the Democrats in the 116th Congress. For the Republicans, Figure 1.5 shows that shocks to leaders' daily propensity to discuss a particular issue generally results in a less than 1 percent increase in the rank-and-file members' daily propensity to discuss that issue. In particular, Republican leaders induced a ∼ 1 percentage point increase in their rank-and-file members' propensity to discuss impeachment and freedom/sacrifice, and border security. Leaders induced a 0.5 to 1 percentage point increase for impeachment, crimes at the border, attacking the Democrats as socialists, USMCA, and lauding the low unemployment rate. Figure 1.5 also shows that members induced a ∼ 2 percentage point increase in their leadership's propensity to discuss impeachment and humanitarian aid at the border. Members exerted a $\sim$ 1 percentage point increase in their leaders' propensity to discuss crimes at the border and attacking the Democrats as socialists. Additionally, they exerted a nearly 1 percentage point increase for trade deals and USMCA, and lauding the low unemployment rate. Again, members' influence is an order of magnitude larger than the leadership's influence. Notably, the magnitudes derived for Republicans leadership and rank-and-file members are similar to those for Democratic leaders and members. This suggests that party leaders and members are similarly responsive to each other with respect to their messaging regarding their propensity to discuss sentiment-topics, regardless of party.

These results show consistent patterns in legislators' social media behaviors. Party leaders exert influence over the messaging agenda in precisely the topics that are consistent with the theory. In fact, the results for the coordination problem are consistent across time periods, parties and the changes in the party which controls the House of Representatives.

Need for Direction by Membership - Information Problem

Next, we examine in detail the behavior of congressional parties for topics where we believe the information aggregation problem dominates. Intuitively, the information aggregation problem dominates the political environment when there are large costs to the party for choosing the wrong policy. This problem tends to arise when there is more uncertainty in the political environment, be it related to the nature of the political problem, the eventual policy outcome, or the electoral ramifications for taking a policy stance. For example, in a government shutdown scenario, whether to continue the shutdown carries large risks. It may galvanize the base of the party taking the strong stance and increase turnout in favor of the party. Or potentially just as likely, this stance may harm the economy and thus dissuade swing voters from supporting the party. In either case, the potential risks are large. In the case when the information problem dominates, the party relies on "the wisdom of the crowd" of the party at large. By aggregating information, the party hopes to coordinate on the "correct" message, even if this risks not coordinating on any message at all. In these cases, the costs of coordinating on the wrong message outweigh the costs of failing to coordinate.

Our results for topics predicted as member driven are consistent with this theory. Specifically, Figure 1.6 shows that Democratic House members exerted the most

Figure 1.3: Republican Topics: Need for Direction
Predicted Leader Driven 115th Congress

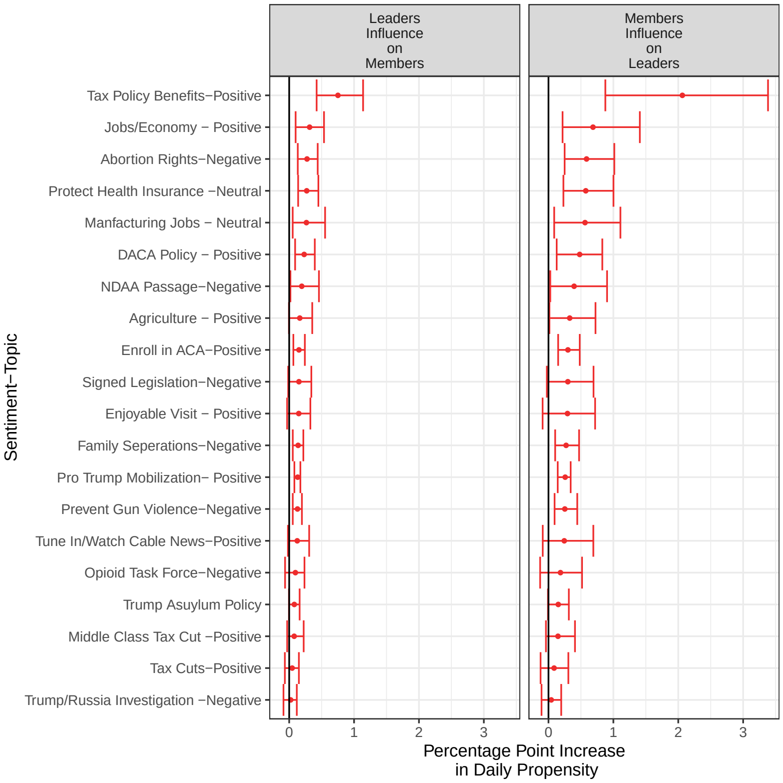


Figure 1.3: Impulse response functions for sentiment-topics predicted to be leader driven for the Republican Party. Bootstrapped 95-percent confidence intervals are shown.

influence over the propensity to discuss Supreme Court nominations (approximately a 4 percentage point increase for each standard deviation shock) and wishing thoughts and prayers after a crisis (a ∼ 2.8 percentage point increase). However, across these same topics, leaders' influence is either statistically insignificant at traditional levels or is near 0. Notably, the effect sizes for members on leaders are an order of magnitude greater than the leadership's influence on rank-and-file members.

The GOP messaging between leaders and rank-and-file is more tightly correlated, but we see that the influence exerted by members is less than influence exerted by Democratic rank-and-file members on their leadership. Rank-and-file members drive a 1.5 increase in both the propensity for leaders to discuss the low unemployment rate and also thoughts and prayers around a tragedy. Notably, as illustrated by Figure 1.7 rank-and-file members exert a ∼ 1 percent increase on the propensity to discuss important meetings. We hypothesize this is an obfuscation messaging strat-

Figure 1.4: Democratic Topics: Need for Direction
Predicted Leader Driven 116th Congress

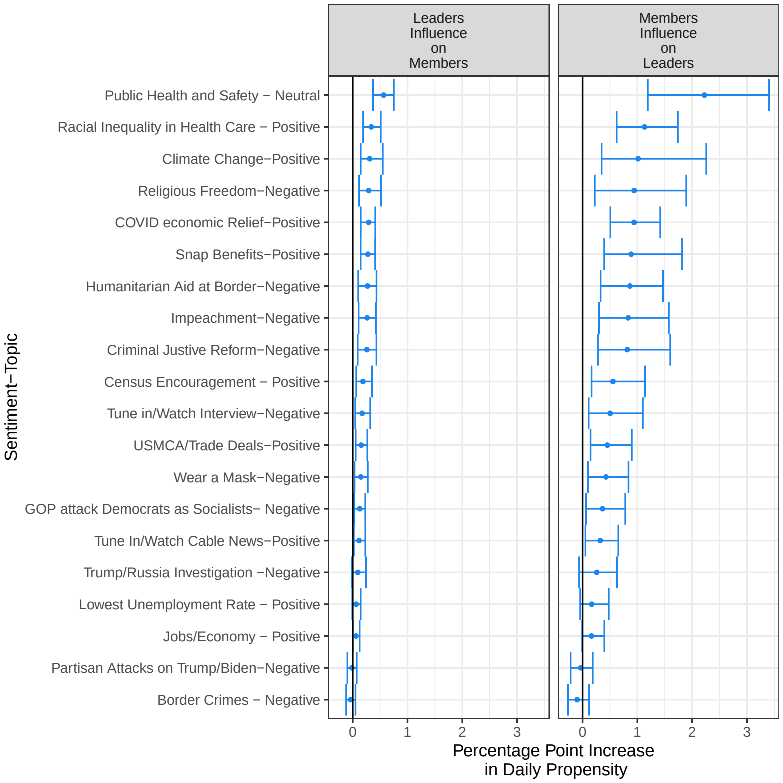


Figure 1.4 : Impulse response functions for sentiment-topics predicted to be leader driven for the Democratic Party. Bootstrapped 95-percent confidence intervals are shown.

egy. Given the majority party runs the risk for being blamed for negative economic and social conditions in the country, this result is preliminary evidence majority parties find it advantageous to engage in measurable amounts of political deflection.

The results for the 116th Congress follow a similar pattern for both parties. Figure 1.8 shows that the Democratic rank-and-file membership exerts a 2 to 3 percent effect on the topics that are in need of direction, whereas leaders exert little influence on these same topics. In the 116th Congress, Democrats became the majority party. Despite this change in institutional control, party communication behavior on social media is consistent with the 115th Congress. Notably, decrying partisan votes – an obfuscation and deflection message – is now one of the key topics where rank-and-file Democratic members exert influence on their party leaders. This is similar to the obfuscation tactics among the GOP rank-and-file when they were in the majority in the 115th Congress. This supports the prediction from the theoretical framework

Figure 1.5: Republican Topics: Need for Direction
Predicted Leader Driven 116th Congress

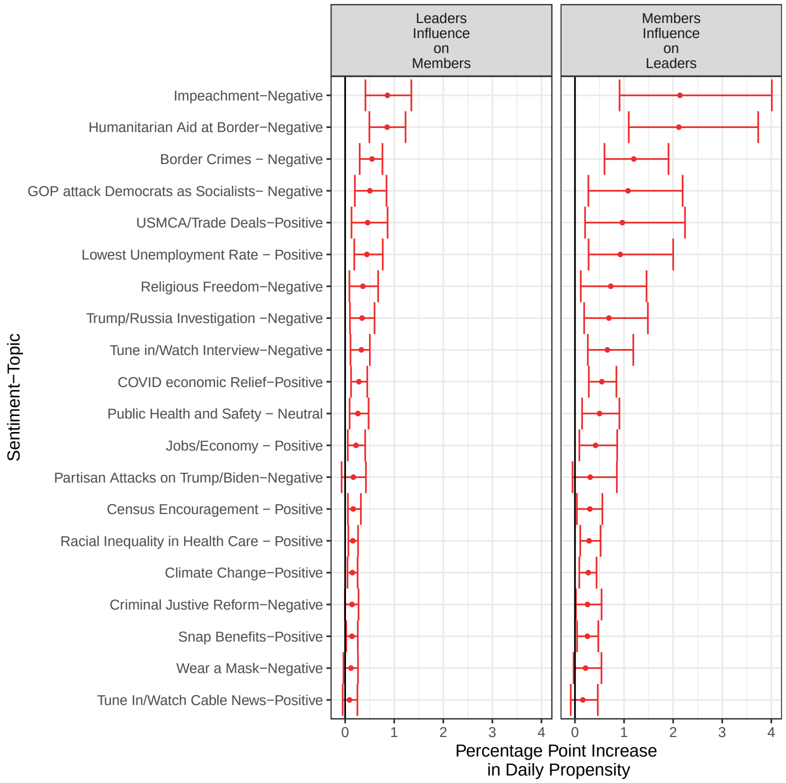


Figure 1.5: Impulse response functions for sentiment-topics predicted to be leader driven for the Republican Party. Bootstrapped 95-percent confidence intervals are shown.

that parties would rather fail to coordinate than coordinate on the wrong message.

In the 116th Congress, the Republican rank-and-file behaves a lot like they did the 115th — and a lot like their contemporaneous Democratic colleagues during the 116th Congress. Figure 1.9 shows that impulses of a standard deviation to the leaders' daily propensity to discuss a particular issue generally results in a approximately 0.5 to 1 percent increase in the rank-and-file members' daily propensity to discuss that issue. As in the 115th Congress, leaders and rank-and-file members both exert influence over these topics, but rank-and-file members' influence is an order of magnitude larger than the leadership's influence. Notable, the magnitudes derived for Republicans leadership and rank-and-file members are smaller than those for Democratic leaders and members. This suggests that party leaders and members are similarly responsive to each other in relative terms between members and leaders, though the magnitude of that influence varies between parties. Additionally, the Re-

Figure 1.6: Democratic Topics: Need for Direction
Predicted Member Driven 115th Congress

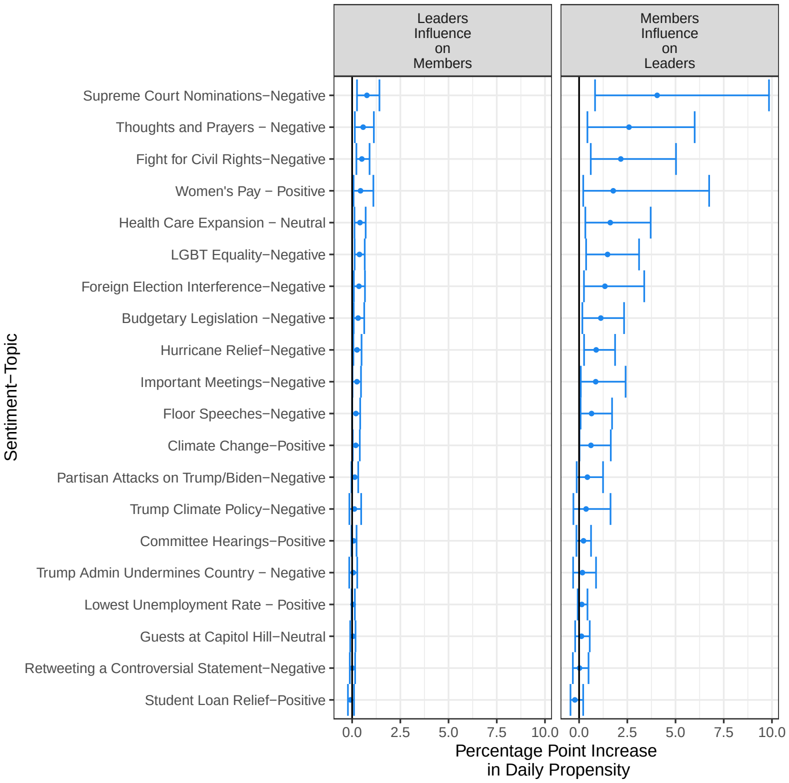


Figure 1.6 : Impulse response functions for sentiment-topics predicted to be leader driven for the Democratic Party. Bootstrapped 95-percent confidence intervals are shown.

publicans, who controlled the presidency, continued to obfuscate, decrying partisan votes and discussing positive constituent visits to their congressional offices.

Discussion

We highlight the consistency of these findings across the parties and the substantive robustness: on issues where House rank-and-file influence discussion, their effect on leaders is larger in magnitude than on issues where leaders lead. This is true across topic types, as illustrated in Figures 1.2, 1.3, 1.4, and 1.5. So, while leaders and rank-and-file influence each other, the measurable effects from rank-and-file are stronger than those on leaders for issues where they respectively had influence.

Finally, in Table A.6.5 we note that leaders exert on average more influence than the most followed accounts in each party. On average, leaders exert double the influence as the most followed accounts from within the same party. This highlights the

Figure 1.7: Republican Topics: Need for Direction
Predicted Member Driven 115th Congress

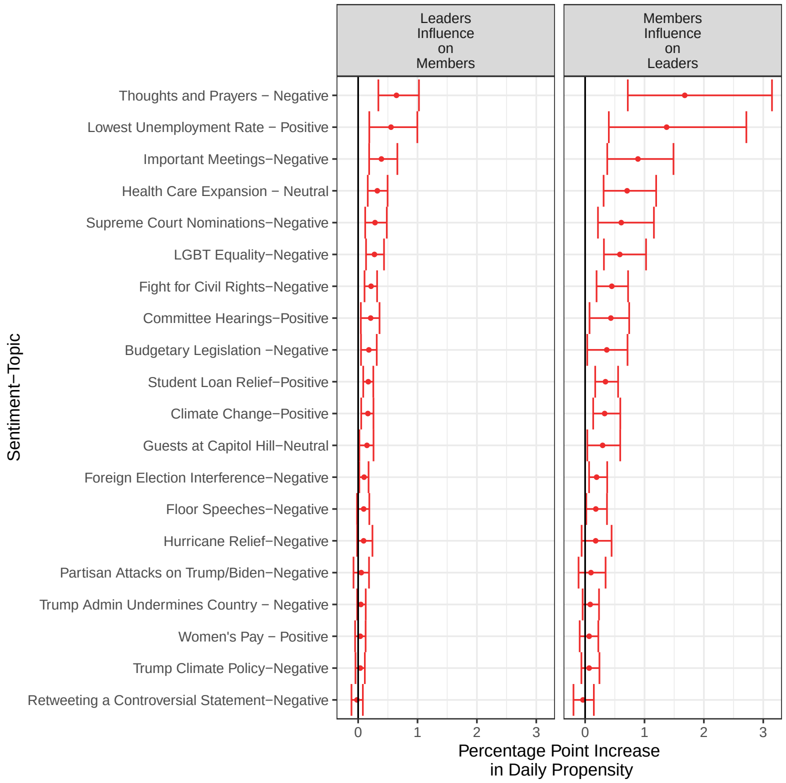


Figure 1.7: Impulse response functions for sentiment-topics predicted to be leader driven for the Republican Party. Bootstrapped 95-percent confidence intervals are shown.

strength of institutional leadership within the party caucus relative to the influence of members of the party who are popular with the public on social media.[11]

1.6 Conclusion

Who controls the legislative messaging agenda has important consequences in a democracy. Currently, the literature on legislative agenda setting suggests that the agenda is driven by national polarization. But other theories, such as formal models of legislative leadership, assert that legislative messaging strategies depend importantly on the information and political environment. In particular these formal theories argue that legislators shift their messaging as they balance coordination and

[11] We also show in Table A.6.5 that leaders exert nearly double the influence on their own members than leaders from the other party exert on the members of the opposing party, suggesting the result is not due to trends on social media. Instead, this result suggests that the role of leaders within their own party explains the result.

Figure 1.8: Democratic Topics: Need for Direction
Predicted Member Driven 116th Congress

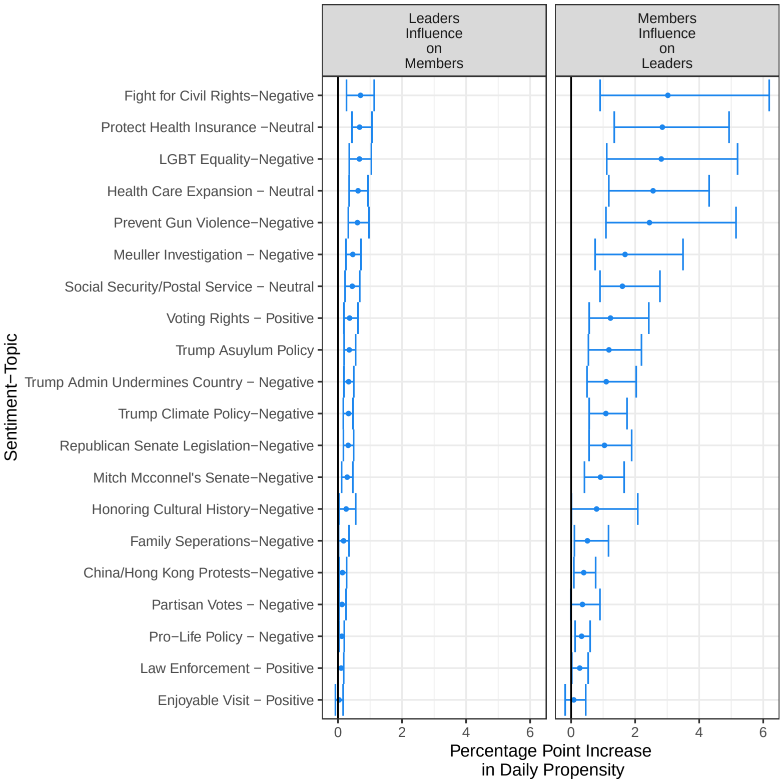


Figure 1.8 : Impulse response functions for sentiment-topics predicted to be leader driven for the Democratic Party. Bootstrapped 95-percent confidence intervals are shown.

information problems. Thus these formal theories predict that when coordination problems are pressing, legislative members follow the policy positions of party leaders.

Our research contributes to the study of legislative leadership, messaging and agenda setting by putting a formal theory of party leadership to the test. We have presented evidence using social media data that the Dewan and Myatt, 2007 theoretical framework of party leadership helps explain patterns of communication and leadership in the U.S. House of Representatives by highlighting the tensions between the need of congressional political parties to coordinate around a unified policy position and the uncertain nature of politics. We present empirical support for our hypothesis that House party leaders exert influence and lead discussion on topics that do not need policy direction, while members exert influence discussion on topics where topics do need policy direction, mediated by information aggregation. To this end,

Figure 1.9: Republican Topics: Need for Direction
Predicted Member Driven 116th Congress

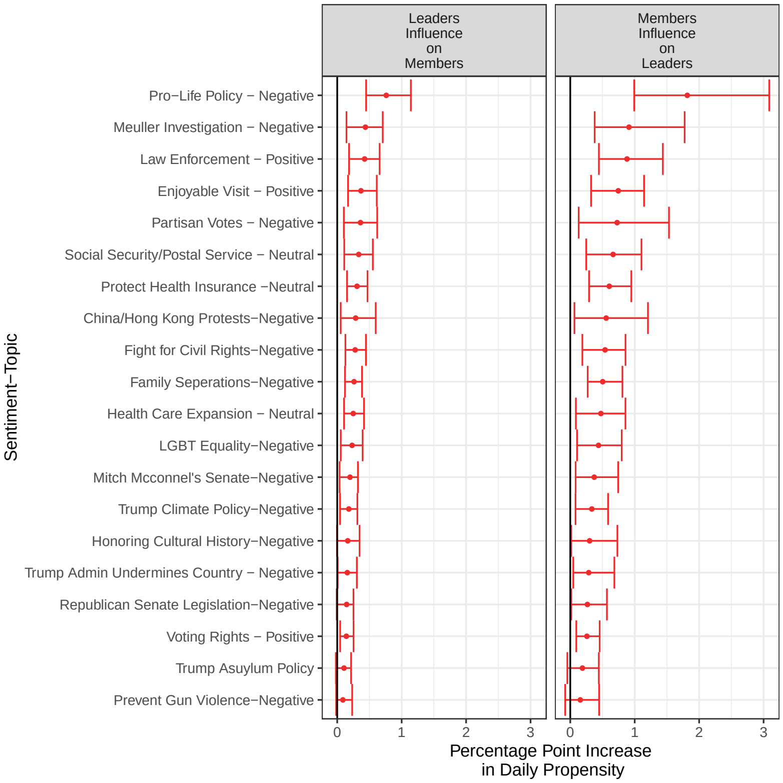


Figure 1.9: Impulse response functions for sentiment-topics predicted to be member driven for the Republican Party. Bootstrapped 95-percent confidence intervals are shown.

we find that, given a large enough shock to House leadership's propensity to discuss a sentiment-topic where the coordination problem dominates, leaders exert a statistically significant influence in the short-run over their rank-and-file members' propensity to discuss that sentiment-topic. Notably, this effect also operates when the information aggregation problem dominates, with influence flowing from rank-and-file to leaders. Moreover, when House rank-and-file members experience a shock to their propensity to discuss a sentiment-topic, leaders are more strongly impacted than in the reverse case. For a standard deviation (~10 percentage point) shock to leadership's propensity to discuss, we might observe 0.5 percent to 2 percent increases in rank-and-file's propensity to discuss. For the reverse, we see a standard deviation (~10 percentage point) shock to House rank-and-file's propensities to discuss a sentiment topic results in a 1 to 3 percentage point increase in leadership's propensity to discuss a sentiment-topic.

This suggests a complex interplay between leaders and members, which is in line with the theory and consistent across parties, changes in partisan control of the legislative institutions, and fundamental changes in the underlying political environment. We find evidence from the IRFs suggesting that leaders exert influence over their members on topics that come to dominate social media discussion. Furthermore, in those cases where members influence leaders, their effect on the messaging of leadership is nearly double that of leadership on rank-and-file members. That is, House leadership and rank-and-file messaging on Twitter influence each other. However, when rank-and-file members drive discussion, their effect is far larger than that of leadership. Thus, using this theoretical model to specify the coordination-information trade-off, we use our data to shed light on the situations where legislative party members resolve tensions between a coordination problem and an information problem.

We believe this theoretical framework provides a blueprint for studying how communication on social media reveals legislative party behavior, and our work demonstrates ways to measure and test a relevant hypothesis derived from the theory. Future work should more precisely classify topics in need of direction versus those that are not. They may also test notions of leadership.

Our research helps demonstrate that social media data is useful for studying legislative behavior and organization. We test formal political theory with social media data using machine learning methods, in line with the recent trend to more closely connect formal political theory with strong quantitative testing (Bueno de Mesquita and Fowler, 2021; Granato, Lo, and Wong, 2021). Using formal political theory to guide our data collection and analytical methods is an important contribution of our research, which we hope provides direction for ways that social media data and advanced quantitative methods can be used to test political theories.

Chapter 2

IF A STATISTICAL MODEL PREDICTS THAT COMMON EVENTS SHOULD OCCUR ONLY ONCE IN 10,000 ELECTIONS, MAYBE IT IS THE WRONG MODEL

2.1 Introduction

Political scientists have studied democratic elections over most of the history of our discipline, producing an extensive, high quality, and steadily improving scholarly literature with few equals across scholarly fields. Statistical studies of actual district-level election returns — including causal effects, counterfactual analyses, forecasts, and full generative models of numerous phenomena — supplemented by a wide variety of other approaches — such as intensive interviews, survey research, participant observation, archival work, and historical analyses — have produced an enviable record of reliable knowledge about the workings of this crucial democratic institution.

Yet, quite often, commonly used statistical models are spectacularly wrong. This is easiest to see in election prediction, where rigorous out-of-sample evaluations are unforgivingly obvious, and a major concern even when prediction is not the immediate goal. Although standard models do remarkably well much of the time, and have taught us a great deal, they are embarrassingly far off with regularity. These mistakes are not ordinary errors of ordinary magnitudes. Our best models indicate that certain events we see regularly should be rarely observed even if we had data from a trillion elections and some from even a trillion-trillion elections.

The intrepid political scientists who give media interviews after elections take one for our team trying to explain this to the public. But pretty much the best they can do is to say something like “Oops!. . . We Did It Again” and to explain that voters get to cast ballots for whomever they want. However, we all know (to paraphrase Britney Spears again) we’re not that innocent. Errors of such magnitude are not merely mistakes. They are bugs in our logic, our models, our forecasts, our conclusions, our textbooks, our advice, and our public pronouncements — similar to what we would think if we built a computer program to forecast the Democratic vote proportion, hit run, and it played a video of a galloping giraffe. This is not a missed forecast; it’s the wrong model. And models that do so badly when they are vulnerable to being proven wrong, as in prediction problems, do not inspire confidence when applied to other tasks more difficult to evaluate and of more interest to social scientists, such as causal inferences or generatively accurate descriptive summaries.

We aim to learn some fundamental characteristics of electoral democracy through a validated generative statistical model capable of estimating many of the diverse quantities political scientists find of interest. These include descriptive quantities — such as the probability of an incumbent losing, the odds of a competitive election, the expected vote of the median house seat, partisan bias, electoral responsiveness, among others — and, with appropriate additional assumptions, causal and other counterfactual inferences. Only a generative model can provide sufficient generality to estimate all these and other quantities, along with accurate uncertainty estimates, which is unlike approaches better for more specific purposes, such as via model-free, distribution-free, machine learning, or semi-parametric approaches. Such a model should also be capable of making election forecasts, but we (and most other political scientists) are not especially interested in forecasting in and of itself (except as citizens to participate in the fun and public interest leading up to an election). After all, from an academic perspective, the best method of forecasting is well known: just wait a bit. However, ensuring that we have a useful model requires that it be made vulnerable to being proven wrong in as many ways as possible, for which forecasting — along with leave-one-election-year-out cross-validation — is essential.

We thus build a new general purpose statistical model and validate it with extensive out-of-sample tests in 14,710 district-level US Congressional elections, 1954-2020. We show that, unlike standard approaches, estimates from this model are correctly calibrated, meaning that its probability estimates are accurate representations of empirical frequencies. Some of the generatively accurate descriptive summaries from this model reveal the rich complexity and dramatic changes in the landscape of US Congressional elections, including a reinterpretation the 1950s as very similar to the present day, except with parties then based on social-psychological groups rather than ideological distinctions. They also suggest an optimistic conclusion about a central feature of American democracy: Although, the marginals sometimes vanish and incumbency advantage sometimes soars, the probability of that incumbents losing their seats has been quite high and essentially unchanged over our entire sample period. Of course, the same model can be used to estimate numerous other quantities.

We describe the standard model and our proposed alternative in Section 2.2, perform many out-of-sample evaluations in Section 2.3, and give substantive findings and even suggest a broader theory of congressional elections consistent with these results in Section 2.4. Section 2.5 describes the broader methodological implications of generatively accurate descriptive summaries.

2.2 Statistical Models of District-Level Elections

We summarize the standard model used in the literature (Section 2.2) followed by our proposed alternative (Section 2.2). We construct our alternative approach to incorporate more substantive knowledge of elections, to simultaneously analyze more elections, and to attend to more of the known statistical issues than previously possible, all within a single Bayesian model. This led us to jointly estimate, integrate over, and represent the uncertainty of 3,567 parameters, including coefficients, missing cell values, uncontested districts, and random effects terms.

One of the reasons our approach has not been tried before is that it would have been computationally infeasible even a few years ago. With highly tuned computational algorithms we developed on a new server (with 20 cores and 128gb of RAM, and software tuned specially to this hardware), we are now able to complete one run of our model on a decade of congressional elections data in only about twenty minutes, although a full analysis of all our data with calibration and strictly out-of-sample evaluation requires about 48 hours of model run time generating about 44gb of output. Along with this paper, we are making available easy-to-use open source software that implements all our algorithms and methods.

Standard

The outcome variable for modeling US congressional elections is the Democratic proportion of the two-party vote, v_{it} for district i and election (time) t. The standard model is a linear-normal regression of v_{it} on a vector of K covariates X_{it}, with estimation conducted for each election year t run independently. For most applications in the last quarter century, an independent normal district-level random effect (constant over hypothetical or real elections but varying over districts) is added to the regression to model the political uniqueness of individual districts (Gelman and King, 1994, implemented in JudgeIt software).[1]

The specific content of the covariates varies some by application but, to fix ideas

[1] Instead of directly estimating γ_i and modeling multiple elections together, which would have been computationally difficult in the 1990s, JudgeIt analyzes one election at a time, after a preprocessing step to estimate how much variation should be attributed to this random effect.

and for the analyses below, we define X_{it} to include a lagged vote share ($v_{i,t-1}$), incumbent party (the party that won the previous election, with 1 for Democrat and 0 for Republican), incumbency status (1 if the Democratic candidate is an incumbent, 0 for open seat, and -1 for a Republican incumbent), uncontestedness (1 if a Democrat runs uncontested, 0 if contested, and -1 if Republican runs uncontested), an indicator for the old confederate states, and a presidential midterm penalty (coded 1 if t is a midterm year and the incumbent party in district i matches the president's party in that midterm and 0 otherwise).

We summarize this model as

$$\begin{aligned} v_{it} &\sim \mathcal{N}(\mu_{it}, \sigma^2) \qquad (2.1) \\ \mu_{it} &= X_{it}\beta_t + \gamma_i \end{aligned}$$

where β_t is a vector of K linear regression effect parameters, $\gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2)$ is an independent normal random effect with variance $\sigma_\gamma^2 > 0$, and σ^2 is the variance of the usual homoskedastic regression independent normal error term.

Proposed Model

We now build on the standard model to develop our proposed approach. We keep the same flexibility in choice of covariates within a fully Bayesian framework, but in three steps we describe our changes. First, in Section 2.2, we provide a qualitative description of model components we designed to reflect knowledge from the literature on elections that had been excluded from the standard approach. Second, in Section 2.2, we put together these components into a single Bayesian model, but for expository purposes focus only on the simple special case where all elections are contested. Finally, in Section 2.2, we allow district elections to be either contested or uncontested.

See Appendix B.1 for the full likelihood, Supplementary Appendix C.4 for computational details, and Supplementary Appendix C.2 for a set of "ablation" studies that, by sequentially removing each model component, demonstrates how all components are essential to the performance we achieve. Supplementary Appendix C.5 considers alternative modeling assumptions.

Novel Model Components

Error terms in statistical models are designed to represent "known unknowns," features that reflect political scientists' knowledge of elections too difficult to code in the covariates. For example, the error term in Equation 2.1 allows for *district uniqueness* by adding a random term γ_i to model the persistence of this uniqueness for any one district i over time, beyond changes due to X. For example, Minnesota's 7th Congressional District has long been more Republican than the nation as a whole, favoring Donald Trump in 2016 and 2020 by about 30 percentage points. Yet, Democrat Colin Peterson won this seat from 1991 to 2021 because of his personal brand and unusual political preferences, opposing abortion and supporting the border wall, but (perhaps accounting for how he wins the Democratic nomination) highly progressive economic views.

We now add to this model four other "known unknowns," modeling features that reflect valuable substantive political information well understood by students of elections or observable in the data but rarely modeled directly. First is *covariate effect stability*: β_t varies relatively little over time. For example, the incumbency advantage might range between two and ten percentage points, with only rare sharp changes over time. Similarly, the coefficient on the lagged vote is usually in the range of $[0.6, 0.8]$. We add this feature to the model by (a) modeling all elections within a "redistricting regime" (i.e., all elections for which the district geography remains unchanged) simultaneously rather than independently, and (b) assuming that each element β_{tk} of vector β_t (corresponding to covariate k, $k = 1, \ldots, K$ and time t) comes from the same distribution $\beta_{tk} \sim \mathcal{N}(\hat{\beta}, \sigma_{\beta_k})$, where $\sigma_{\beta_k} < \infty$; in contrast, estimating each equation separately and independently, as in the standard approach, is equivalent to setting $\sigma_{\beta_k} \to 0$. (The notation $\hat{\beta}$ is a shorthand reference to empirical Bayes, meaning that this distribution shrinks different covariate effects in the same redistricting regime toward the estimated mean without favoring one's a prior guess; this is equivalent to a fully Bayesian model with the mean in the null space; see Girosi and King 2008.) The idea here is to borrow strength for the estimate of each parameter in each year from the estimation of the same parameter in other years, but without the rigidity and potential bias that would come from a more "informative" prior. This will be especially valuable in smaller legislatures, such as many state assemblies and senates and the class up for election in the US Senate.[2]

Second, we allow for *positive cross-district covariances* by adding a random national swing term, η_t, that allows all districts in one election to be affected in roughly the same way by the same national event, over and above the information in X. For

[2] We could elaborate this assumption by allowing β_t to trend linearly, as a random walk, or as a function of other covariates, but we find no evidence for these alternative approaches in our data.

example, the 1994 Republican national congressional campaign strategy (known as the "Contract With America") seemed to be a successful heresthetical maneuver (Riker, 1990; Shepsle, 2003) that moved all the districts in the Republican direction by approximately the same amount. Although we cannot know ex ante what any one national swing will be, we can estimate the variation caused by the national swings, which we know occur regularly, and represent this uncertainty in the model with a common random effect for all districts. The result is the well known "approximate uniform partisan swing" pattern common across time periods, electoral systems, and even countries (Katz, King, and Rosenblatt, 2020).

Third, we model *district-level political surprises*, including intentional heresthetical maneuvers and unintentional exogenous political events that affect one district's vote at a point in time differently than others and are not included in X. Consider for example the election in Texas' 22nd district in 2006. Tom Delay was the popular Republican House majority leader from the district, regularly winning election by 35 or more percentage points. During the campaign, he was indicted and abruptly resigned. Worse for his party, the deadline to field a candidate on the ballot line had passed and so his party could only field a write-in candidate late in the campaign. The result was that this overwhelmingly Republican district elected a Democrat over the Republican write-in candidate by over 8 percentage points. Equation 2.1 already includes the usual normal error term that can be used to model surprises, but a normal distribution indicates that deviations from a prediction this large would happen so infrequently that it would almost never be observed. Of course, as every election observer is aware, these surprises happen regularly, even if we do not know which ones will occur. As we explain below, we will therefore swap out the normal distribution for one that can more appropriately represent these political surprises, also keeping predictions within the [0,1] interval.

Finally, the normal distribution used in the standard approach turns out to be inadequate for two reasons. The first reason is that, although it often works well when the mean or average vote outcome is of interest, it fails miserably for most other aspects of the distribution, such as for uncertainty estimates or the probability of a close election or of one party winning. The second reason is that the normal tail implies that big surprises should almost never occur, meaning that it also gets the concentration around the mean wrong. To fix these problems, we use the additive logistic Student t (ALT) distribution, which, unlike the normal, constrains the vote proportion to the [0,1] interval and also has appropriately fatter tails to represent surprises. In addition, the ALT distribution has the simultaneous advantage of having more of its density concentrated near the mean, making the mean (and the covariates that account for its variation) more informative at the same time as it is accounting better for surprises.[3] The ALT distribution thus allows more informative predictions to coexist in the same model with the possibility of huge surprises.

The Model, with Fully Contested Elections

We now combine all the features described above in one model, reusing the notation (and redefining symbols) from Section 2.2. For expository simplicity, we imagine until the next section that all districts are contested. Thus, let

$$v_{it} \sim \text{ALT}(\mu_{it}, \phi_t^2, \nu_t), \tag{2.2}$$

$$\mu_{it} = X_{it}\beta_t + \gamma_i + \eta_t \tag{2.3}$$

where the variance is decomposed by the ALT for additional flexibility into scale ϕ and degrees of freedom ν_t parameters (as $\nu_t \longrightarrow \infty$, the ALT approximates the additive logistic normal). The systematic component for the conditional expected value includes three independent random effect terms for covariate effects, district uniqueness, and national swing, respectively,

$$\beta_{tk} \sim \mathcal{N}(\hat{\beta}_k, \sigma^2_{\beta_k}), \quad \gamma_i \sim \mathcal{N}(0, \sigma^2_\gamma), \quad \eta_t \sim \mathcal{N}(0, \sigma^2_\eta),$$

for $k = 1, \ldots, K$ covariates, $i = 1, \ldots, n$ observations, $t = 1, \ldots, T$ elections, and diffuse priors chosen for estimation convenience (see Appendix C.4).

For intuition, we consider the voting data on the logistic scale by letting $y_{it} \equiv \ln[v_{it}/(1 - v_{it})] = \mu_{it} + \omega_{it} = X_{it}\beta_t + \gamma_i + \eta_t + \omega_{it}$, with error term $\omega_{it} \equiv \ln[v_{it}/(1 - v_{it})] - \mu_{it}$, which Equation 2.2 indicates is is t distributed. This enables us to see, first, that national swings induce a positive covariance between any two districts i and j ($i \neq j$) for each election year t: $\text{Cov}(y_{it}, y_{jt}|X_{it}, X_{jt}, \beta_t) = \sigma^2_\eta > 0$. This setup also makes clear that the random district uniqueness term γ_i induces a positive covariance for election outcomes in any one district i at two times t and t' (within the same redistricting decade), over and above differences due to X: $\text{Cov}(y_{it}, y_{it'}|X_{it}, X_{it'}, \beta_t, \beta_{t'}) = \sigma^2_\gamma > 0$.

As with the standard approach, some covariates one might put in this model vary over i and t (e.g., the lagged vote, v_{it}), some vary only over i (e.g., the confederate

[3] Roughly, the ALT is the implied distribution on v (and so restricted to the [0,1] interval) when the t distribution is applied to the (unbounded) logistic transformation of the vote $\ln v_{it}/(1 - v_{it})$. For technical details, and extensive evaluations in multiparty elections, see Katz and King (1999).

states indicator), and some vary only over t (e.g., presidential approval). A random effect can also be included, which can be useful when little information exists such as for covariates of the last type when T is small.

The Model, Allowing for Uncontested Elections

In the standard approach, the vote in uncontested elections is often recoded to fixed values such as $v_{it} = 0.25$ for Democrats running uncontested and $v_{it} = 0.75$ for Republicans running uncontested, or sometimes uncontested elections are deleted entirely. We instead formally distinguish between the observed vote v_{it} and the *effective vote* v^*_{it}, defined as the vote proportion that would be observed if the election had been contested (e.g., King and Gelman, 1991a). The effective vote is observed $v^*_{it} = v_{it}$ in contested elections but unobserved if one party runs unopposed. We then impute unobserved values (for uncontested elections) during Bayesian estimation simultaneous with the rest of the model. This approach includes all the information available and accounts for all uncertainty in the imputation.

To model v^*_{it} when unobserved, we replace the outcome variable v_{it} in Equation 2.2 with the effective vote, and add a "censoring assumption": candidates who run unopposed would have won even if the election were contested. This assumption is intuitive, probably accounts for why the district was uncontested in the first place, and is a special case of the assumption made by Katz and King (1999). We then replace Equation 2.2 with

$$v^*_{it} \sim \text{ALT}(\mu_{it}, \phi^2_t, \nu_t), \tag{2.4}$$

and write the likelihood function for an election district that is fully contested as $\text{ALT}(v_{it} \mid \mu_{it}, \phi^2_t, \nu_t)$, for a district where a Democrat runs uncontested as $\psi_{it} \equiv \int_0^{0.5} \text{ALT}(v^* \mid \mu_{it}, \phi^2_t, \nu_t) dv^*$, and for a district where a Republican runs uncontested as $1 - \psi_{it}$. The integral implements the censoring assumption.

The model contains one additional feature: When the lagged effective vote is used as a covariate, it too can be unobserved, which adds another level of modeling complexity. We describe this feature, along with the full likelihood, in Appendix B.1.

2.3 Evaluation

We now evaluate both the standard linear-normal approach and our proposed additive logistic t model with contemporaneous correlations, or LogisTiCC for short. We do this by summarizing the models' statistical properties (Section 2.3), comparing the probabilities of rare events from each approach to actual elections (Section 2.3), and studying the models' confidence interval coverage (Section 2.3).

Statistical Properties

As political scientists have long understood, the linear-normal model can reveal important information about elections, when its specification is correct or close to correct. The standard modeling approach is not formally a limiting special case of the LogisTiCC although it can be thought of as an approximation in some situations. For one, as with all potentially misspecified models, point estimates from the linear-normal model will choose the distribution closest to the true data generation process (in the sense of the Kullback-Leibler information criterion; see White 1996) even if the data come from the LogisTiCC. In addition, if the linear specification is correct, both the normal and the LogisTiCC models will produce similar (and approximately consistent) estimates of (the same) β.

Unfortunately, given the covariance structure of the proposed model, estimates from the normal will be highly inefficient relative to the LogisTiCC, if data come from the model we are putting forward that would seem to better represent the knowledge of election experts, and standard errors of β will be incorrect. However, most quantities of interest other than β, such as even the probability of a candidate winning an election, will be statistically inconsistent under the normal but consistent with the LogisTiCC.

As we demonstrate, a key problem with the linear-normal model is its incorrect independence assumptions, leading to substantial false precision in its uncertainty estimates (confidence intervals and standard errors that are too small). In contrast, the LogisTiCC allows for dependence among elections held in the same district at different times and among elections held in different districts on the same day. Correcting for this false precision leads to appropriately larger confidence intervals: the ratio of the nominal width of LogisTiCC-to-normal confidence intervals is about 1.4 for district-level predictions and about 5 for aggregate predictions such as the vote for the median house seat. (See Supplementary Appendix C.1 for details.)

Rare Event Probabilities

We analyze 28 years of US Congressional elections from 1954 to 2020, including a total of 14,710 district-level contests, with forecasts limited to the 10,778 contests that exclude the first year of each redistricting decade. This large dataset enables us to conduct numerous rigorous evaluations (cf. Grimmer, Knox, and Westwood, 2022), all of which we do out of sample (so that no data from the election being predicted is used during calibration or estimation). In each analysis, we use either a one-step-ahead or leave-one-out forecast, depending on context.

To begin, consider the probability of extraordinarily rare events under each model. For illustration, we use the notion of *moral certitude* from the Enlightenment, which is that events with probabilities smaller than 1 in 10,000 should be disregarded. (Because demographers of the time observed that the probability of a healthy person dying in the next day was smaller than 1 in 10,000 and does not seem to affect people's behavior in their daily lives, people act as if they are "morally certain" that these rare events will never occur; see Kavanagh 1990; Buffon 1777.) Updating this (quaint) idea, we make predictions for all elections in our dataset (except the first year in each redistricting decade) and count the number of elections for which the vote proportion observed out-of-sample appears outside a 99.99% (i.e., $1 - 1/10,000$) forecast credible interval. If the interval is correct, we should observe about 1 in 10,000 outside the interval.

Figure 2.1 gives a count of these extraordinarily rare events (on the vertical axis) by election year (on the horizontal axis) and for the normal model (in gold) and the LogisTiCC (in black). As can be seen, the data dramatically violate the normal model's predictions in a disturbingly large number of elections. In the entire dataset of 10,778 elections, we would expect to see only about *one* 1-in-10,000 event, but this claim is wrong by a factor of more than sixty, in that surprise events the model is morally certain will not occur actually happened in 61 elections (and as many as 12 of the 435 elections in a single year, 1958) (see also Gelman et al., 1995, Ch. 8). The figure also annotates some of the points with the exact probability that we would expect to see these results under the model. These forecasts are stunningly bad. The late Richard McKelvey was fond of arguing that a fix for over-claiming in empirical work would be to require anyone reporting a p-value to take a bet with the implied odds (i.e., the reciprocal of the p-value to one) against someone finding evidence to the contrary. Using this logic, a one dollar bet against the linear-normal model's claimed level of certainty would give an equal chance of winning quadrillions of times more money than exists in circulation in all the world's currencies.

In stark contrast, the black line in Figure 2.1 shows that only one of the 10,778 out-of-sample observed election results are much of a surprise to the proposed LogisTiCC model. All but one year has zero events and just one (in 1996) has one

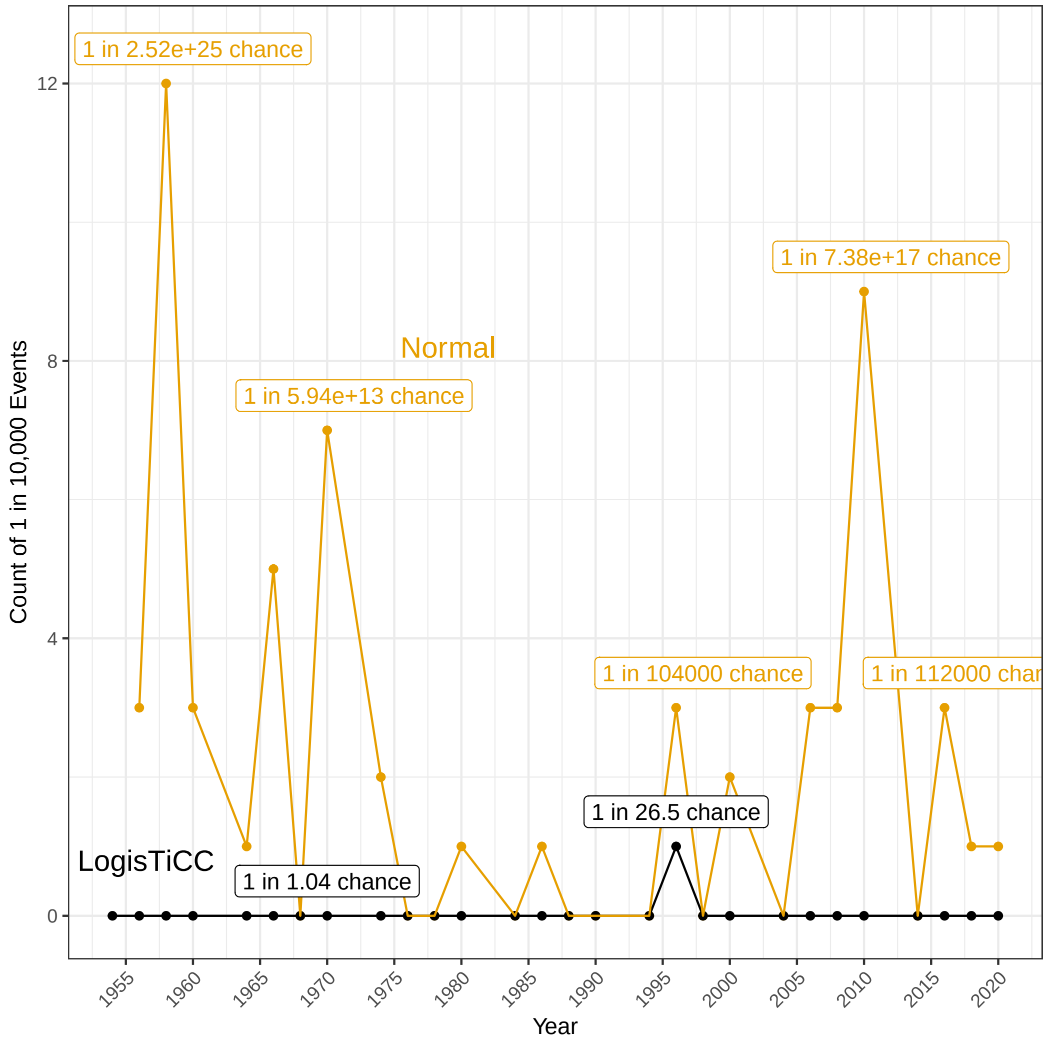


Figure 2.1: Moral Certitude: Count of elections outside a 99.99 credibility interval for each election year (with selected points labeled with the probability each model gives of seeing this many 1-in-10,000 events). Separate calculations appear for the normal model (in gold) and our proposed LogisTiCC model (in black).

event with a modest probability of 1 in 26.5, which is about what we would expect if the world generated all the data according to this model.

Thus, for this measure of extraordinarily unlikely events, the out-of-sample performance of our proposed model vastly exceeds that of the standard approach. We now show that this result is general in that the probabilities from our model, but not the normal, are well *calibrated*, meaning that for example when the model predicts that a certain event will occur with a 30% probability, that event actually occurs in about 3 of every 10 elections, and so on. We do this, for each election and model, by first computing the (out-of-sample) probability of a competitive outcome (which we define as $v_{it} \in [0.45, 0.55]$). We then sort these probabilities into bins, $[0, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$,. . . , separately for each model, and plot them in Figure 2.2, as follows. For each model, we plot a dot with a horizontal coordinate as the average of the estimated probabilities of elections in a bin and the vertical coordinate as the number of (out-of-sample) elections in the same bin that are in fact observed to be competitive. Dots for a perfectly calibrated model should fall approximately on the 45 degree line.

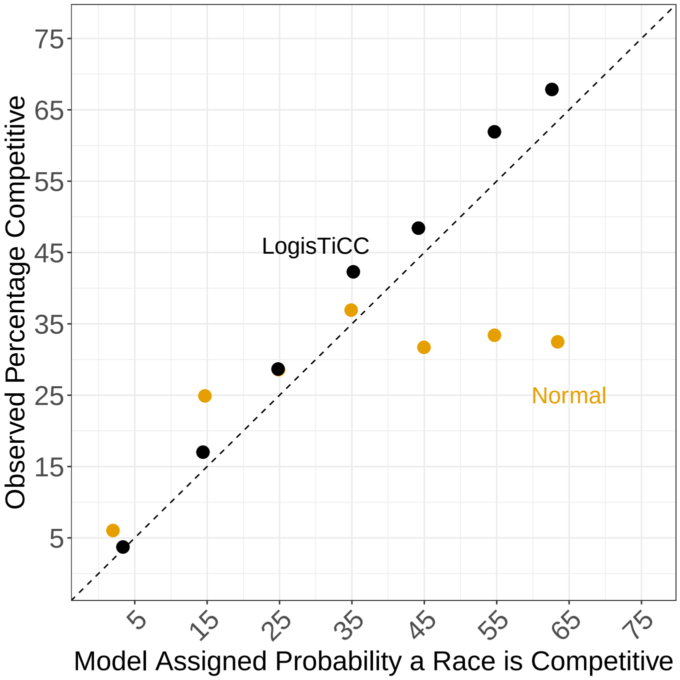


Figure 2.2: Calibration: Predicted out-of-sample probabilities (horizontally) by observed frequencies (vertically).

As Figure 2.2 demonstrates, the dots computed from the LogisTiCC bins (in black) are all close to the 45 degree line, and hence well calibrated. In contrast, those from the normal (in gold) substantially deviate from the 45 degree line of equality as the predicted probability of a competitive election gets higher. In other words, the normal model fails most dramatically in elections that are most politically important, the competitive ones.

Coverage

We now study, in three ways, the properties of credible intervals computed from the standard and proposed models.

First, we plot in Figure 2.3 a time series of one of the most consequential quantities of interest in US politics — the Democratic proportion of the vote of the median seat in the House of Representatives (see the red stars). Then, for each year and model, we omit this year from the dataset and compute a point forecast and 95% out-of-sample credible interval around it. These appear in gold for the normal and black for the LogisTiCC. In addition to the LogisTiCC intervals being longer than for the normal because of the normal's false precision, the LogisTiCC intervals should be interpreted differently. First, recall that a t-based interval has both fatter tails to accommodate surprises and more concentration of density near the mean than the normal (making the mean prediction more informative). Second, the LogisTiCC intervals are accurate (See Figure 2.2) whereas the normal intervals are overconfident. This can be seen because in these out-of-sample tests, we would expect a well calibrated model to miss only about 1.4 elections, but the normal misses 20 of 27. In contrast, LogisTiCC's predictive confidence interval captures the observed outcome every time.

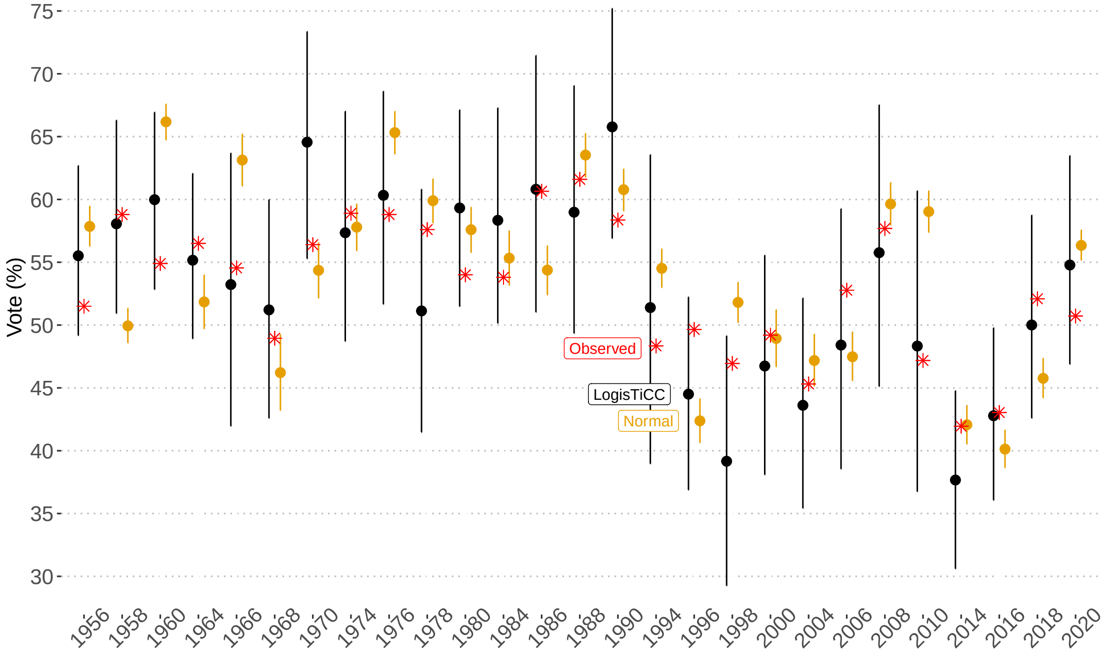


Figure 2.3: Expected Vote Share of the Median House Seat (95 Percent Credible Interval).

Second, for each model, we compute a 95% out-of-sample credible interval around every individual district's vote share and tally up the percentage of districts that interval captures. Our results appear in Figure 2.4, with time on the horizontal axis and the percent coverage on the vertical axis (again with normal in gold and

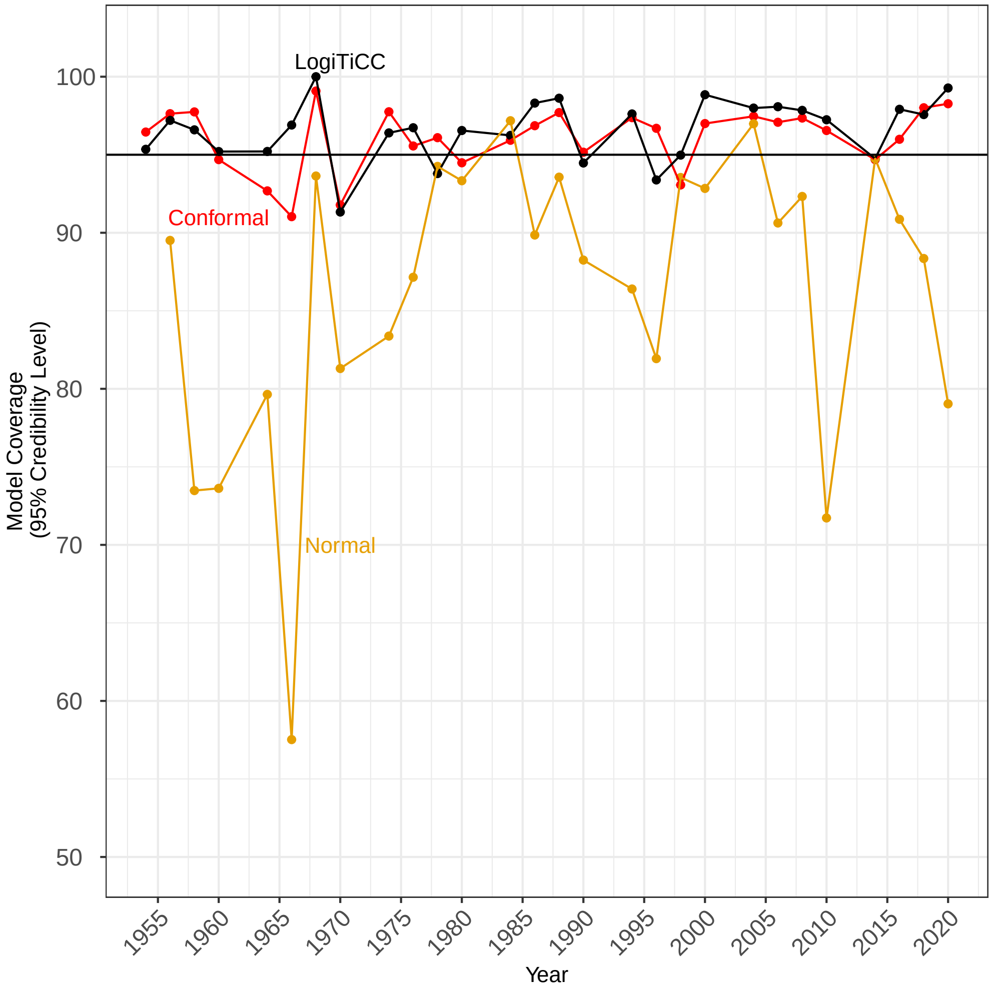


Figure 2.4: Coverage under Each Model at the 95 Percent Level.

LogisTiCC in black). A properly calibrated model should capture 95% of districts which, aside from estimation error, should be at the flat black line near the top of the figure. This is the case for the LogisTiCC, which has well calibrated intervals. In contrast, the normal interval substantially deviates from capturing 95% of the elections in all but a few years.

Finally, we evaluate our distributional assumption (a compound error term with random effects and an additive logistic *t* distribution). To do this, we use methods of "conformal inference" that offer guarantees of accurate distribution-free finite sample coverage even under model misspecification, for any predictive model, and so we use it to check for misspecification in our model (Vovk, Gammerman, and Shafer, 2005). (Intuitively, the method works by computing confidence intervals based on errors from previous years' forecasts, assuming primarily that the data generation process is exchangeable conditional on the covariates.) In Figure 2.4 we add conformal confidence intervals (in red). We first confirm that the conformal intervals have accurate coverage, as designed, which we can see as the red line varies around the flat 95% line across the years. More relevant for our purposes is the comparison between the fit of the red and black lines to the 95% line. This comparison indicates that the LogisTiCC has approximately the same high quality coverage as these distribution-free intervals. These results thus provide evidence for the veracity of our distributional assumptions and for our Bayesian model as a generative model of US congressional elections data.

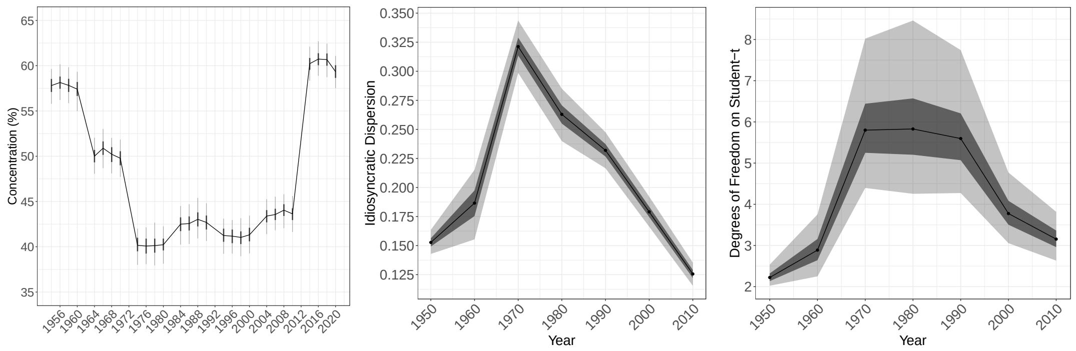
2.4 Electoral Implications

We use our model to compute generatively accurate descriptive summary statistics. First, in Section 2.4, we characterize election variation as falling into three regimes, at the start, middle, end of the 66 years of our study, and how elections throughout are powerfully driven mostly by national rather than local swings. Second, Section 2.4 builds on the first section with an empirical theory of congressional elections consistent with our empirical results and prior literature that tries to strip out several under-appreciated normative assumptions. Section 2.4 then focuses on a key feature of American democracy, the probability of an incumbent loss, and shows that it is essentially constant over time, despite well known huge changes in the incumbent's expected vote advantage.

The Three Regimes of Election Prediction Variability

Our model decomposes election variability into district uniqueness, national swing, covariate effect stability, and political surprises, in addition to well known covariate effects. As Section 2.3 shows, these parts of the model provide far better fit to congressional elections data, making for accurate out-of-sample forecasts, uncertainty intervals, and calibrated probabilities. We now turn to the large scale patterns this modeling strategy reveals in congressional elections, leaving most of the substantive implications to the following sections.

First, we begin with an intuitive summary measure of the overall patterns in congressional elections data that we call vote *concentration*, the proportion of the vote probability mass in the interval [0.45,0.55], for mean predictions of 0.5. As Figure 2.5a shows, the early and late periods have high vote concentration, meaning that any one prediction conveys more certainty and more information, whereas the middle years have substantially lower concentration values, indicating that predictions in



(a) Concentration (vote probability mass within ±5 points of 50% expected vote)

(b) Scale, ϕ

(c) Degrees of Freedom, ν

Figure 2.5: Model Features.

that period were of less (or more variable) value. These are not small differences: A prediction of $v = 0.5$ plus or minus five percentage points in the 1950s and the 2010s captures about 60% of likely voting outcomes, whereas in the 1970s-1990s the same interval only captures 40% of these outcomes.

Second, our results show that the national swing is far more important than the variation due to district uniqueness (even after accounting for the covariates), which is one reason for strong time series patterns in voter concentration. To see this, we compute the ratio of the standard deviation of the vote (on the logit scale) due to variations in national swing relative district uniqueness: $\sigma_\eta/\sigma_\gamma = 0.2/0.036 = 5.6$ (a ratio we find to be largely stable over time). Campaign observers have long known that exogenous events and heresthetical maneuvers by individual congressional candidates in their district campaigns can be important, but this result shows that exogenous national events and national-level heresthetical maneuvers are more than five times as consequential as the sum of all the individual district campaigns. All politics may well be local in its effect, but national level political issues have a far bigger effect both nationally and locally than local issues (see also Hopkins 2018 and Caughey and Warshaw 2022: Sec. 3.3).

Finally, we decompose the vote concentration results from Figure 2.5a by noting that the ALT distribution partitions the overall variance into two parameters, the "scale" ϕ, which quantifies the amount of variation, and "degrees of freedom" ν, which controls the shape of the predictive distribution. Time series estimates of these parameters appear in Figures 2.5b and 2.5c, respectively. In both cases, we see a clear inverted U shape, revealing low variability in electoral outcomes at the start

(1950s–60s) and the end of the series (2000s–2010s) and much higher variability in the middle years (1970s–1990s). The degrees of freedom parameter is similarly low at the start and end of the period, indicating sharper deviation from the normal with both longer tails and more concentration of density around the mean prediction, and higher values near the middle, indicating lower concentration.[4]

An Empirical Theory of American Democracy

The literature on American elections is increasingly scientific, but it has not always made its underlying normative assumptions transparent, which may have led to unrecognized biases and missed opportunities. We first clarify this point and then turn to a reevaluation of our empirical evidence.

Avoiding Normative Assumptions

Here we highlight the sometimes unrecognized philosophical assumptions in the literature. To do this, we begin with a simple characterization of American representative democracy as *a set of electoral rules that enables politicians to seek office by making public appeals and voters to choose among the politicians*. Importantly, the electoral rules constrain neither the arguments politicians make nor the calculus voters use in choosing candidates.

Political scientists and political philosophers have long layered on top of this simple definition various normative assumptions that they either consciously justify as important or effectively treat as facts. For example, scholars frequently ask whether voters pay attention to the important issues of the day, but they too often presumptuously define "importance" when in fact that's the voters' job. War, gun control, trade, unemployment, inflation, taxes, abortion, energy policy, and others, may sound important to political philosophers, but nowhere in American electoral rules do the normative preferences of a bunch of academics get to determine how voters make their decisions.

Similarly, when we impose our normative preferences for what counts as consistent positions across issues, voters may have a range of values of issue constraint from low to high. In fact, however, issue constraint is always "high" by definition, once we recognize again that the voters get to decide how much to count different issues in their voting decision. If voters decide only personality is important, or being

[4] See Supplementary Appendix C.6 for additional empirical evidence of the three regimes. Note also that $\nu \approx 6$, the largest value in Figure 2.5c, still deviates substantially from an additive logistic normal, and both deviate from the normal.

pro-choice is consistent with support for the death penalty, no rule of American democracy is violated. Of course, philosophers can take normative positions, and political scientists can evaluate them systematically, but when we take on board normative views as if they are fixed features of the world, we can wind up with misleading conclusions.

These normative assumptions are so embedded in our empirical analyses that we can even miss that they are assumptions. The problem may be easier to see in older literature, on which much of our present empirical work is built. For example, consider the American Political Science Association's famous report, "Toward a more responsible two-party system" (APSA, 1950), which set the agenda for a generation of American politics researchers. The leading political scientists of the time wrote that when party positions and voter decision making are not based on the issues scholars deemed important, then "Party responsibility at the polls thus tends to vanish. This is a very serious matter, for it affects the very heartbeat of American democracy". They even clarified that "Those who suggest that elections should deal with personalities but not with programs suggest. . . that party membership should mean nothing at all" (APSA, 1950). (We should give the authors of this report a break, written as it was before most of the methodological developments in the social sciences, but, from a modern perspective, the report reads as breathtakingly reckless, with recommendations for numerous major reforms squeezed into single sentences, and all based on unevaluated normative assumptions and little systematic evidence.)

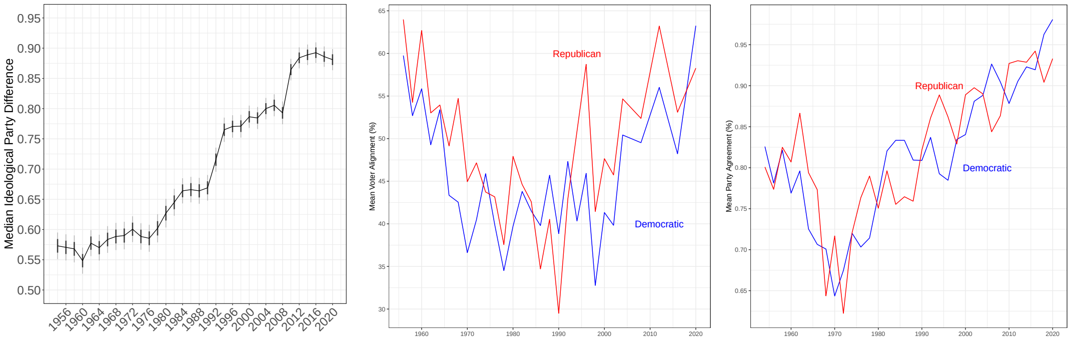
We might also ask whether these normative assumptions are merely reasonable viewpoints that no one would disagree with? After all, few have objected in the literature. For that matter, who would object to the claim that voters should cast ballots based on government programs rather than personality or temperament? Well, as it happens, we live in a representative democracy, not a direct democracy, and in most other situations where a person needs to be selected to do a job, temperament is a crucial factor. Personality evaluations are routinely made for job searches throughout the economy, choosing a romantic partner, picking an instructor, and in many other situations. Even if we could agree on the important issues of the day, new issues always arise after election day that cannot be the basis for voter decisions. In other words, one reasonable normative perspective is that voting should be based at least in part on subjects other than policy issues and programs. And whether we agree with this normative claim or not, it is perfectly consistent with the rules of American democracy: The decision makers are voters, not political philosophers.

Empirical Evidence

In Section 2.4, we described Figure 2.5a as showing that the distribution of vote predictions was just as concentrated around its mean in the 1950s as it is now, and much less concentrated for the years in between. From a casual reading of the literature, this result seems awfully surprising: Where does it say that the political parties were as coherent, internally organized, and distinct from each other in the 1950s as they are today? The 1950 APSA report was designed to fix the lack of coherence in the parties, after all. How can it be that "the era of consensus," with Eisenhower as president and the parties in broad agreement over the cold war, economic prosperity, and support for international alliances like NATO, was as partisan as the 2010s and 2020s, with the gulf in ideological differences so large they seem impossible to span?

But wait, it is worse! Consider a direct measure of ideological polarization over time in Figure 2.6a, measured by a time series plot of the difference in DW-NOMINATE scores between the median Democratic and Republican members of the House (see McCarty, 2019). This figure shows a nearly monotonic increase in ideological polarization over the entire period, very low in the 1950s and very high in recent years. So why then would Figure 2.5a imply that the 1950s were highly partisan? The answer is that the 1950s were highly partisan, but the distinction between the parties was not based on the notions of ideology that political scientists and political philosophers happen to think are important. In fact, Figure 2.6a does not show that party polarization was at a low point in the 1950s; it highlights the failure of the political science concept of ideology to accurately describe this earlier period.

Scholars in the 1960s were aware of these patterns but they used them mostly to declare their dissatisfaction with how voters make their decisions. The leading empirical book of the time, *The American Voter* (Campbell et al., 1980), showed empirically that voters were intensely partisan, but not very informed on the issues these political scientists decided were important. Of course, by definition, the voters were highly informed on the issues they chose to pay attention to, which can be seen by their highly predictable voting patterns. Voting decisions were based largely on partisan identification, which in turn was based on stable and measurable factors like group identities, such as race, religion, and union membership, and parental



(a) Ideological Legislator Alignment

(b) Party ID–Vote Agreement

(c) Legislator-Leader Party Agreement

Figure 2.6: Ideological vs. Partisan Alignment.

socialization.

We can also convey these basic empirical facts in simple time series plots. We do this in Figure 2.6b, by plotting percent agreement between party ID and the vote in ANES surveys, and Figure 2.6, for the percent agreement between members of the House and their party leaders (among roll call votes where leaders of one party oppose those of the other party). Both figures are characteristically U-shaped, mirroring our concentration graph in Figure 2.5a. (Note that the nadir of the time series comes earlier in Figure 2.6 than 2.6b, consistent with the idea that changes in voter behavior are mostly elite driven.)

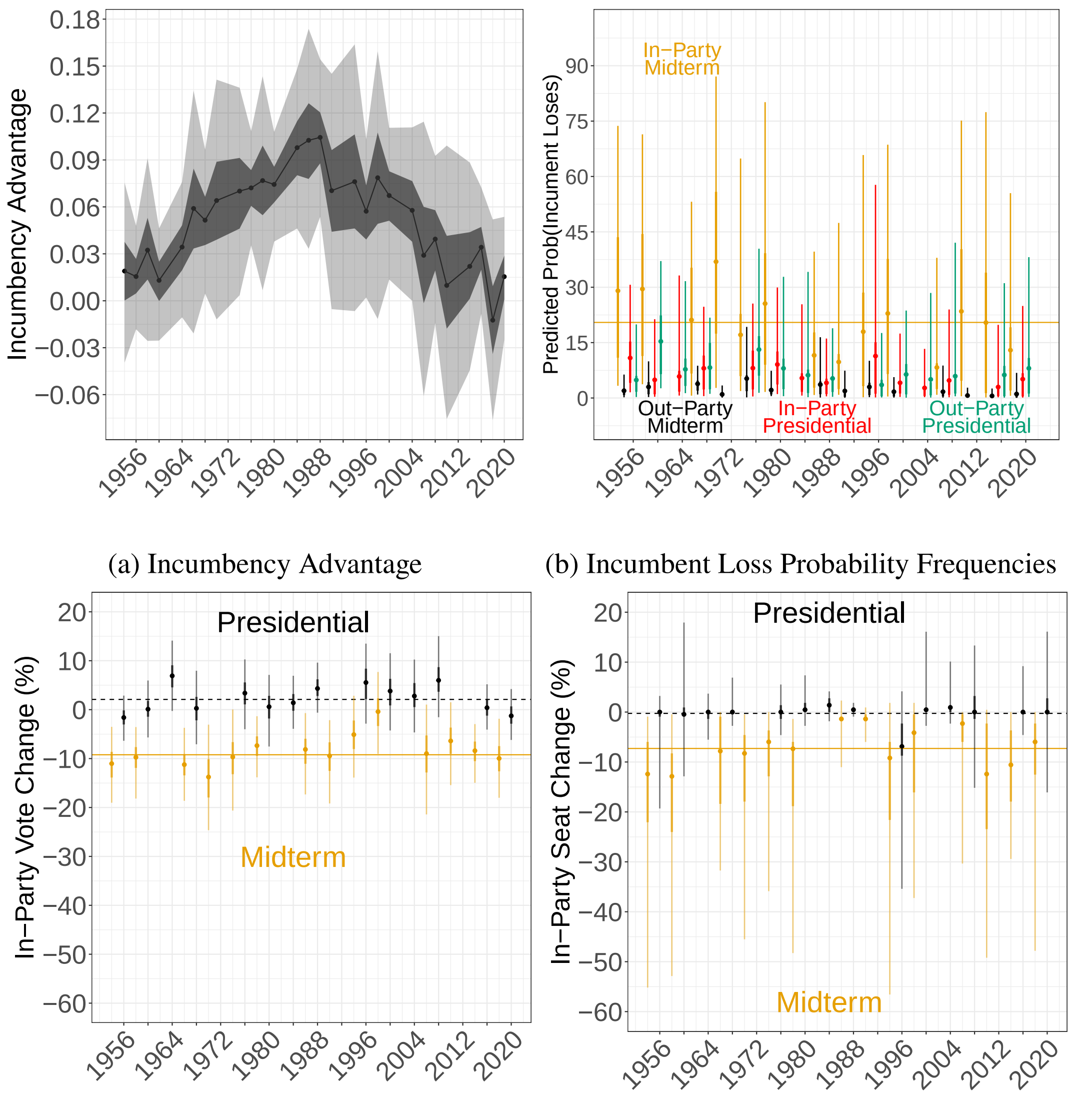
Finally, note the asymmetry in the graphs we present here: for party differences, we give results among voters (Figure 2.6b) and legislators (Figure 2.6), but for ideological differences, we only present differences among legislators (Figure 2.6a). Why no graph for ideological differences among voters? The reason is that ideology is an idea invented by philosophers and used by political scientists; it was relatively unknown among voters until recently. In fact, questions about ideology were not even asked in the American National Election Survey until 1972 and even then prefaced with an explanation: "You may have recently heard a lot of talk about left/right. . . " Ideology is a normative idea that academics impose on voters, not necessarily one that voters chose to use themselves.

Changes in Incumbency Advantage, Stability in Incumbent Loss Probabilities

Section 2.4 shows the consequences, in terms understanding or misunderstanding empirical results, of substituting our own normative preferences for those of voters. In this section, we show the consequences of choosing a quantity of interest that we happen to find of interest and missing a related one of central importance for democracy. A fundamental question for any democracy is the responsiveness of its legislators to constituent preferences, and whether elections produce consequences for violating the voters' will. Mayhew (1974) famously noticed that this guarantee appeared to be breaking down in the 1970s given the decline in the number of competitive elections and what appeared to be an increase in estimates of the electoral value of incumbency (see also Abramowitz and Webster, 2016; Ferejohn, 1977). Studies of these "vanishing marginals," and corresponding increases in incumbency advantage (Gelman and King, 1990; Jacobson, 2015), were a major concern to generations of scholars. However, win margins and expected increases in incumbent votes, as important as they are in and of themselves, are only indirect indicators of the relevant quantity — *the probability that an incumbents will lose his or her job* in the next election. And it is the probability of losing one's job that is likely to be the motivating factor in keeping incumbents responsive to constituents and the whole democracy working. We show here that the broad regime changes in American politics described in Section 2.4 counteract the expected advantages of incumbency, leading to long term stability in the risk of incumbents losing their seats. Moreover, this probability of loss is not only stable, it has been high over the last two-thirds of a century and across the three different electoral regimes we identify in Section 2.4, precisely because of the patterns identified there.

We begin with the familiar electoral advantage of incumbency, plotted over time in Figure 2.7a. For each year, the figure reports the expected vote for an incumbent minus that for a nonincumbent, with all else held constant. If we add appropriate identification assumptions, as in (Gelman and King, 1990), the vertical axis of this figure can be interpreted as an estimate of a causal effect, the expected increase in the vote for a party that comes solely due to nominating the incumbent for reelection as compared to the best available nonincumbent willing to run. This incumbency advantage was about two percent in the 1950s and 60s, increased to about ten percentage points in the 1980s, and then dropped back down again to around two percent by the third regime after 2000 (as noted by Jacobson, 2015).

Most of the information in incumbency advantage estimates comes from the difference between the vote for incumbents and open seat candidates of the same parties. Each of these two components are strong functions of the national swing in any one year, which itself is of course closely related to the probability of an incumbent loss. This means that the value of the incumbency advantage, based on the difference,



(a) Incumbency Advantage

(b) Incumbent Loss Probability Frequencies

(c) Vote Penalty

(d) Seat Penalty

Figure 2.7: Measures of Electoral Competitiveness.

is mostly unrelated to the national swing. Thus, for clarity in Figure 2.7b, we give estimates from our model of the probability of incumbent loss for in-party members during midterm years, where there is a well known large predictable negative national swing. The gold dots in this figure represent the average incumbent loss probability for all in-party midterm incumbents each year, with thick bars corresponding to the central 50% of the district loss probabilities and thin bars capturing 95% of them (i.e., these are not confidence intervals, representing uncertainty; they instead describe the distribution of district-level probabilities).

As Figure 2.7b reveals, the in-party midterm loss probabilities are large and variable, but do not trend over time. The average probability of an in-party incumbent loss during a midterm, represented by the horizontal gold line, is a substantial 20.6%. The vertical lines through the dots indicate that many incumbents have much higher probabilities of losing their jobs, which is indicated by the high end of the asymmetric intervals around the gold dots. The other three logical subsets have much lower average loss probabilities; these include in-party presidential in red, out-party midterm in black, and out-party presidential in green. Although these other three subsets have very small average loss probabilities, the competitiveness of the presidential election means that incumbents will sometimes wind up facing voters with a remarkable one-in-five chance of losing their jobs. Of course, nonincumbents in open seat races have much higher probabilities of losing and incumbent challengers's chances of losing are higher still. If you are or hope to be a tenured professor, think of how much more you might pay attention to the chair of your department, review committee, and students if every four or eight years one in five tenured professors where summarily fired. Your laurels would not be very restful. Of course, this is excellent news for the incentives American democracy provides to its elected legislators to be responsive to their constituents.

Why, then, does incumbency advantage change so dramatically in Figure 2.7a even as the probability of incumbent loss remains so stable in Figure 2.7b? Indeed, these seemingly contradictory results are both computed from the same run of the same generative model. The answer comes from the results in Section 2.4: When incumbency advantage is low, near the beginning and end of our 66 year data set, variation is low and voter concentration is high, meaning that even a 2 percentage point incumbency advantage has some substantial value. When the expected advantage of incumbency rises to roughly 10 percentage points in the middle of the period, the concentration and thus the value of that expected vote decreases by twice as much (from about 60% at the start and end, to only 40% in the middle, of Figure 2.5a). This increasing variability means that incumbents see little actual reduction in their probability of losing office. Getting a bonus of 10 percentage points (because you are an incumbent) may seem comforting, but if this "bonus" also comes with a much larger random component its value is degraded (see also Jacobson, 2015).

Finally, we summarize the consequences of these probabilities for the in-party's loss of votes in Figure 2.7c, and seats in Figure 2.7d. The average loss during midterm elections, represented by the gold flat lines, reflects about an 8.1 percentage point vote loss (±2.3 points, a 95% CI) and 8.8 percentage point seat loss (±1.7 points). These average effects are substantial, but do not miss the occasionally large and highly asymmetric confidence intervals in Figure 2.7d, meaning that we should also expect occasional extremely large in-party seat losses.

Consistent with Figure 2.7b, we also see little to no in-party vote or seat change during presidential years, which is reflected in the black dots and lines in Figures 2.7c and 2.7d.

2.5 Generatively Accurate Descriptive Summaries

We attempt in this paper to build *generatively accurate descriptive summaries* of our data, reducing the tremendous complexity of American politics and congressional elections to understandable summaries computed from a single internally consistent statistical model. While the cost of working with generative models is the modeling assumptions, the benefits include rigorous out-of-sample validation (see Section 2.3) and a far richer range of substantive political science questions that can be tackled, a topic we take up in this section.

Description is sometimes regarded as separate from inference and unaffected by the usual threats to proper statistical analysis (i.e., often as long as you say you're doing "mere" description, anything goes). In practice, however, the best descriptive summaries are those vulnerable to being proven wrong (and then ideally not actually wrong) and tailored to the many precise questions of substantive interest. In fact, descriptive summaries are essential to addressing the breathtaking range of questions of interest to social scientists. Scholarship should not be limited to quantities that happen to be computationally or statistically convenient, or those in whatever methodological area happens to have made progress lately (such as causal inference in recent years; see Supplementary Appendix C.7).

We outline in this section some of quantities that can be estimated from a generatively accurate model and explain how they can be used to enrich political science research. As inference is simply "using facts we know to learn about facts we do not know," we characterize the types of quantities we may wish to estimate by first detailing both the "unknowns" that may be of interest and then the "knowns" we have available to condition on. We characterize the unknowns in three ways. First, the *location* of an unknown is where the values are of the outcome variable that we want to know. This may involve a "forecast", i.e., into the future; "farcast," i.e., to an election in the present or past not in our dataset (such as for a different office or country); "nowcast," i.e., to unobserved features of elections in our dataset (such as the posterior distribution for a district vote, only one value of which is observed, as in posterior predictive checks); or even a "faroutcast," which refers to values of the outcome variable under counterfactual conditions (such as if no incumbents had run). Second is the *level* of aggregation of the quantity of interest, such as for district-, state-, regional-, or national-level statistics, or features of non-geographic groupings like all Democratic districts or all those without an incumbent. Finally, our quantities of interest involve a *concept*, such as partisan bias, electoral responsiveness, the probability that an incumbent will lose, the expected vote in a district, or the district vote of the median legislator.

Quantities of interest always condition on three types of features that are either known or, in the case of counterfactuals, assumed known. These include (1) the choice of covariates and their values for unknown quantities; (2) keeping, removing, or zeroing out random effects; and (3) keeping, removing, or adjusting surprises (such as to focus only on the expected value or other features of the posterior). We explain how to make decisions for choosing quantities of interest, such as those given in Section 2.4, and how to mix and match the location, level, and concept of the unknowns with the covariates, random effects, or error term surprises to condition on.

Consider estimating the probability that a Democrat wins a particular district i in election year t. The simplest case is a "nowcast." Here we are interested in the ex ante probability that the Democrat would this district election, which in fact we have already observed (ex post). To do this, we set the values of X to their observed district values. For example, suppose the lagged vote for the Democrat incumbent is 74%. We might then consider setting $\beta_{tk} = \hat{\beta_{tk}}$ for all k covariates, and $\gamma_i = \hat{\gamma}_i$ and $\eta_t = \hat{\eta}_t$ for the random effects. Of course, we do not know any of these numbers for certain ex ante and, therefore, we would choose instead to include estimation uncertainty in our estimate of the probability that the Democrat would win this district. Thus, instead of fixing these parameters at their point estimates, we draw them from their posteriors, centered on the estimated values. Then, suppose we take 1,000 draws from this posterior; to generate our model-based "nowcast" of v_{it}, we then multiply each of these draws by the relevant X_{itk} and add the draw of the district effect and national swing. We then run this sum through the inverse logit function to get a hypothetical draw on the scale of votes. This generates 1,000 hypothetical draws of v_{it}. To estimate the probability a Democrat wins the district, we then simply count up the fraction of the draws greater than 0.5.

For forecasting, the choices about how to construct generatively accruate descriptive statistics is more flexible, and thus more complicated. Consider the simplest case. If all we want is a one-election-ahead forecast, we still have a number of important decisions to make. For example, how do we set the covariate values? If it is the next election, we have observed the lagged vote, so that is straightforward to use, but what about incumbency? Do we know yet if the current incumbent will run again? If so, we could use that value. But if we are making the forecast well before the election, and do not know this yet, what value should we use? We could assume they all run, or some randomly selected proportion run again. Perhaps, however, it is better to consider what would happen if the district were open? Regardless, the analyst must choose some value relevant to the question at hand, and must realize that this choice changes the question we are answering. In fact, the differences in forecasts across these assumptions may be of considerable substantive value.

We also have important decisions about the district effect, γ_i. If this is were only one election ahead, and we do not think much else has changed, then we may want to fix $\gamma_i = \hat{\gamma}_i$ as we did in our "nowcast" above. However, we surely do not know β_{tk}. So instead we need to use draws of it from $\beta_{tk} \sim \mathcal{N}(\hat{\beta}_k, \sigma^2_{\beta_k})$. This will add additional uncertainty to our forecast, but is otherwise similar to our "nowcast". And we are unlikely to know the national swing and so must make a parallel choice about η_t. Given all these assumptions, we can then generate our hypothetical election draws and calculate the fraction of times the Democrat wins as our prediction as a probability.

Perhaps the most difficult set of choices comes from in making a "faroutcast," as for example, when we want to forecast what would happen in a new legislative map following the implementation of a proposed (or perhaps recently passed) redistricting plan. Here, the covariate choices are not obvious. First, we generally will not know where incumbents will be, but perhaps we can make some educated guesses. Alternatively, we might assume all seats are open to obtain a baseline probability that a Democratic candidate could win the seat. Harder yet, is what to do about lagged vote, which is giving the model a measure of the normal vote in the district. We could use precinct level returns for the previous election, subtract out the incumbency advantage and uncontestedness, aggregate into the new districts, and then add back in the incumbency advantage for districts where the decisions of incumbents and challengers is known. Or perhaps we could use presidential vote, or some average of statewide votes re-aggregated in the new district map. These constructed measures would be needed in the original model or in a separate model that imputes lagged vote from some statewide measures. Also, as with our forecast, we do do not know β_{tk}, γ_i, or η_t. And as before we could then generate our hypothetical election draws and calculate the fraction of times the Democrat won.

The flexibility of the model easily enables one to calculate even more sophisticated quantities of interest. For example, one of the largest sources of uncertainty in election predictions is the national swing. We can thus draw η_t directly from its posterior. Alternatively, we can model the parameters of the η_t prior with national-level covariates, such as unemployment, presidential approval, or whether the country is at war. Yet another option is to fix it at the value of some previous election that seems similar to the current one.

By combining the location, level, and concept for a quantity of interest and fine tuning by making choices about the covariates, random effects, and surprises, accurate generative models like the one we describe here can reveal a vast amount about American elections, far richer than any one specific estimate or data analysis on its own.

2.6 Concluding Remarks

Commonly used models of district-level election results have enabled political scientists to learn a wide variety of information about American legislative democracy. But the observable implications of these models fail spectacularly quite often in ways that should almost never happen. We build on this existing approach by adding features of elections political scientists have learned over the years, and building on new statistical and computational technology not previously available. We validate our approach with extensive out-of-sample (and distribution-free) tests in 14,710 district-level elections. Our generative model is general in that it can be used, with the appropriate additional assumptions and covariates when necessary, to estimate almost any quantity of interest in the literature, and others, all with calibrated (i.e., accurate) probabilities and honest uncertainty intervals.

We apply the model to estimate one of the most central requirements of any representative democracy — the extent to which legislators have a serious chance of losing reelection. We reveal this number to be quite high and remarkably constant over more than half a century, a time period which we show has seen dramatic changes in many other important characteristics of electoral politics such as the incumbency advantage. We then build a more general model of American democracy consistent with these findings.

Further growth in computational power may one day enable feasible estimation of joint generative models that enable a richer substantive portrait of the electoral system, such as conducting modeling at the precinct-level to include redistricting periods, or encompassing other elections such as for the US senate, president, and state legislatures. With a continual focus on rigorous out-of-sample validation, and larger generative models, it may even be possible, one day, to estimate these simultaneous with other sectors of society such as the economy, demography, public policy, and public health, or potentially data from other countries.

Chapter 3

TENSOR-BASED IMPLEMENTATION FOR MULTI-LAYER CONTEXTUAL TOPIC MODELING : APPLICATIONS TO U.S. CONGRESSIONAL LEADERSHIP AND AGENDA SETTING POWER

3.1 Introduction

Given new sources of text data, researchers in social science have increasingly turned to natural language processing as a tool for understanding and analyzing important behavior of both political elites and the public. Text data offers a unique opportunity to generate high frequency data with increasingly large data sets, some of these data sets ordering on the magnitude of tens of millions of observations with thousands of features (see: Steinert-Threlkeld, 2018, Salganik, 2017). Initially, researchers employed statistical models for text analysis to classify open-ended responses to survey questions. Early attempts to classify such responses involved handcoding the responses; however, such an approach faces two key problems in modern social science research agendas. First, given the increasing size of text data, hand-coding data is increasingly impractical. Second, such methods face risk to the external validity of the outputs, even under ideal conditions. Of course, in practice, labelling occurs in less-than-ideal conditions: to achieve labels at meaningful scale, researchers often employ cost-effective methods such as mturk. Unfortunately, these methods often produce low-quality data. Even in the best case, human hand-coders may simply make errors as they attempt to classify labels, either through unintentional error or by failure to comprehend the underlying nature of the labelling task.

Recently, researchers in social science have adopted methods of text analysis that often rely on classifying text documents based on applications in computer science and natural language processing. Computer scientists have rapidly developed methods of Natural Language Processing over the last two decades. Social scientists have increasingly employed these methods as a means of dimension reduction in order to better analyze the structure of text data. The most popular methods include mixed-membership models, where topics are classified as probabilistic mixtures over groups. In social science applications, the two most popular methods are

Structure Topic Models (STM) and Latent Dirichlet Allocation (LDA). Blei, Ng, and Jordan, 2003 developed LDA as a fully unsupervised method for identifying topics from clusters of words and then classifying documents into mixtures of these topical clusters. The model's computational tractability and flexibility has largely driven its widespread adoption across a variety of political science contexts.

Given the increasingly large amount of text data now easily accessible to researchers, especially social media data made available through public Application Programming Interfaces (APIs), there is a pressing need for increasingly tractable estimation methods such that they have practical convergence times and make use of reasonable memory budgets. Using the latest topic modeling methods, for example, STM or LDA could tractably estimate 100,000 documents within a reasonable amount of time. However, both of these models are usually estimated via variational Bayes and Expectation Maximization. These methods are computationally expensive and in large-data applications, often face problems with costly memory usage and impractically slow convergence. If social scientists wish to study a dataset containing all tweets relating to a massive political phenomena, such as the social media response to mass protests or national elections on Twitter, datasets could reasonably reach 50 million or more documents. Such methods could take weeks or months to converge; more pressingly, without the aid of advanced CPU architectures with production-grade memory budgets, the methods are completely infeasible to estimate on the entire dataset. In these cases, certain lines of research will be blocked to analysts with only standard workstation capabilities.

To this end, using tensor-based spectral decomposition of lower-order moments allows for tractable estimation of large data on a variety of single-level latent layer models, such as LDA. That said, both LDA is a clustering method that is largely agnostic to domain knowledge from the researcher. STM requires metadata to measure prevalence. When we consider text as data, it is important to note that political actors may discuss the same topics, such as climate change or immigration, but their underlying contexts towards those topics could cleave both within and between political parties and other relevant groups, such as lobbying groups and special interests. STM addresses these problems by incorporating meta data into the classification of topics. In order to account for contextt orientation in a weakly supervised fashion in text and within the context of LDA, Lin and He, 2009 proposed a model of Joint Sentiment Topic analysis. In certain contexts where functional flexibility is important to the research application and context is a key component of the lexical space, Joint Sentiment Topic (JST) is potentially preferable to LDA and STM. Building on the work of Anandkumar et al., 2013, (which guaranteed the theoretical tractability and identification of the model) and Huang et al., 2015 (which proposed a practical implementation), and Kangaslahti et al., 2023 (which proposed a fully online and end-to-end GPU implementation) this paper will propose a tensorized implementation and small extension of JST modeling to contexts beyond sentiment and show application to a dense data-set of public floor speeches in the U.S. House of Representatives. The paper builds largely on the architecture that was proposed in Kangaslahti et al., 2023 by incorporating context and applying it to a large corpus of political text.

This new estimation approach offers benefits in addition to speed. First, the batched approach allows the model to scale to large data that would be otherwise infeasible to fit a topic model, by reducing the memory overhead. Not only that, but by taking advantages of a GPU-based implementation, the batched approach allows the model to be estimated even on a standard workstation end-to-end on GPU backend. By reducing the memory costs and implementing an estimation routine that converges in reasonable time, the method will reduce the user-related costs of employing these methods on large data. The would make these estimation methods accessible to a wider array of researchers, even those who do not have access to the most advanced servers or sophisticated work stations.

This paper makes two contributions – first, the paper extends tensor-based estimation to a model which incorporates the underlying context of words in a hierarchical fashion, Multi-Layer Contextual Topic Modeling (MLCT). Then it shows that this method improves on speed and identifying partisan differences in speech, while maintaining lexical coherence. This is particularly useful for political domains because the domain effects could be strong given differing political contexts- by utilizing a weakly supervised notion of context, MLCT allows for a more flexible analysis of contextual patterns in political speech.

3.2 Current Methods

Given the increasingly large amount of text data now easily accessible to researchers, especially social media data made available through public APIs, there is a pressing need for increasingly tractable estimation methods. Among the most popular of these methods is Latent Dirchilet Allocation (LDA) – a fully unsupervised method of uncovering topical clusters and mixtures over those labels (Blei, Ng, and Jordan, 2003).

Political scientists have innovated with respect to text methods in order to better understand temporal aspects of speech and aspects relating to the individual actors issuing the statements under analysis. For example, Quinn et al., 2010 treats the unit of speech not as the individual speech, but instead the day's aggregated speeches in the U.S. Senate are treated as the unit of observation. They estimate the topical composition of the aggregated speeches within each day representing a topical mixture. Incorporating this temporal aspect of speech while exploiting the mixture-based nature of LDA classification, this model produced estimates of the daily attention to distinct political topics, to track what the Senate was talking about over a long time series. Employing another Bayesian Hierarchical framework, Grimmer, 2010's "expressed agenda model" measures the attention paid to specific issues in senators' press releases, hoping to derive the messaging space and strategy of individual senator's press and communications offices.

As Grimmer and Stewart, 2013 observe, methodologists need to both be in conversation with computer scientists and develop new models of their own to address the evolving set of questions that may be addressed by text data. With increasing computational power, political scientists have begun to explore powerful word-based methods such as word-embeddings which incorporate context, unlike unsupervised topic models. Already, such models have been successfully employed to discern partisan valence of speech, as well as contextual orientation (Rheault and Cochrane, 2020; Rudkowsky et al., 2018).

Because they incorporate context, there is great excitement around many of the latest word-embedding methods in computer science, such as BERT (Devlin et al., 2019). Even more has been made of Large-Language Models and their promise for a universal method for many task. These methods are exciting as they incorporate grammars in a computationally tractable way and can produce outputs for many types of task. However, these innovations come at a cost: unlike many of the most unsupervised or weakly supervised methods, these frontier models require extensive pre-training, often on pre-labelled data. Even with fine-tuning, the models will struggle in highly context-dependent settings that are outside the training set. In the case of BERT or ChatGPT, the model has been pre-trained on a large, expansive internet crawl. Reassuringly, Rodriguez and Spirling, 2022 validate that these pre-trained methods still work well across a variety of contexts within that training set. Datasets that are highly historical, contemporaneous, or generated via social media fall outside that training set. As it happens, these are exactly the types political and social science contexts that researchers wish to study.

What's more, for supervised methods preferred by many computer scientists, the available social science data is both voluminous and unlabelled and contemporaneous, with shifting underlying dynamics.

Datasets containing billions of tweets relevant to the Black Lives Matter Movement, COVID, or #MeToo are potentially ripe for social science study – but these datasets are both large-scale and extremely dynamic. Given the scale of many of the latest data sources and the dynamic nature of the political science contexts where these data are most valuable, frontier models face hurdles in terms of expense of labelling and training, as well as infeasibility of hand-labelling dynamic data in real time.

In addressing these vital contexts, weakly supervised methods for topic classification have shown their potential to further understand social behavior. In addition to these feasibility constraints, prior research has shown that computer-assisted methods better recover conceptually meaningfully aspects of political science text data better than their fully automated (such as word embeddings) or fully expert-coded counterparts. (see Grimmer and Stewart, 2013, Grimmer and King, 2011). Researchers need ways to incorporate known structure on these vast datasets. To this end, STM allows for the incorporation of metadata into textual analysis by allowing for a linear functional form (Roberts et al., 2014b) over metadata that allow analysts to uncover latent topics while studying how the prevalence and content of topics change with respondent and document metadata. Similarly, JST allows for the incorporation of a prior over a subset of words. Both methods offer potential ways forward in terms of studying these applications in political science.

3.3 Model Comparisons

Ensuring Model Consistency

In order to ensure a consistent model comparison, we ensure that we are comparing models holding as much equal as possible. We use the same corpus of tweets and vocabulary across all three models. Additionally, we compare the models at the same number of topics or senti-topics. Finally, we selected a data environment where we believed all of the topics models should perform well – a dataset of legislative tweets. This data has rich topical structure, obvious labels and and easily measured partisan separation. Finally, we compare these weakly supervised methods to a fully unsupervised LDA baseline.

Model Metrics

We compare the models on two dimensions. The first is how well the models recover coherence topics. One important aspect of text analysis is recovering conceptually meaningful topics or clusters. We report a standard measure of topical coherence to obtain an objective quantitative comparison. Second we compare the ability of each model to recover partisan valence of the legislators from their speech. Given that the party label of legislators is known, we should expect that these models should be able to recover it. We expect STM to perform best on this measure, as party labels can be directly incorporated into the generative model as metadata. That said, if JST performs at least similarly, that would be suggestive of its utility as an insightful alternative when metadata quality is of low quality or unavailable to the researcher. This is because it is recovering conceptually meaningful structure in the data.

Coherence: First, we compare the models based on coherence. We use the normalized pointwise information metric combined with a cosine formulation. We follow Röder, Both, and Hinneburg, 2015 who find that a Cosine Similarity metric on the Normalized Pointwise Mutual Information metric (NPMI) of top words is the best performing coherence metric in a survey of direct and indirect measures of topical coherence.

The basis of the metric is the NPMI vector, calculated for each words in the vector of top 20 words per topic:

$$\text{npmi}(w_i; w_j) \quad = \quad \frac{\log_2 \frac{P(w_i, w_j) + \epsilon}{P(w_i)P(w_j)}}{-\log_2 P(w_i, w_j) + \epsilon} \quad \forall w_j \text{ in the top 20 word topic } k$$

The metric is then constructed as follows. First, an vector is computed for each word w_i in the top twenty words against all remaining words $w_j \neq w_i$, yielding an NPMI vector for each word as described in the preceding paragraph. Then, the metric calculates the average of the cosine similarities of each of the the NPMI vectors for each top word.

Cluster Validity: Second, we examine Principle Components of the mixtures generated by LDA, STM, and MLCT to assess how well they recover known partisan labels in the data. We report on standard metric of clustering validity, (Meilă, 2003)'s variation of information:

$$\text{VI}(\hat{L}; L) = -r\left[\log(b/d) + \log(b/r)\right]$$

where

- $\hat{L}$ are the party labels uncovered from a Principal Components Analysis (PCA) decomposition on estimated topic mixtures from a topic model
- L the true party labels
- b the number of legislators where the labels L and $\hat{L}$ disagree on party label
- d the number of Democratic legislators both labels L and $\hat{L}$ agree are Democrats
- r the number of Republican legislators both labels L and $\hat{L}$ agree are Republican

The intuition behind this metric is to measure the distance between sets of partitions. When the automated party labels are close to the true labels, the metric will be smaller.

3.4 Tensor Preliminaries

Innovations in spectral decomposition methods from low-order tensors has allowed for increasingly parsimonious estimation of latent variable models. To connect to previous work in this area, we introduce the same tensor notations borrowed from and consistent with Anandkumar et al., 2014. A real p-th order tensor $A \in \bigotimes_{i=1}^{p} \mathbb{R}^{n_i}$ is a member of the tensor product of Euclidean spaces $\mathbb{R}^p$. For a vector $v \in \mathbb{R}^n$, we use $v^{\otimes p} := v \otimes v \otimes \cdots \otimes v \in \bigotimes^p R^n$ to denote its p-th tensor power. In this paper, make use of third-order tensors. Intuitively, the object we are working with is the tri-occurrence of words. (That is, when words appear together up to three times in one document).

Anandkumar et al., 2014 notes that tensor decomposition is delicate, in general. Tensors may not even have unique decompositions. Fortunately, the orthogonal tensors that arise in the present model have a structure which permits a unique decomposition under a mild non-degeneracy condition, which we described below. We are able to estimate an orthogonal decomposition to acquire the eigenvectors and eigenvalues of a 3-order tensor. Note that despite the complications of these decompositions, they have attractive properties noted in Anandkumar et al., 2012, Anandkumar et al., 2014, and Huang et al., 2015. The tensor structure employed in these papers – as well as in this work – is the symmetric orthogonal decomposition.

This decomposition expresses the low-order tensors calculated in this paper as linear combination of relatively small-dimensional forms; each form is the tensor product of a vector and the collection of vectors which form an orthonormal basis. These methods exploit a key property of such symmetric tensors which possess these decompositions. That is, these tensors have eigenvectors that, after some careful algebraic manipulation, reduce to the topic-word probability matrix that is the central object of estimation for topic models. Anandkumar et al., 2013 and Anandkumar et al., 2014 focus on theoretical considerations for such models, and we direct the reader to these works for a more technical treatment of parameter recovery. This work will focus on implementation and application of these methods to a latent variable model based on LDA which incorporates hierarchical context.

We take the basic model from (Anandkumar et al., 2013). We have a random vector $h = (h_1, ..., h_k)^T \in \mathbf{R}^k$, where h are the k latent factors. In Anandkumar et al., 2013 and Huang et al., 2015, h represents the topic-document distribution. In the case of Tensor MLCT, h is the *context-topicc* distribution over topics. In this case, k is the number of contexts multiplied by the number of topics.

Tensorizing MLCT

Suppose we have the following exchangeable random vector: $\{w_1, w_2, w_3, ..\} \in \mathbf{R}^M$. Exchangeability is a key simplifying assumption in both LDA and JST. Blei, Ng, and Jordan, 2003 notes that exchangeability is essentially a notion of "conditionally independent and identically distributed," for LDA-based topic models. In this case, the conditioning is with respect to an underlying latent parameter (h) of a probability distribution, which gives the mixture of size k (that is ($S \times T$) total) context-topics. We assume that the number of documents is weakly greater than the number of context-topics ($M \geq k$), a reasonable assumption given the observed nature of text data. Next, $w_1, w_2, w_3, .. \in \mathbf{R}^M$ are conditionally independent given h. Finally, let there exist a matrix $O \in \mathbf{R}^{Mx(S*T)}$ with row vectors for each word-topic probability ϕ, which are each $1 \times ((S * T)$ for all $v \in \{1, 2, 3, 4, ..., N\}$ where N is the number of words in the vocabulary:

$$E[w_v|h] = Oh$$

Thus, we have the following notation for JST:

- T, Number of Topics

- S, Number of Contexts
- $P(l) = \pi_l$ is the probability document is in context l.
- $P(z|l) = \theta_{l,z}$ is probability that a documents is in the z'th topic conditional being in context l.
- $P(w|z,l) = \phi_z^l$ is probability we observe a word conditional that it is in the z'th topic and in context l.

$$O = \left[\phi_1^1, \phi_2^1, ..., \phi_T^1, \phi_T^2, ..., \phi_T^S\right] \tag{3.1}$$

$$h = \begin{bmatrix} \pi_1\theta_1 \\ \pi_2\theta_2 \\ \vdots \\ \pi_S\theta_S \end{bmatrix} \tag{3.2}$$

with

$$\theta_l = \begin{bmatrix} \theta_{l,1} \\ \theta_{l,2} \\ \vdots \\ \theta_{l,T} \end{bmatrix} \tag{3.3}$$

.

Relabeling to acquire contextual structure: Due to the lack of ordering for spectral methods at hand, the model is identified only up to the atomized context-topics at the word level. Despite this limitation, we can still incorporate and account for the contextual structure of the data in our estimation procedure. To simplify the notation, the paper will consider the relabeled concentration parameters $\alpha_{l,z}$ for each of the *context-topics*. It denotes the mixing parameter $\alpha_{\gamma_0} = \sum_l^S \sum_t^T \gamma_l \alpha_{l,z}$.

Finally, we allow heterogeneity in $\alpha_{l,z}$, but assume homogeneity in γ_l. This significantly reduces the number of calculations in cross-moments of the population moments, reducing the problem to LDA, while still allowing for heterogeneity in the parameters through the underlying topics in each context. Standard JST and LDA implementations usually assume completely homogeneous priors, yet Tensor MLCT still retains some of the flexibility of Tensor LDA as described in Anandkumar et al., 2013.

Non-Degeneracy Assumption: The key non-degeneracy assumption which guarantees parameter recovery in these models under this method is that O is full rank (see Anandkumar et al., 2012, Anandkumar et al., 2013, and Anandkumar et al., 2014). As noted in Anandkumar et al., 2014, this implies that variance-covariance matrix of the data is positive semi-definite with rank k. This is essentially a non-degeneracy assumption and relatively mild, especially in our applications where the data are relatively rich in variation. In datasets where there is colinearity or insufficient variation and the assumption fails, Hsu and Kakade, 2012 shows combining observations (such as in Quinn et al., 2010) can boost the rank of the requisite matrices at hand. Given the size of the vocabularies and number of documents in many of the large-scale text datasets, it is extremely unlikely this condition will fail to hold in more practical applications where this method might be useful to researchers.

As in LDA, we assume h is a distribution itself. We let $\alpha_{z,l} \in \mathbf{R}_+^k$. We interpret h as the distribution over the joint context-topic labels. We assume context-topics are independent and Dirichlet distributed. We have $\alpha_{z,l} = \alpha_{1,1} + \alpha_{2,1} + \ldots + \alpha_{2,1} + \ldots + \alpha_{T,S}$ and $\alpha_{\gamma_0} = \sum_{l=1}^{S} \sum_{z=1}^{T} = \alpha_{z,l}$, so.

$$p_{\alpha,\gamma}(h) = \frac{\Gamma(\alpha_{\gamma_0})}{\prod_l \prod_z \Gamma(\alpha_{z,l})} \prod_{l=1}^{S} \prod_{z=1}^{T} h_{l,z}^{\alpha_{l,z}-1} \tag{3.4}$$

Note that first we have

$$E[w_1] \quad = \quad OE[h] \tag{3.5}$$

Under the simplifying assumption that $\gamma_1 = \gamma_2 = \gamma_3$. As we have exactly the form of LDA, we apply theorem 3.5 and theorem 4.3 from Anandkumar et al., 2014 to directly to recover the parameters $\alpha_{l,z}$, as we have the form following for the second empirical moment:

$$\begin{aligned} E[w_1 w_2] &= E[E[w_1 w_2 | h]] \\ &= OE[hh^T]O^T \\ &= \frac{1}{\alpha_{\gamma_0}(\alpha_{\gamma_0} + 1)} O(\alpha \otimes \alpha) O^T \end{aligned}$$

That is, we assume homogeneity in the mixing parameter for contexts, while retaining heterogeneity in the mixing parameter *context-topics*.

Because singular value decomposition is un-ordered, we cannot explicitly learn the contexts parameters of the model. However, we can still account for the underlying context-topic structure of the data, as in JST. That is, we can incorporate context into the analysis exactly as JST does for sentiment, and recover the context-topic level parameters.This leads to gains in speed over Gibbs sampling based JST and gains in topical coherence over LDA.

3.5 Forming moments

Anandkumar et al., 2013 are able show that the model parameters are identified under a mild rank assumption of the O matrix for LDA. Given the introduction of sentiment/context, we make one further assumption to ensure identification of the re-labeled parameters. As in Lin et al., 2012, this paper employs a weakly-supervised implementation and so we augment the observed document-term matrix and weight by contextual orientation. We note that while Lin et al., 2012 imposes a purely sentiment-based context, this can be generalized to any weak prior over contexts. Thus, the main input for the algorithm is a constructed data matrix of word counts weighted by contextual orientations. For the application to social media data, we use the paradigm list of sentiments employed by Lin et al., 2012. For the U.S. House Speech data, we impose a prior over 80 words in 4 contexts (20 per context). The contexts are economics, foreign policy, social policy, and judicial appointments. Under these assumptions, the Tensor MLCT is atomized, and the estimation procedure reduces to LDA on the augmented sentiment-word document matrix. We can then apply the same estimation procedures as in Anandkumar et al., 2013 and Huang et al., 2015, and Kangaslahti et al., 2023 but on the augmented context-word document matrix (of size $N \times (S * T)$).

In order to connect the word context orientations to empirical observations, the document-term matrix is expanded to a context-word document matrix by weighting by the above contextual priors. For the words not in the prior, the paper assumes a uniform distribution over contexts. For the U.S. House Social media data, we follow Lin et al., 2012, we assume three sentiments corresponding to neutral, negative, and positive. Words in a given sentiment are weighted 0.9, with weights 0.05 for the remaining two. Given varying domain contexts, it is possible that words might exhibit differing sentiment orientations depending on context. This distributional assumption avoids imposing excessively strong priors on the data. Note that we can explicitly recover the individual parameters of h by integrating over all the topics derived for a context – this is despite the fact that spectral methods do not preserve ordering, since we've imposed the weakly supervised structure. That said, in analyzing text data, we can still recover the individual context-topics. Importantly, the Tensor MLCT accounts for the underlying context-topical structure, unlike LDA.

Finally, following Kangaslahti et al., 2023, we center the data to reduce the number of matrix calculations for the higher-order moments. By centering the data, we mean to say that we de-mean the data so that it has a mean of zero. $(X - mean(X))$. Centering the data greatly simplifies the calculation of the higher-order moments because cross-moments cancel out. This reduces the computational overhead significantly. This gain in computational tractability comes at the cost of generating dense matrices, increasing memory overhead. Most methods exploit sparsity of document-term matrix to reduce memory overhead. However, in this case, this model is batched. That is, we incrementally estimate the singular value decomposition of the second moment on pre-prescribed subsets of data, rather than estimate the singular value decomposition on the entire dataset at once. This reduces the maximal memory imprint to $2 * (N_b)$ samples are kept in memory at any given time, where N_b is the number of samples in the batch. Thus, there are gains in flexibility, as the method can be implemented either on a slower CPU backed or on a GPU backend with tighter memory constraint. The researcher can decide which method best suits their use-case, depending on the available computational resources and the size of their dataset.

Finally, let the centered augmented document term matrix be denoted by C with rows $c_t := (c_{1,t}, c_{2,t}, ..., c_{M,t}) \in \mathbf{R}^{M*S}$ denoting the centered frequency vector for $t-$th document where M is the number of words in the vocabulary, and let N be the number of documents. Finally, we will let $k = S * T$ denote the total number of context-topics. Given this set-up, we have the following empirical moments:

$$\begin{aligned}
M1 &:= \frac{1}{N}\sum_{t=1}^{N} c_t && (3.6)\\
M2 &:= \frac{(\alpha_{\gamma_0}+1)}{N}\sum_{t=1}^{N} c_t \otimes c_t && (3.7)\\
M3 &:= \frac{(\alpha_{\gamma_0}+1)(\alpha_{\gamma_0}+2)}{2N}\sum_{t=1}^{N} c_t \otimes c_t \otimes c_t && (3.8)
\end{aligned}$$

where $\otimes$ is the tensor dot product. Following Anandkumar et al., 2013, we have that the moments can be factorized as

$$\begin{aligned}
E[M_1] &:= \sum_{l=i}^{S}\sum_{z=1}^{T}\frac{\alpha_{z,l}}{\alpha_{\gamma_0}}\mu_i && (3.9)\\
E[M_2] &:= \sum_{l=i}^{S}\sum_{z=1}^{T}\frac{\alpha_{z,l}}{\alpha_{\gamma_0}}\mu_i \otimes \mu_i && (3.10)\\
E[M_3] &:= \sum_{l=i}^{S}\sum_{z=1}^{T}\frac{\alpha_{z,l}}{\alpha_{\gamma_0}}\mu_i \otimes \mu_i \otimes \mu_i && (3.11)
\end{aligned}$$

where $\mu = [\mu_1, ..., \mu_k]$ and $\mu_i = Pr(x_t|h = i), \forall t \in [l]$. In other words, μ is the context-topic word matrix.

Anandkumar et al., 2013 showed that μ is recoverable from the singular value decomposition of the third order tensor. The choice of estimating the model from low-order moments up to third moment is for two reasons. The first is practical–estimating moments much beyond the third moment would require either explicitly computing the empirical analogue of an increasingly high-dimensional tensor, which quickly becomes computationally infeasible. The second is that the co-occurence of two terms might not be sufficient to pin down topical meaning. For example, if a researcher observes that the words *blackberry* and *apple* co-occur, they would not be able to discern the topical meaning unless they observed that these words co-occur with the word *phone* or *fruit*. Thus, the third order moment is both feasible to estimate, while reasonably capturing this critical variation in the co-occurrence of terms.

This paper builds on the implementation that Huang et al., 2015 proposes and the online, GPU end-to-end procedure engineered in Kangaslahti et al., 2023. That is,

```
Algorithm 1: High-Level Estimation Procedure
Result: Document Context-Topic Matrix and Word Context Topic Matrix
    1. Construct the augmented context-word weighting matrix
    2. Calculate whitening Matrix $W$ and whiten centered-data from step 1.
    3. Using stochastic gradient descent, estimate the spectral decomposition of
       third order moments tensor $M_3$ implicitly from whitened counts.
    4. Recover context-topic mixture for documents using variational inference.
```

following similar procedures, M_2 is formed implicitly from a singular value decomposition of the centered data matrix. Then the M_3 tensor is implicitly formed using the whitened counts of the centered data. Whitening renders the tensor symmetric and orthogonal (in expectation). Most importantly, it reduces the dimensionality of the third moment from $O(n^3)$ to k^3, the number of topics. Given the nature of speech in political environments, the number of topics is almost always going to be an order of magnitude smaller than the number of words. However, the exact implementation is improved over Huang et al., 2015 is a few key ways. First, the data is centered, reducing the complexity of the computation of the higher-order cross moments. Second, a batched PCA is employed to estimate the decomposition of M_2, rather than k-truncated SVD in order to parsimoniously decompose the data given it is no longer sparse. Finally, the gradient calculation is simplified given the context-topic setting, leading to extremely efficient recovery of the M_3 decomposition. Finally, given we have recovered the context-topic distribution over words, we employ standard variational inference to recover document-level parameters. The paper proceeds by walking through each step of the algorithm,

Using the singular values and singular vectors from the centered data, we construct a whitening matrix W such that

$$W^T M_2 W = I$$

where I is the identity matrix. The whitening matrix is computed via incremental Principal Components. This allows for a batched implementation that reduces the memory constraints of the method. Note this is currently the most costly bottleneck both in terms of time and memory. Areas of future for spectral NLP methods are likely to come from finding more efficient ways to estimate the singular values from PCA on large, non-sparse data. From the centered data, we have:

$$W = U\Sigma^{-\frac{1}{2}}$$

where U and Σ (the variance matrix of the centered data) are the top k singular vectors and singular values of the centered data.[1] In order to estimate the implicit third moment, the method calculates the whitened counts

$$y_t = \langle W, c_t \rangle$$

We will use these whitened counts to construct the implicit third-order tensor. Using that implicit tensor, the method uses a stochastic gradient descent to find the spectral decomposition of the third-order moments.

Stochastic Gradient Descent

Following Huang et al., 2015 and Kangaslahti et al., 2023, this paper implements a fully online method to recover context-topical parameters. We let $v = [v_1|v_2|...|v_k]$ be the true eigenvectors of the third-order moment. We denote the sample size with n_x. Now that with the whitened tensor in hand, the method follows Huang et al., 2015 in implementing a Stochastic Tensor Gradient Descent (STGD) algorithm for tensor CP decomposition.

Now, note the whitened third-order tensor for the centered data ($k \times k \times k$) is

$$\mathcal{T} = \frac{(\alpha_0 + 1)(\alpha_0 + 2)}{2n_x} \sum_{t \in X} y_t \otimes^3$$

.

Finally, the method solves the minimization problem for STGD following Huang et al., 2015, the method solves the optimization problem:

$$\arg \min_{\mathbf{v}:||v_i||_F^2=1} \{|| \sum v_i \otimes^3 - \mathcal{T}||_F^2 + \theta|| \sum v_i \otimes^3 ||_F^2\}$$

[1] Note that given this setting, the singular vectors of the centered data and M_2 are the same, while singular values of the centered data are the squared values of the singular values of M_2. Thus, the second moment is never explicitly formed, saving considerable overhead in terms of memory.

where the first term measures the Frobenius norm between the constructed tensor and the eigenvectors of the third moment. This is meant to encourage similarity between the true quantity and the observed quantity from the whitened tensor. The second term is an orthogonality penalty, meant to encourage orthogonality between the eigenvectors.

Thus, the method minimizes the loss function

$$L(\mathbf{v}) = \frac{1}{n_x} \sum_{t=1}^{n_x} \frac{1+\theta}{2} || \sum_{i=1}^{k} \otimes^3 v_i ||_F^2 - \left\langle \sum_{i=1}^{k} \otimes^3 v_i, \mathcal{T} \right\rangle$$

Simplifying the expression from Huang et al., 2015, denote each stochastic update as $\hat{v}_i^j$. Then taking the derivative of the loss function, the stochastic updates j can be written as

$$\hat{v}_{j+1} \leftarrow \hat{v}_j - \beta^j \frac{\partial L^t}{\partial v_i}\bigg|_{\hat{v}_j}$$

.

Expploiting the centering of the data, we can simplify the expression from Huang et al., 2015, each stochastic update is

$$\hat{v}_{j+1} \leftarrow \hat{v}_j - 3(1+\theta)\beta_j \hat{v}_j \left(\hat{v}_j^T \hat{v}_j * \hat{v}_j^T \hat{v}_j \right) + \frac{3\beta_j(\alpha_0+1)(\alpha_0+2)}{2n_x} y^T (y\hat{v} * y\hat{v})$$

where $*$ is the Hardamard product (column-wise product) and all remaining matrix operations are normal matrix products (for the full derivation, see Appendix). The learning rate β_j scales linearly downward with each iteration j in order to discourage large jumps in late in the search space. Finally, eigenvalues from third-order moment in hand, the method estimates the context-topic word matrix $\hat{\mu}$ by unwhitening the eigenvectors of M_3:

$$\hat{\mu} = W^{T^{\dagger}} \hat{v}$$

where † denotes the pseudo-inverse.

Context-topic Document Distribution Recovery

Finally, to recover the context-topic distribution over documents, the paper employs variational inference based on that first proposed by Blei, Ng, and Jordan, 2003. The authors propose a simplified family of latent variable models defined by

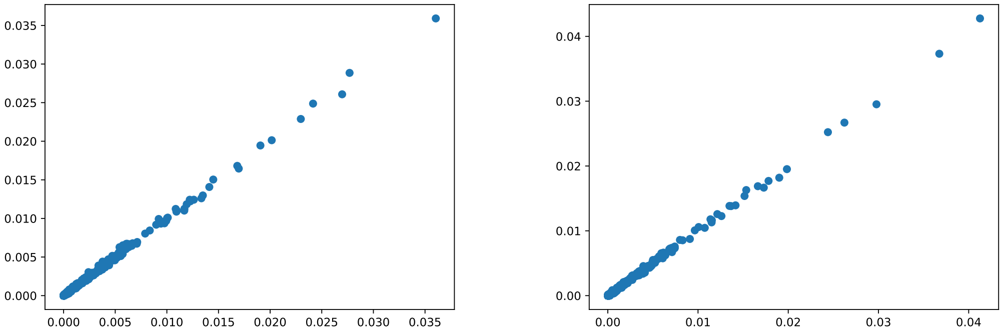
$$q(\theta, \mathbf{z}|\psi, \phi) = q(\theta|\psi) \prod_{t=1}^{N} q(z_t|\phi_t),$$

Note, that the second term above is estimated by $\hat{\mu}$, greatly simplifying the calculation, as we have already estimated the word-level parameters. Thus, the method optimizes the Kullbach-Leibler Divergence:

$$\hat{\psi} = \arg\min_{(\psi)} \int_\theta \sum_z q(\theta, \mathbf{z}|\psi, \phi) \cdot \log\left(\frac{q(\theta, \mathbf{z}|\psi, \phi)}{p(\theta, \mathbf{z}|d, \alpha, \beta)}\right) d\theta$$

In words, the method tries to find the best lower-bound that gives us a a variational distribution most similar to the true posterior in order to estimation the document-level context-topic matrix. We can then find a bound using Jensen's inequality (For further implementation details, see Appendix A.1 of Blei, Ng, and Jordan, 2003).

3.6 Simulations



(a) Simulated Topic 1

(b) Simulated Topic 2

Figure 3.1: Simulated Topic-Word Probabilities under a Two-Topic Model recovered using Stochastic Gradient Descent to recover the Tensor Singular Value Decomposition.

As a simple validation check, we simulate draws of MLCT model and compare Tensor MLCT results for the topic word matrix to the true values. Computationally, we simulate a Tensor MLCT model with 1 context, 2 topics, and 1000 documents, and 100 words. For computational convenience, we simulate on CPU and use the non-streaming estimation routine. As illustrated by Figure 3.1, we find a nearly 1-1 match, suggesting the model is capable of recovering the true Topic probabilities in simulated conditions.

3.7 Application: Model Comparison on U.S. House Twitter Data

Data

For the basis of a model comparison using real data, we follow Ebanks et al., 2021 and collect the Twitter handles of 440 representatives from June 29th, 2019 to March 23, 2020 based on the official Twitter handles list[2] collected by C-SPAN.[3] Although the true model parameters are not known as the data are observed, this exercise allows for a more easily understood comparison of model outputs. We treat the LDA model as the baseline model of comparison, as it is currently standard in the literature for topic modeling and the closest framework to JST. For this analysis, we restrict to exactly the first 300,000 tweets of the 117th Congress in the House, including only original posts and excluding re-tweets. This data is high-frequency text data, the kind of which is well-suited for NLP dimension reduction methods such as LDA and JST. What's more, we know from the context of tweets they are likely to have very small mixing parameters – tweets are likely only contain one or two topics.

Model Comparison

In order to benchmark the model, an LDA model, a standard JST model, and a Tensor MLCT model is fitted on a large Twitter dataset. Ultimately, we expect to see large gains in speed over larger datasets, with minimal cost to mathematical coherence, as well as facial coherence. The standard LDA topic model using a variational Bayes Implementation written in Python Pedregosa et al., 2011.[4]The online and batched version of LDA is implemented to provide the fastest possible benchmark against both versions of JST. We then compare to a Gibbs Sampling implementation of JST written in `C++`.[5] The Tensor MLCT method is written in `python` using a `Tensorly` backend. This backend is optimized for performance both on CPU and GPU processors. For now, since most standard workstations are designed for a CPU backend and a goal of this implementation is to offer a user-friendly method for large data analysis, the method is tested on a standard CPU backend typical of a usual workstation. Finally, the models are optimized on topical coherence. We follow Röder, Both, and Hinneburg, 2015 who find that a Cosine Similarity metric on the Normalized Pointwise Mutual Information metric (NPMI) of top words is the best performing coherence metric in a survey of direct and indirect measures of topical

[2] We did not include election, personal, or private accounts in our datasets.

[3] `https://twitter.com/cspan/lists/members-of-congress/members`

[4] We use the `sklearn.decomposition.LatentDirichletAllocation` written in `python`

[5] Written by Lin and He, 2009

coherence.

The basis of the metric is the NPMI vector, calculated for each words w_i in the vector of top 20 words:

$$\mathrm{npmi}(w_i; w_j) = \frac{\log_2 \frac{P(w_i,w_j)+\epsilon}{P(w_i)P(w_j)}}{-\log_2 P(w_i, w_j) + \epsilon} \quad \forall w_j \text{ in the top 20 words}$$

The metric is then constructed as follows. First, an vector is computed for each word w_i in the top twenty words against all remaining words $w_j \neq w_i$, yielding an NPMI vector for each word as described in the preceding paragraph. Then, the metric calculates the average of the cosine similarities of each of the the NPMI vectors for each top word. On this basis, we optimize and compare each model.

Model	Time to Convergence	Optimal Coherence (NPMI)	Optimal $T * S$)
LDA	2:17:05	0.49	80
JST/MLCT	1:14:23	0.59	30*3
Tensor MLCT	00:25:13	0.53	28*3

Table 3.1: Model Comparison: All three models were run on the same system architecture. Proccessor: Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz, 64gb RAM, 10 cores.

Table 3.1 shows the speed with which each model converged, as well as the coherence achieved. Note that Tensor MLCT converges significantly faster than either LDA or Gibbs-sampling based JST. In fact, Tensor MLCT dominates JST and LDA on speed, and dominates LDA on coherence. But it is comparable, but slightly lower coherence, than plain JST. This is likely due to the approximation of the singular values and singular vectors of the M_2 moment using Incremental PCA[6]. Given this is also the main bottleneck in terms of speed of estimation, a key extension of this method will be improving the incremental PCA method to one that is fully online and batched. That said, the significant gains in terms of total model estimation time outweigh the minimal loss in topical coherence.

We next show the ability of the weakly-supervised MLCT to recover a known characteristic, such a partisanship. We then devise a means to compare its relative ability to recover that characteristics against other popular methods such as STM and LDA.

[6] We use the `gensim` implementation of batched, incremental PCA in order approximate the decomposition of the second moment Pedregosa et al., 2011.

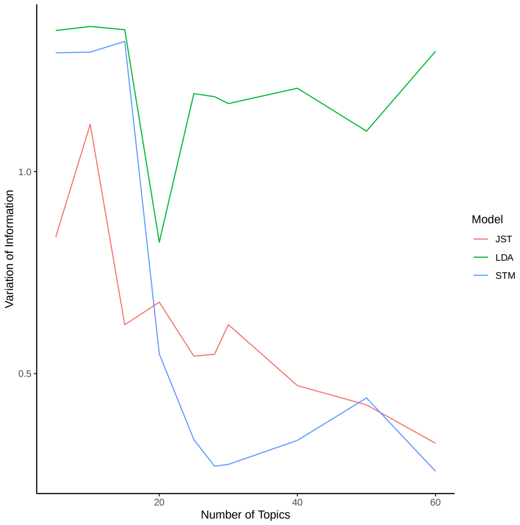
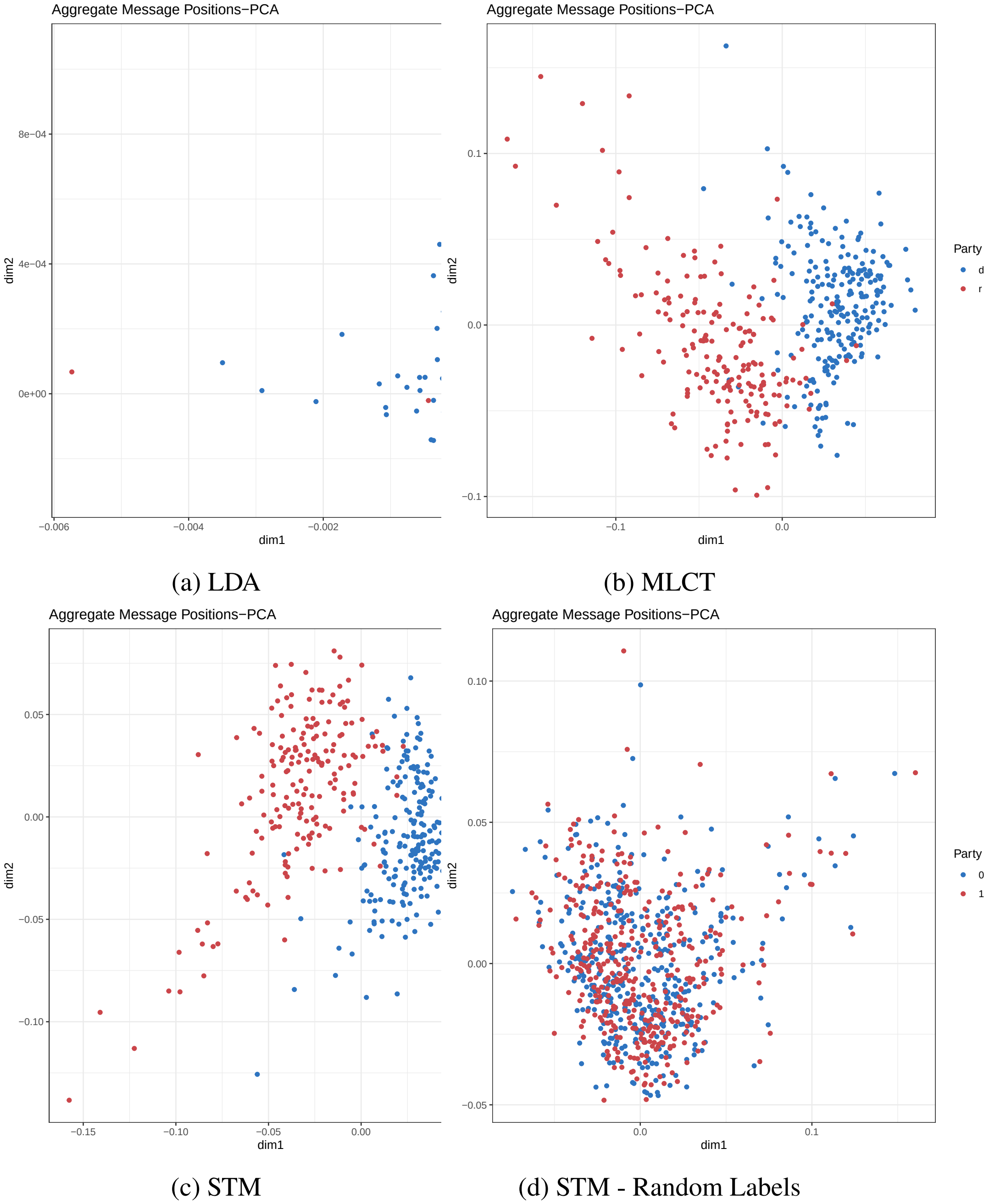


Figure 3.2: Clustering Metrics MLCT vs. STM vs. LDA.

To measure the ability of each model to recover a known characteristics of the data, we scale the average of the mixtures produced by each model for each legislator to recover partisan separation. If the evidence shows that weakly supervised methods are able recover this known characteristic, this suggests they are able to effectively recover a conceptually meaningful aspect of the data. Figure 3.3 illustrates a visualization of the first two principle components of the average mixture by legislator over the entire time period covered by the dataset. As 5 topics was optimal for both STM and MLCT under the coherence metric, we fixed the topic numbers at 5 to provide visual evidence and to illustrate the result. That said, the relative results between models are robust to choice of number of topics, although cluster quality degrades for all models outside the optimal topical range. Moreover, we show variation of information coefficients for all topics below. It is precisely the labels generated by these clusters that we use to compare to the true party labels. Even upon visual inspection, it is clear from Figure 3.3b and Figure 3.3c that JST and STM are capturing the partisan characteristic in the data, which strongly points to their ability to capture meaningful aspects of the data. At the same time, LDA and STM where the party labels are randomized perform poorly at recovering the partisan characteristic. We show in Figure 3.2, the results are robust to topic number.

Figure 3.2 shows that STM generally outperforms MLCT, which comports with expectations – that is, the party labels generated by STM are generally closer to the truth than JST, an unsurprising result given that party labels are included in the training of the model. That said, both models vastly outperform an LDA benchmark.



(a) LDA

(b) MLCT

(c) STM

(d) STM - Random Labels

Figure 3.3: Recovery of Partisan Characteristic using Topic Models.

This is consistent with the visual evidence from the first two principal components presented above.

3.8 Application: Long Run Trends in US Leadership and Agenda Setting in the House

Data, Pre-procecssing, Topic Selection

In order to study the long-run agenda setting trends in the U.S. House, we apply the MLCT model to a dataset of U.S. Congressional Speeches from 1877 to 2015 Gentzkow, Shapiro, and Taddy, 2018. We train the model on 16,852,571 speeches from the House and Senate during this period. Although we train the model on the full corpus (to help best identify the full population of speeches and for comparison in future research on this historical dataset), we restrict our analysis to the U.S. House after the emergence of the formal party leadership system in 1899. We ultimate analysis 6,629,112 speeches from the U.S. House.

We processed the corpus of Congressional speeches using Rapids, an Nvidia supported GPU framework, following the best practices outlined in the literature Grimmer and Stewart, 2013; Hopkins and King, 2010. We followed closely the preprocessing procedures outlined in Kangaslahti et al., 2023; Ebanks et al., 2021, with particular attention to the GPU end-to-end backend for computational speed. Pre-processing is crucial for producing interpretable and valid results (Grimmer and Stewart, 2013; Hopkins and King, 2010) , and we optimized feature selection by stemming and trimming the words in the final corpus used for MLCT esimation.

To arrive at our final set of features, we followed this process:

- Remove any document shorter than 3 tokens.
- Stemmed all words using PorterStemmer to preserve semantic meaning.
- Trimmed the included features by excluding any words appearing in fewer than 2 percent of the document, in such a way that scales with the number of documents in the corpus. We also excluded words that appeared more often than an upper bound of 50 percent of documents.
- Identified bi-grams in the data and tokenized them.

The social science literature has extensively explored the sensitivity of critical substantive findings to pre-processing. Brute force bag-of-words models, when combined with with rigorous empirical validation, generally provide the bulk of explanatory power for text data Hopkins and King, 2010. As long as pre-processing captures all relevant features, our inferences derived from NLP can be used to analyze social phenomena. However, the tuning of pre-processing choices generally depends on the nature of the application. For the following application in this section, our unit of observation is a Congressional speech, a generally dense document.

In Figure 3.4, we show the gamut of MLCT models that we fit to the Congressional speech data. We report the Umass coherence by orthogonality penalty, θ, by learning rate β, and number of topics, k. Upon facial inspection, the the results reported

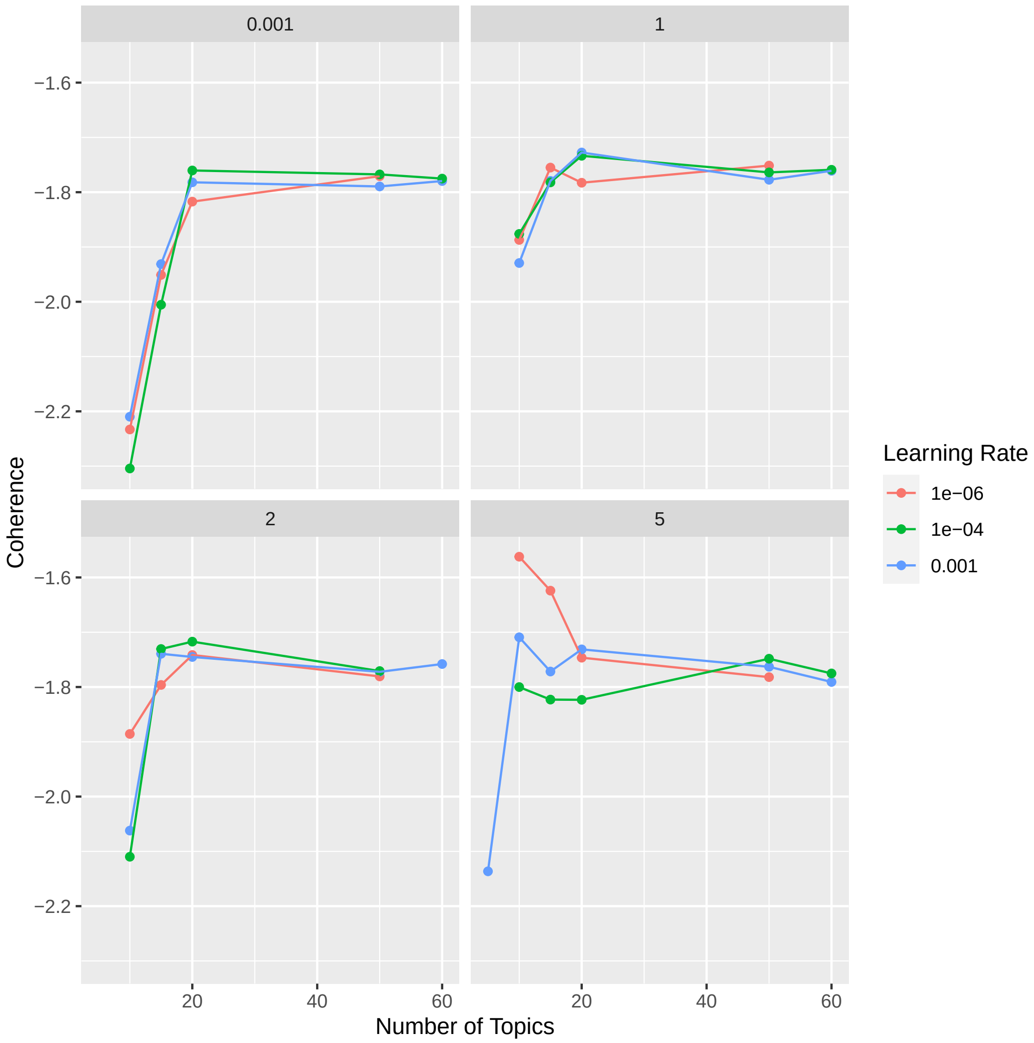
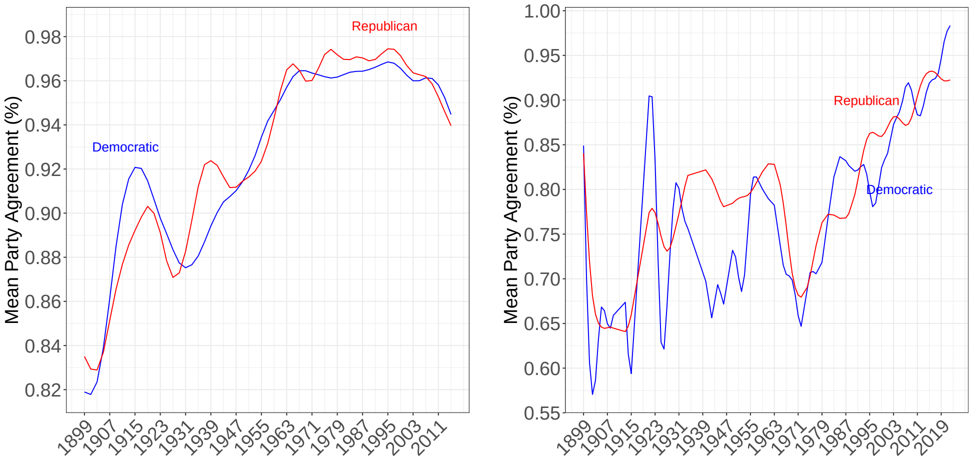


Figure 3.4: Topic Selection Tensor MLCT.

for $\theta = 5$ were generally unstable and facially incoherent. Of the models with lower orthogonality penalties, the model with 20 topics, a learning rate of 0.0001, and an orthogonality penalty of 3 had the maximal coherence. We set the mixing parameters over contexts and topics both to 0.1. Additionally, this model had general facial coherence. Thus, the remainder of the analysis on the application is performed on the outputs of this MLCT model.

Here, we extend the analysis from Figure 2.6 in Chapter 2 from the introduction of the modern party leadership system in 1899 through to 2015. We report the extended level of legislator-leader agreement on historical votes and legislator-leader agreement on speeches given on the House Floor by reporting the correlations of vote and speech agreement. The speech data gives an additional dimension to consider



(a) Long-Run Speech Agreement

(b) Long-Run Vote Agreement

Figure 3.5: Levels of Within-Party Legislator-Leader Agreement on Roll-Call Votes and Within-Party Legislator-Leader Agreement on Floor Speech Topics.

strategic interactions in Congress over roll-call votes alone. Roll-Call votes are tightly controlled by the party leaders, who are unlikely to bring votes where they are likely to lose and are unlikely to bring votes that will divide their party. Thus, we might expect roll-call votes to overstate the levels of agreement between party leaders and legislators in the U.S. House.

Yet, this series shows a strikingly different story and remarkable levels of agreement in public speeches by members of the U.S. House, even accounting for context. The series show some remarkable similarities. First, the Democratic Party and Republican Party caucuses in the U.S. House are remarkably similar over time, a long-run stable stable trend. The parties tend to move in tandem, across vastly different measures of partisan agreement, in votes and in speech. Secondly, the levels of agreement are generally high throughout the historical period. The parties are generally aligned in their agendas as expressed in public speeches, with a low of 80 percent, with post-civil rights highs of 96 percent, which holds steady through the modern period. Since 2010, both parties have exhibited slight declines in their party speech alignment to around 94 percent.

On votes, parties historically agreed with their leaders around 80 percent of the time, with declines during the Teddy Roosevelt Administration and during the passage of the Civil Rights Act. In the first, pro-business and populist Republicans where divided on their votes (related to tariffs, gold, and anti-trust policy), while in the 1960s Southern and Northern Democrats were deeply divided over the passage of

Civil Rights.

Yet along neither measure are the parties particularly divided for long periods of time, and they exhibit long-run levels of relatively high agreement on votes and speech, around 80 percent along both measures.

Analysis

Tensor MLCT provides a tractable means of analyzing larger datasets, allowing for new ways to study social and political behavior. For example, in the U.S. House Speech application, dynamic scaling of the context-topical space yields a novel way to visual the development of legislators' strategic communication positions over time. In Figure A.5.4, legislators' social media positions during the 117th Congress are shown over time, following the same procedure as in Ebanks et al., 2021.

Such an analysis allows researchers to analyze the dynamics of discussion in a tractable way: Although previous text methods enable scaling of text data into partisan spaces, the Tensor MLCT method offers additional flexibility in terms of these types of applications. For example, this kind of scaling traces the stability of discussion between and within political parties. Additionally, this method also allows for novel ways to study the relation between legislative leaders and the rank-and-file members in the strategic communication space by uncovering a structure of topical cohesiveness above merely word counts.

To illustrate this point, consider our application to long-run Congressional agreement and alignment over time. Without topical structure, word frequencies or embeddings alone might not capture the full contours of contemporaneous and historical political debate, as it is likely to fall out the predefined training set. We uncover remarkable long-run stability in leader-legislator agreement over time, even when we might expect that traditional measures, such as roll-call votes would overstate agreement. In fact, we find the opposite.

However, MLCT will identify which topics are about combating climate change with new fuel standard, as opposed to focusing on job losses in the context of climate policy, versus the historical relevance of the gold standard to monetary policy. MLCT will identify that Democratic officials discuss child separations in the context of immigration, while Republicans discuss immigration in the context of border security. In fact, it is the ability to parsimoniously uncover the context of words in relation to this larger topical structure that renders topic models so popular in text analysis.

Taken together, this additional context enables researchers to study the full population of Congressional speeches, at scale, to study Long-Run agenda setting in principled fashion the extends beyond roll-call votes.

3.9 Conclusion

This paper introduces a new spectral Tensor-based implementation method for a standard model of weakly supervised context-topicc analysis, the Multi-Layer Contextual Topic model. The model shows considerable improvements in speed over the standard LDA model and a standard JST model (estimated using variational Bayes and Gibbs Sampling, respectively). At the same time, Tensor MLCT exhibits only minimal loss in coherence relative to JST and retains an edge over LDA on this measure. Additionally, a qualitative assessment of the context-topiccs and tweet labels suggests the model achieves a reasonable level of external validity.

This new estimation methods offers benefits in addition to speed. First, the batched approach allows the model to scale to large data that would be otherwise infeasible to fit a topic model by reducing the memory overhead. Not only that, but by taking advantage of a CPU-based implementation, the batched approach allows the model to be estimated even on a standard workstation without a GPU backend. In addition to strong CPU performance, the model can run fully online on a end-to-end GPU backend, similar to Kangaslahti et al., 2023.In both cases (CPU and GPU), by reducing the memory costs and implementing an estimation routine that converges in a reasonable time, the method will reduce the user-related costs of employing these methods on large data. To this end, the code implementation underlying this paper will be incorporated into the `TensorLy` library of packages for tensor-based methods in `python`. This will enable a larger audience to make ready use of the implementation on a backend optimized for performance with tensor-based calculation.

In extending text applications of this method, we can next study the context-topiccal structure to ask questions related to mass-movements on Twitter – for example, the MeToo movement, Black Lives Matter, and Covid-related protests. These datasets have over 50 million observations each. Tensor-based methods provide a feasible and accessible computational approach to studying these social movements going forward.

Importantly, in this paper, we use the model to uncover otherwise unseen long-run trends in Congressional party leader-legislator agreement. We uncover remarkable long-run stability in leader-legislator agreement over time, even when we might expect that traditional measures, such as roll-call votes would overstate agreement. In fact, we find the opposite. Party leaders and legislators tend to agree at very high levels, and this finding is consistent for the post-war period.

Importantly, this class of latent variable models is useful for applications both within and beyond Natural Language Processing. Within the realm of models of natural language processing, tensor-based estimation routines of models that go beyond the bag-of-words approach and incorporate more semantic structure could be a practical extension, such as to correlated topic models. As of yet, such models are computationally infeasible on large-scale datasets given their current implementation. Incorporating more complex grammars could allow for the estimation of more realistic models of language, improve prediction,and make NLP methods usable in a wider array of research domains.

Additionally, implementing tensor versions of additional latent variable models could allow for applications to new contexts. For example, some researchers have proposed using a Gaussian mixture model (another type of latent variable model) to study elections (King and Gelman, 1991b). Such models would become significantly more tractable with the spectral tensor estimation approach. In fact, Anandkumar et al., 2014 has shown tensor based methods can recover the parameters from various classes of latent variable models, including a Gaussian mixture. Notably, these works mostly focus on theoretical considerations. Thus, implementing these estimation routines in a tractable, accessible fashion could augment the computational tools available to social science researchers as datasets grow increasingly large and complex.

BIBLIOGRAPHY

Abramowitz, Alan I and Steven Webster (2016). "The rise of negative partisanship and the nationalization of US elections in the 21st century". In: *Electoral Studies* 41, pp. 12–22.

Aldrich, John and David Rohde (Jan. 1998). *Measuring Conditional Party Government*. Working Paper. Duke University. URL: https://www.researchgate.net/publication/251842516%5C_Measuring%5C_Conditional%5C_Party%5C_Government.

– (2001). "The logic of conditional party government". In: *Congress Reconsidered*. Ed. by LC Dodd and BI Oppenheimer. Washington, DC: CQ Press, pp. 269–92.

Anandkumar, Animashree et al. (2012). "Two SVDs Suffice: Spectral decompositions for probabilistic topic modeling and latent Dirichlet allocation". In: *CoRR* abs/1204.6703. arXiv: 1204.6703. URL: http://arxiv.org/abs/1204.6703.

Anandkumar, Animashree et al. (2013). *A Spectral Algorithm for Latent Dirichlet Allocation*. arXiv: 1204.6703 [cs.LG].

Anandkumar, Animashree et al. (2014). "Tensor Decompositions for Learning Latent Variable Models". In: *Journal of Machine Learning Research* 15.80, pp. 2773–2832. URL: http://jmlr.org/papers/v15/anandkumar14b.html.

APSA (1950). *Toward a More Responsible Two-Party System. A Report of the Committee on Political Parties of the American Political Science Association*.

Barbera, Pablo et al. (2019). "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data". In: *American Political Science Review* 113.4, pp. 883–901. DOI: 10.1017/S0003055419000352.

Bendix, William (2016). "Bypassing Congressional Committees: Parties, Panel Rosters, and Deliberative Processes". In: *Legislative Studies Quarterly* 41.3, pp. 687–714.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning research* 3.Jan, pp. 993–1022.

Bueno de Mesquita, Ethan and Anthony Fowler (2021). *Thinking Clearly with Data: A Guide to Quantitative Reasoning and Analysis*. Princeton University Press.

Buffon, George Louis Leclerc de (1777). "Essai d'arithmétique morale". In: *Euvres philosophiques*.

Bürkner, Paul-Christian (2018). "Advanced Bayesian Multilevel Modeling with the R Package brms". In: *The R Journal* 10.1, pp. 395–411. DOI: 10.32614/RJ-2018-017. URL: https://doi.org/10.32614/RJ-2018-017.

Campbell, Angus et al. (1980). *The american voter*. University of Chicago Press.

Caughey, Devin and Christopher Warshaw (2022). *Dynamic Democracy: Public Opinion, Elections, and Policymaking in the American States*. University of Chicago Press.

Curry, James M. (2015). *Legislating in the Dark: Information and Power in the House of Representatives*. Chicago: The University of Chicago Press.

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: `10.18653/v1/N19-1423`. URL: `https://aclanthology.org/N19-1423`.

Dewan, Torun and David P. Myatt (2007). "Leading the Party: Coordination, Direction, and Communication". In: *American Political Science Review* 101.4, pp. 827–845. DOI: `10.1017/S0003055407070451`.

Ebanks, Daniel et al. (2021). *Leadership Communication and Power: Measuring Leadership in the U.S. House of Representatives from Social Media Data*. Preprint. Los Angeles: APSA Preprints.

Fenno, R.F. (2003). *Home Style: House Members in Their Districts*. Classics Series. Longman. ISBN: 9780321121837. URL: `https://books.google.com/books?id=p4UrAQAAMAAJ`.

Ferejohn, John A (1977). "On the decline of competition in congressional elections". In: *American Political Science Review* 71.1, pp. 166–176.

Gamm, Gerald and Steven Smith (2020). "The dynamics of party government in Congress". In: *Congress Reconsidered, 11th Edition*. Ed. by LC Dodd and BI Oppenheimer. Washington, DC: CQ Press, pp. 197–224.

Gelman, Andrew and Gary King (Nov. 1990). "Estimating Incumbency Advantage Without Bias". In: *American Journal of Political Science* 34.4, pp. 1142–1164. URL: `tinyurl.com/yymdaj5r`.

– (May 1994). "A Unified Method of Evaluating Electoral Systems and Redistricting Plans". In: *American Journal of Political Science* 38.2, pp. 514–554. URL: `j.mp/unifiedEc`.

Gelman, Andrew, Xiao-Li Meng, and Hal Stern (1996). "Posterior predictive assessment of model fitness via realized discrepancies". In: *Statistica sinica*, pp. 733–760.

Gelman, Andrew et al. (1995). *Bayesian Data Analysis*. Chapman and Hall.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy (2018). *Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts*. `https://data.stanford.edu/congress_text`. Palo Alto, CA.

Girosi, Federico and Gary King (2008). *Demographic Forecasting*. Princeton: Princeton University Press. URL: j.mp/dsmooth.

Granato, Jim, Melody Lo, and MC Sunny Wong (2021). *Empirical Implications of Theoretical Models in Political Science*. Cambridge University Press.

Grimmer, Justin (2010). “A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases”. In: *Political Analysis* 18.1, pp. 1–35. DOI: 10.1093/pan/mpp034.

Grimmer, Justin and Gary King (2011). “General purpose computer-assisted clustering and conceptualization”. In: *Proceedings of the National Academy of Sciences* 108.7, pp. 2643–2650. DOI: 10.1073/pnas.1018067108. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1018067108. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1018067108.

Grimmer, Justin, Dean Knox, and Sean Westwood (2022). “Assessing the Reliability of Probabilistic US Presidential Election Forecasts May Take Decades”. In: *Working Paper*.

Grimmer, Justin and Brandon M. Stewart (2013). “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”. In: *Political Analysis* 21.3, pp. 267–297. DOI: 10.1093/pan/mps028.

Hall, Thad E. and Betsy Sinclair (2018). *A Connected America: Politics in the Era of Social Media*. Oxford University Press.

Harbridge, Laurel (2015). *Is Bipartisanship Dead? Policy Agreement and Agenda-Setting in the House of Representatives*. New York: Cambridge University Press.

Hopkins, Daniel (2018). *The increasingly United States: How and why American political behavior nationalized*. University of Chicago Press.

Hopkins, Daniel and Gary King (2010). “Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability”. In: *Public Opinion Quarterly*, pp. 1–22. URL: http://j.mp/jVFIVg.

Howard, Nicholas O. and Mark E. Owens (2020). “Circumventing Legislative Committees: The US Senate”. In: *Legislative Studies Quarterly* 45.3, pp. 495–526.

Hsu, Daniel and Sham M. Kakade (2012). *Learning mixtures of spherical Gaussians: moment methods and spectral decompositions*. arXiv: 1206.5766 [cs.LG].

Huang, Furong et al. (2015). “Online Tensor Methods for Learning Latent Variable Models”. In: *Journal of Machine Learning Research* 16.86, pp. 2797–2835. URL: http://jmlr.org/papers/v16/huang15a.html.

Imbens, Guido W (2022). “Causality in Econometrics: Choice vs Chance”. In: *Econometrica* 90.6, pp. 2541–2566.

Jacobson, Gary C (2015). “It’s nothing personal: The decline of the incumbency advantage in US House elections”. In: *The Journal of Politics* 77.3, pp. 861–873.

Kang, T. et al. (2018). “Issue consistency? Comparing television advertising, tweets, and e-mail in the 2014 Senate Campaigns”. In: *Political Communication* 35 (1), pp. 32–49.

Kangaslahti, Sara et al. (2023). “TensorLy-LDA: Analyzing Social Media Conversations at Scale with Online Tensor LDA”. In: *Working Paper*.

Katz, Jonathan N, Gary King, and Elizabeth Rosenblatt (2020). “Theoretical foundations and empirical evaluations of partisan fairness in district-based democracies”. In: *American Political Science Review* 114.1, pp. 164–178. URL: `GaryKing.org/symmetry`.

Katz, Jonathan N. and Gary King (Mar. 1999). “A Statistical Model for Multiparty Electoral Data”. In: *American Political Science Review* 93.1, pp. 15–32. URL: `bit.ly/mtypty`.

Kavanagh, Thomas M (1990). “Chance and Probability in the Enlightenment”. In: *French Forum*. Vol. 15. 1, pp. 5–24.

Kim, In Song, John Londregan, and Marc Ratkovic (2018). “Estimating Spatial Preferences from Votes and Text”. In: *Political Analysis* 26.2, pp. 210–229. DOI: `10.1017/pan.2018.7`.

King, Gary and Andrew Gelman (Feb. 1991a). “Systemic Consequences of Incumbency Advantage in the U.S. House”. In: *American Journal of Political Science* 35.1, pp. 110–138. URL: `bit.ly/SystCs`.

– (1991b). “Systemic Consequences of Incumbency Advantage in U.S. House Elections”. In: *American Journal of Political Science* 35.1, pp. 110–138. ISSN: 00925853, 15405907. URL: `http://www.jstor.org/stable/2111440`.

Koop, Gary, M. Hashem Pesaran, and Simon M. Potter (1996). “Impulse response analysis in nonlinear multivariate models”. In: *Journal of Econometrics* 74.1, pp. 119–147.

Lalonde, Robert (1986). “Evaluating the Econometric Evaluations of Training Programs”. In: *American Economic Review* 76, pp. 604–620.

Leamer, Edward E (1983). “Let’s take the con out of econometrics”. In: *American Economic Review* 73.1, pp. 31–43.

Lin, C. et al. (2012). “Weakly Supervised Joint Sentiment-Topic Detection from Text”. In: *IEEE Transactions on Knowledge and Data Engineering* 24.6, pp. 1134–1145.

Lin, Chenghua and Yulan He (2009). “Joint Sentiment/Topic Model for Sentiment Analysis”. In: *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2009)*. DOI: `https://doi.org/10.1145/1645953.1646003`.

Mayhew, David R. (1974). “Congressional elections: The case of the vanishing marginals”. In: *Polity* 6.3, pp. 295–317.

McCarty, Nolan (2019). *Polarization: What everyone needs to know®*. Oxford University Press.

Meilă, Marina (2003). "Comparing Clusterings by the Variation of Information". In: *Learning Theory and Kernel Machines*. Ed. by Bernhard Schölkopf and Manfred K. Warmuth. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 173–187. ISBN: 978-3-540-45167-9.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Quinn, Kevin M. et al. (2010). "How to Analyze Political Attention with Minimal Assumptions and Costs". In: *American Journal of Political Science* 54.1, pp. 209–228. DOI: `https://doi.org/10.1111/j.1540-5907.2009.00427.x`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5907.2009.00427.x`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2009.00427.x`.

Rheault, Ludovic and Christopher Cochrane (2020). "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora". In: *Political Analysis* 28.1, pp. 112–133. DOI: `10.1017/pan.2019.26`.

Riker, William H. (1990). "Heresthetic and rhetoric in the spatial model". In: *Advances in the Spatial Theory of Voting* 46, p. 50.

Roberts, Margaret E. et al. (2014a). "Structural Topic Models for Open-Ended Survey Responses". In: *American Journal of Political Science* 58.4, pp. 1064–1082.

– (2014b). "Structural Topic Models for Open-Ended Survey Responses". In: *American Journal of Political Science* 58.4, pp. 1064–1082.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the Space of Topic Coherence Measures". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. Shanghai, China: Association for Computing Machinery, pp. 399–408. ISBN: 9781450333177. DOI: `10.1145/2684822.2685324`. URL: `https://doi.org/10.1145/2684822.2685324`.

Rodriguez, Pedro L. and Arthur Spirling (2022). "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research". In: *The Journal of Politics* 84, pp. 101–115.

Rudkowsky, Elena et al. (2018). "More than Bags of Words: Sentiment Analysis with Word Embeddings". In: *Communication Methods and Measures* 12.2-3, pp. 140–157. DOI: `10.1080/19312458.2018.1455817`. eprint: `https://doi.org/10.1080/19312458.2018.1455817`. URL: `https://doi.org/10.1080/19312458.2018.1455817`.

Salganik, Matthew J. (2017). *Bit by Bit: Social Research in the Digital Age*. Open Review Edition. Princeton, NJ: Princeton University Press.

Shepsle, Kenneth A (2003). “Losers in politics (and how they sometimes become winners): William Riker’s heresthetic”. In: *Perspectives on Politics* 1.2, pp. 307–315.

Steinert-Threlkeld, Zachary C. (2018). *Twitter as Data*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press. DOI: 10.1017/9781108529327.

Tiefer, Charles (2016). *The Polarized Congress: The Post-Traditional Procedure of Its Current Struggles*. Lanham, MD: University Press of America.

Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Wallis, Kenneth F. (1987). “Time Series Analysis of Bounded Economic Variables”. In: *Journal of Time Series Analysis* 8.1, pp. 115–123.

Wallner, James I. (2013). *The Death of Deliberation: Partisanship and Polarization in the United States Senate*. New York: Lexington Books.

White, Halbert (1996). *Estimation, Inference, and Specification Analysis*. New York: Cambridge University Press.

Yan, Hao et al. (2019). “The Congressional Classification Challenge: Domain Specificity and Partisan Intensity”. In: *Proceedings of the 2019 ACM Conference on Economics and Computation*. EC ’19. Phoenix, AZ, USA: ACM, pp. 71–89. ISBN: 978-1-4503-6792-9.

Appendix A

SUPPLEMENTARY INFORMATION FOR CHAPTER 1

A.1 Introduction

In the following pages we provide technical details about the important steps in our paper's methodology: summary statistics and visualizations of our Twitter data; technical details and sensitivity analyses for our topic modeling; information useful for understanding the sensitivity of our PCA modeling decisions; summary statistics from our network modeling; and finally, details and sensitivity analysis of our dynamic analysis.

Upon publication, we will make our code and documentation available, along with a great deal of additional material that readers can use to examine our modeling decisions and the robustness of our results to those decisions, including detailed log files and estimation details. We will also provide our raw data, subject to Twitter's current policies about data sharing.

A.2 The Distribution of Tweeting Behavior

In this section we provide summary statistics on the Twitter activity of the Members of the U.S. House of Representatives, during the time period covered in our study. Table A.2.1 gives summary statistics for the entire dataset, by party. In Figure A.2.1 we show the data on tweets by member in a histogram.

Table A.2.1: Distribution of Tweeting Behavior: Entire Dataset

Party	Mean	Median	Minimum	Maximum	Standard Deviation
Democratic Party	894.45	797	43	3,200	520.46
Republican Party	528.31	457	11	2,732	417.05
All	727.17	597	11	3,200	509.33

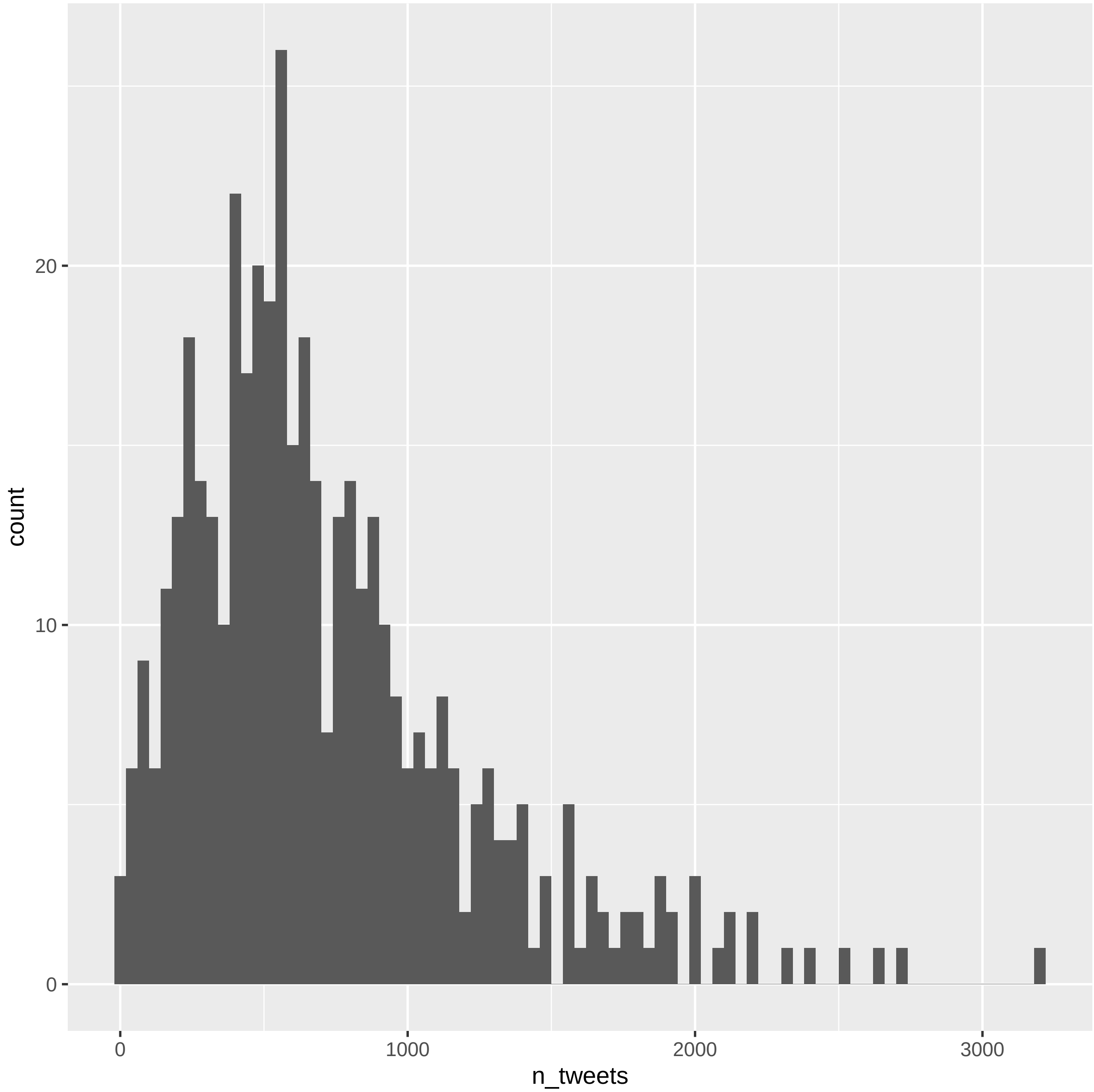


Figure A.2.1: Distribution of Tweets by Member

A.3 Theory

Game Setting and Example

Here we summarize the model setting. In the next section of the Supplementary Information we provide intuition for how the model fits our setting using the 2019 government shutdown as an example. In this model, there are n party rank-and-file members who are deciding to advocate either policy stance A or B. The optimal policy choice depends on a state variable, θ. The state is the underlying political situation. It represents the party mood regarding an unexpected politically sensitive issue. Finally, members receive private signals m_i about the true state of the world, which are normally distributed.

In order to coordinate on a policy, a policy must have a sufficient threshold of support, p_A and p_B for policies A and B respectively. Conceptually, this is the informal level of consensus needed for the party to advocate a platform. Then, x is the number of party rank-and-file advocating policy A. Party members earn the following payoffs depending on their choice of policy stance and on the underlying state θ and support for policy A, x :

$$\begin{cases} u_A(\theta) = \exp\{\frac{\lambda\theta}{2}\} & \text{if } \frac{x}{n} > p_A\text{, adopt policy } A \\ u_B(\theta) = \exp\{-\frac{\lambda\theta}{2}\} & \text{if } p_B > \frac{x}{n}\text{, adopt policy } B \\ u_A = u_B = 0 & \text{if } p_A \geq \frac{x}{n} \geq p_B\text{, coordination failure} \end{cases} \tag{A.1}$$

Dewan and Myatt (2007) assume legislators play a threshold strategy and that they vote for policy stance A instead of the status quo, B, if and only if their private signal $m_i > m$ for some threshold m. They assume this private signal is distributed normally with mean θ and variance $\frac{1}{\psi}$. In the payoff structure, the sensitivity to the benefits of coordinating (electoral success, the continuation of good public policy) are captured by λ, the party's *need for direction.* This concept represents the importance of choosing the right messaging strategy and the gravity of choosing incorrectly. Conditional on state of the world θ, party rank-and-file advocate for A with probability $p = \Pr[m_i > m|\theta]$, which is distributed normally with standard normal CDF Φ by the distributional assumption on the signal m_i. The authors note that as n increases, $\frac{x}{n}$ approaches p by the Law of Large Numbers. The authors then note that assuming large n, policy A succeeds if $p > p_a$. Given the normality assumption on m_i, this condition is equivalent to $\theta > \theta_A$ where θ_A satisfies $p = \Phi[\sqrt{\psi}(\theta_A - m)]$. Similarly, the party adopts policy B if $\theta_B > \theta$ where θ_B satisfies $p = \Phi[\sqrt{\psi}(\theta_B - m)]$ This results in the following outcome structure:

$$\text{Outcome} = \begin{cases} \text{Coordinate on } A & \text{if } \theta > \theta_A \\ \text{Coordinate on } B & \text{if } \theta_B > \theta \\ \text{Coordination failure} & \text{if } \theta_A \geq \theta \geq \theta_B \end{cases} \tag{A.2}$$

$$\text{where} \quad \begin{cases} \theta_A = m + \frac{\pi_A}{\sqrt{\psi}} \\ \theta_B = m + \frac{\pi_B}{\sqrt{\psi}} \end{cases} \tag{A.3}$$

where substitutions $\pi_A = \Phi^{-1}(p_A)$ and $pi_B = \Phi^{-1}(1 - p_B)$ have been made for clarity. The authors note that conceptually, π_A and pi_B measure the heights of the *barriers to coordination*.

Given this setting, the game sequence proceeds as follows:

1. Rank-and-file members receive a private signal $m_i|\theta$ for i in $1, ..., n$ that is conditioned on the true state of the world distributed with variance $\frac{1}{\psi}$, the *sense of direction*.

2. Leaders of the party decide to give a speech or not relaying their signal to the party.

3. Rank-and-file members adopt a policy stance they individually decide to advocate.

4. If the critical thresholds of rank-and-file members advocate for the same policy stance (π_A and π_B), the party successfully coordinates. These thresholds are called *barriers to coordination*. Otherwise, the party fails to coordinate.

5. Borrowing terminology from Dewan and Myatt (2007), rank-and-file members are willing to follow their leaders' signals based on a leadership index R:

$$R = \frac{\text{Barriers to Coordination} \times \text{Sense of Direction}}{\text{Need for Direction}} \tag{A.4}$$

6. The equilibrium strategies are characterized by R, which makes the concept of leadership precise in our context: When $R > 1$, rank-and-file members adopt the same signal as their leaders. For $R < 1$, rank-and-file members adopt a threshold that is biased towards the leaders' preferred threshold, increasing in R. That is, as R approaches 1, rank-and-file member play strategies biased in favor of their leaders' preferred strategies.

In our case, we interpret the private signals m_i as a member's observation of the party's mood, which is derived from interpersonal conversation, social media stances from other party members, and party conference meetings and calls.[1] We interpret the leader's speech as the leadership of the parties tweeting out their talking points and messaging strategy to their members. We interpret the policy stances as the policy stances advocated on Twitter. In order to identify Dewan and Myatt (2007) we restrict the strategy space to what they consider a natural class of strategies, threshold strategies.

We interpret the policy stances on Twitter themselves as the the key strategic behavior. On Twitter, House party leadership and rank-and-file membership publicly and strategically communicate their policy stances. When R is high, we expect rank-and-file members to follow their leaders. When it is low, we expect rank-and-file members to be less likely to follow their leaders. Thus, the leadership index R suggests intuition for patterns of communication behavior we might expect. Using this intuition from this framework, we derive hypotheses regarding House party leadership behavior and the tendency of rank-and-file House members to follow their leaders.

A.4 Topic Analysis

In this section we discuss the details of the Joint Sentiment Topic model and our implementation. In the next section we provide technical details for the Joint Sentiment Topic model. The subsequent sections provides graphical material on the sensitivity of our results to modeling decisions.

Joint Sentiment Topic

We implemented a Joint Sentiment Topic (JST) model (Lin and He 2009) to obtain the topic diversity for members of the U.S. House of Representatives. Lin et. al (2012) describe their method as follows. Take a corpus of tweets C, which is a collection of D tweets $\{t_1, t_2, t_3, ..., t_D\}$. Each tweet itself is a collection of N_t words. Let the words in each tweet be denoted by $\{w_1, w_2, ..., w_{N_t}\}$. Now, each potential word in any tweet is indexed by a vocabulary, with V total terms $\{1, 2, 3, 4, ..., V\}$. Now, let J signify the total number of sentiment labels and L the total number

[1] In order to link this theory to our empirical setting, we first note that House member Twitter accounts are managed both by staff and the legislator. We assume that the incentives of the congressional communication staff are aligned with the legislator they represent. Conversations with several House communication staffers suggest social media activity is coordinated at the office level under the direction of their principal.

of topics. Explicitly, the underlying data-generation process for the documents is summarized as follows:

1. For each sentiment label j in $\{1, 2, 3, ..., J\}$
   a) For each topic k in $\{1, 2, 3, ..., L\}$ draw $\phi_{j,k}$ from $Dir(\lambda_j \times \beta^L_{j,k})$
2. For each tweet t, choose a distribution $\pi_t \sim Dir(\gamma)$
3. For each sentiment label j under tweet t, drawn a distribution $\theta_{j,k}$ $Dir(\alpha)$
4. For each word w_i in tweet t,
   a) Draw sentiment j_i from Multinomial(π_t)
   b) Draw topic label k_i from Multinomial(θ_{t,j_i}) which is conditioned on sampled sentiment j_i.
   c) Draw word from per-corpus word distribution conditioned on sentiment label j_i and topic label k_i, i.e. choose a word from Multinomial(ϕ_{j_i,k_i}).

The hyperparameter α can be interpreted intuitively as the the prior observation counts for the number of times topic k associated with sentiment label j is sampled from a tweet. The hyperparameter β can be interpreted as the prior belief on the frequency at which words sampled from topic k are associated with sentiment label j, respectively, *ex ante*. Following this logic, λ can be treated as the prior belief on the number of times sentiment label j is sampled from a tweet before observing any tweets.

Observe that as β goes to 0, the model converges to a model of a single sentiment-topic. That is, one sentiment-topic label has probability 1, with all other labels being assigned 0. On the other hand, as β grows large, the limiting distribution is uniform over sentiment-topics. We expect that tweets, given their concise nature, are likely only to relate to very few topics at once, so we set these priors relatively small, following standard practice (such as in Lin and He, 2009).

2

[2] The model incorporates a prior over λ using a lexicon which suggests sentiment orientations for some 7000 common words. For more details, see Lin and He (2009) and Lin et al. (2012). We use an R wrapper written around the authors' original C++ code, found here: `https://github.com/linron84/JST` to estimate the model. We run the model for 1000 iterations after a burn-in of 1000. The model is computationally expensive, and it runs for about 9 hours prior before converging.

Topic Selection

Table A.4.2: Emblematic Tweets

Handle	Tweet	TopicTitle
@repdelbene	whats stake trillion dollar federal funding countless policy business decision made based census data fill census sure youre counted	Census Encouragement - Positive
@reprokhanna	climate change isnt intergovernmental panel climate change report effort front imperative minimize impact climate change human life impact human life climate	Climate Change-Positive
@stevescalise	family small business dems held relief hostage day play politics try sneak liberal wish list emergency finally agree largely deal made schumer block worth	GOP attack Democrats as Socialists-Negative
@repcuellar	homeland security questioned border patrol chief carla provost current crisis cbp facing southern discussed border patrol retention well border patrol processing	Humanitarian Aid at Border-Negative
@reparmstrongnd	finally chance ask ig horowitz question fbi investigation trump campaign fbi knew steele unreliable fbi omitted info obtain warrant comey mccabe perpetrated fraud fisa court investigating trump campaign	Trump/Russia Investigation -Negative

We select the number of topics based on the inflection point beyond which increases to coherence are small. Based on this criterion, we select 60 topics. To arrive at this number, we tuned the model starting from 5 topics and 10 topics increasing in increments of 10 up to 60 topics. Figure A.4.2 shows that that topical coherence along an NPMI metric is maximized at 60 topics (which results in 180 Senti-Topics). Due to computational feasibility constraints, we can estimate at most 60 topics, but

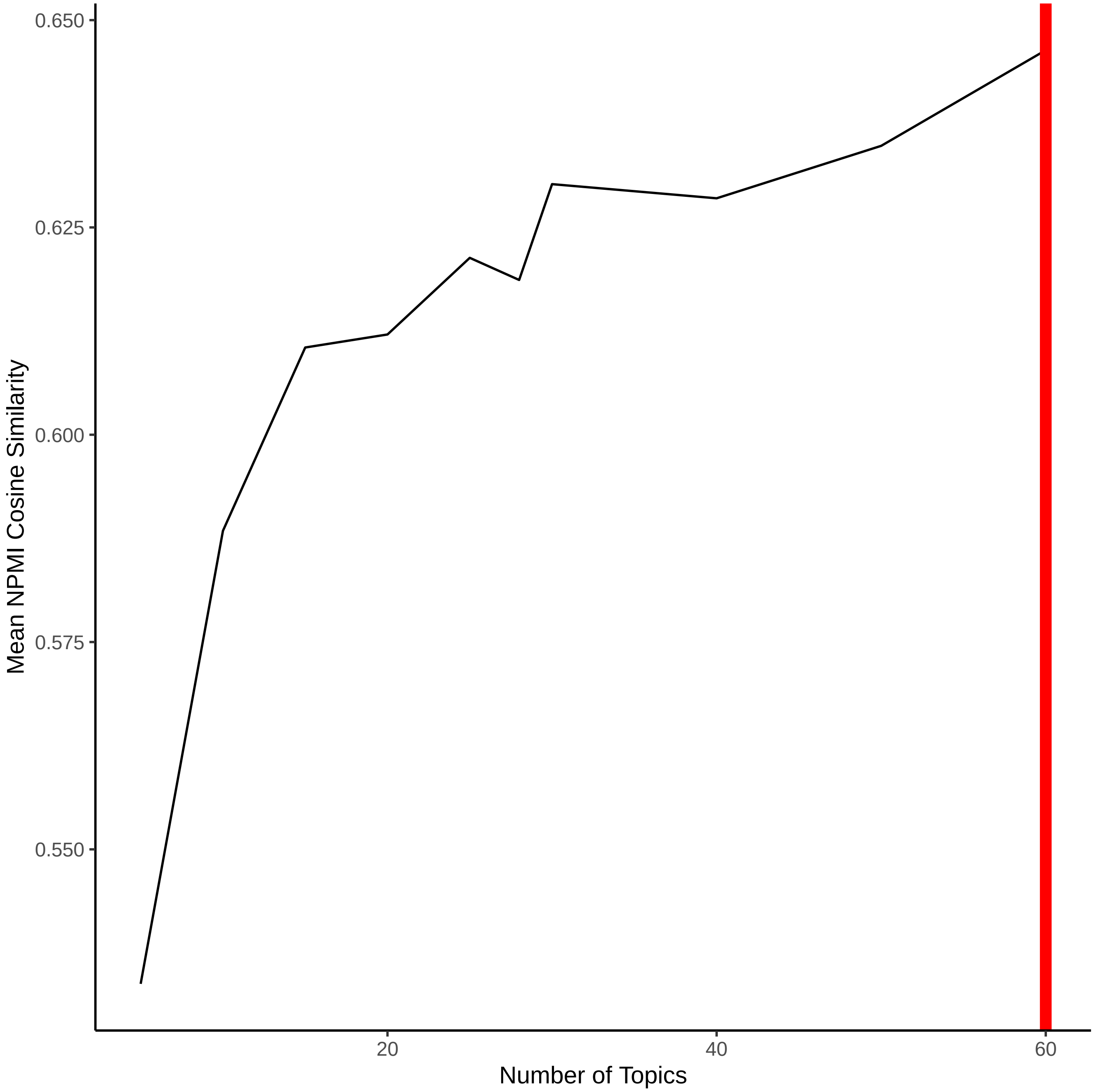


Figure A.4.2: Coherence Score by Number of Topics.

in addition to strong quantitative coherence, we show they have facial validity, as well.

A.5 PCA Analysis and Summary

First, we show the topics which we identify as member-led in tables A.5.3 and A.5.4. These tables report the percent contribution to the overall variation in the data when we decompose the topic data using Principle Components Analysis.

Table A.5.3: PCA Topic Contributions - Member Driven 115th

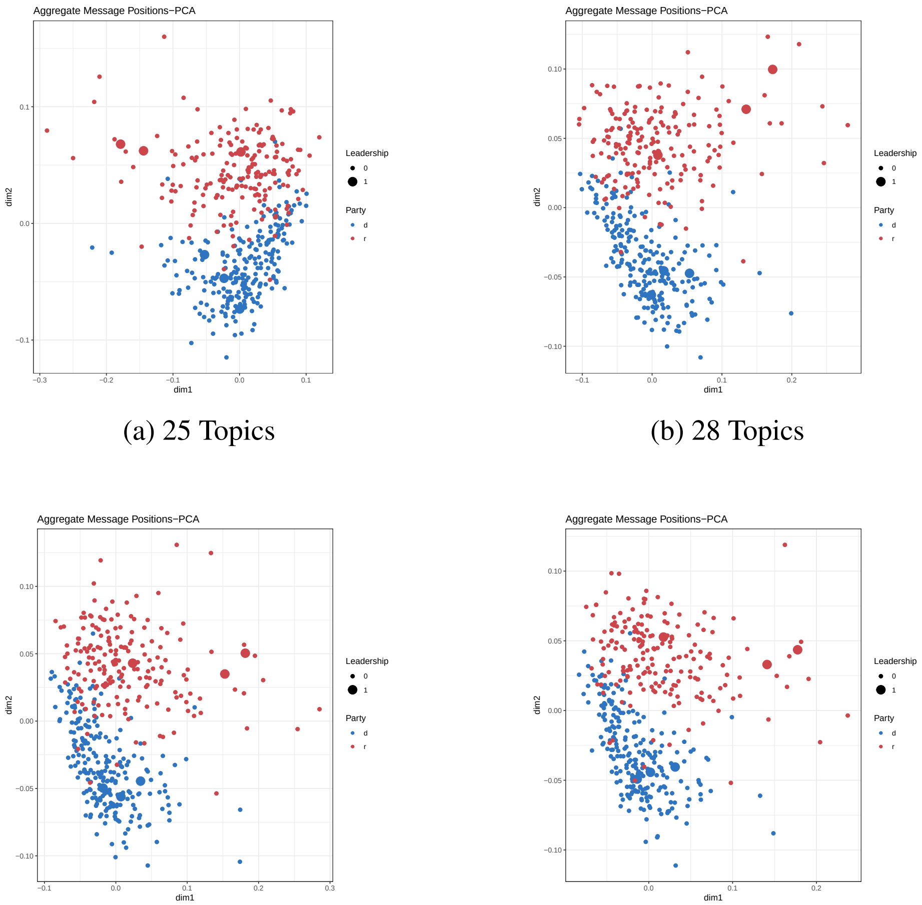
Topic	Contribution
Lowest Unemployment Rate - Positive	1.26
Guests at Capitol Hill-Neutral	1.24
LGBT Equality-Negative	1.23
Fight for Civil Rights-Negative	1.15
Retweeting a Controversial Statement-Negative	1.11
Climate Change-Positive	1.11
Partisan Attacks on Trump/Biden-Negative	1.01
Important Meetings-Negative	0.99
Trump Admin Undermines Country - Negative	0.97
Budgetary Legislation -Negative	0.86
Committee Hearings-Positive	0.86
Hurricane Relief-Negative	0.86
Trump Climate Policy-Negative	0.85
Health Care Expansion - Neutral	0.84
Foreign Election Interference-Negative	0.82
Women's Pay - Positive	0.82
Supreme Court Nominations-Negative	0.73
Thoughts and Prayers - Negative	0.71
Floor Speeches-Negative	0.68
Student Loan Relief-Positive	0.67

Sensitivity to Topic Number - Democratic Party

The following graphs show that the policy positioning which form the basis of the Need for Direction classification scheme are robust to changes in number of topics. The relative positioning and separation in the topical space is invariant to choice of topic number.

Table A.5.4: PCA Topic Contributions - 116th Member Driven

Topic	Contribution
Family Seperations-Negative	1.18
Pro-Life Policy - Negative	1.14
China/Hong Kong Protests-Negative	1.11
Republican Senate Legislation-Negative	1.10
Prevent Gun Violence-Negative	1.08
Trump Admin Undermines Country - Negative	0.96
Fight for Civil Rights-Negative	0.93
Meuller Investigation - Negative	0.88
Trump Asuylum Policy	0.86
Enjoyable Visit - Positive	0.74
LGBT Equality-Negative	0.71
Social Security/Postal Service - Neutral	0.70
Health Care Expansion - Neutral	0.70
Trump Climate Policy-Negative	0.68
Mitch Mcconnel's Senate-Negative	0.67
Partisan Votes - Negative	0.60
Voting Rights - Positive	0.56
Law Enforcement - Positive	0.55
Honoring Cultural History-Negative	0.55
Protect Health Insurance -Neutral	0.54



(a) 25 Topics

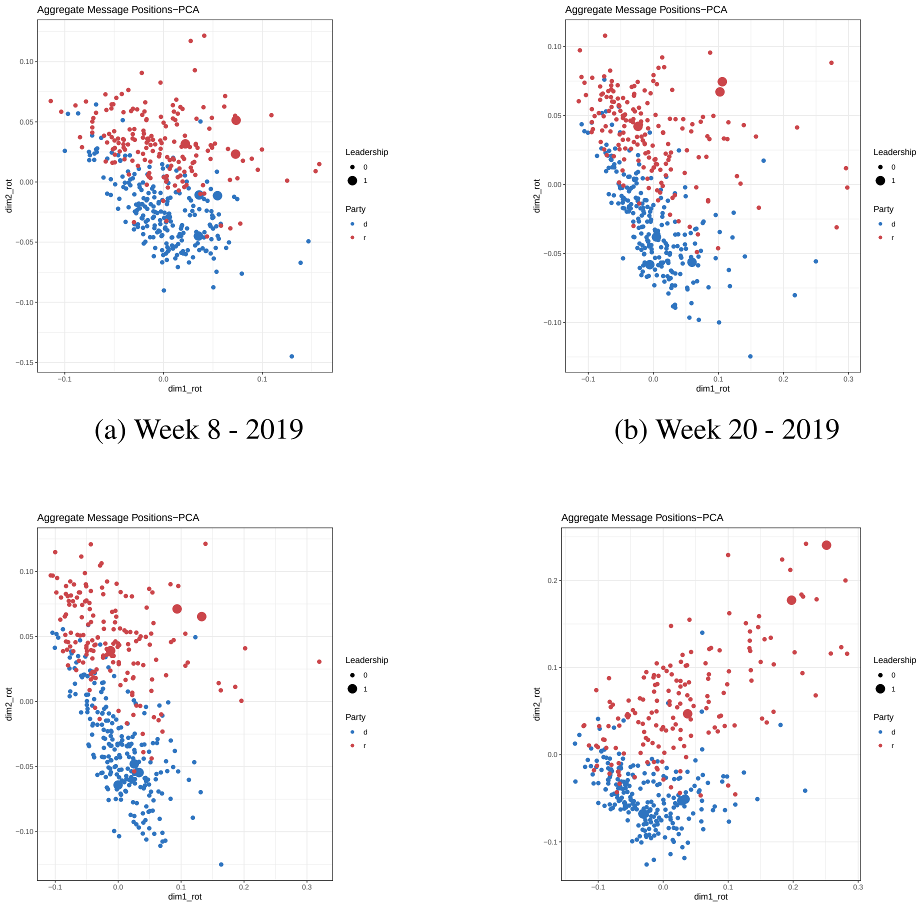
(b) 28 Topics

(c) 30 Topics

(d) 40 Topics

Figure A.5.3: PCA Embeddings for Policy Stances, Varying by Topic Number.

Dynamic Policy Stance Analysis



(a) Week 8 - 2019

(b) Week 20 - 2019

(c) Week 32 - 2019

(d) Week 2-2020

Figure A.5.4: Changes in Time of Policy Stances.

A.6 Time Series and Vector Autoregression

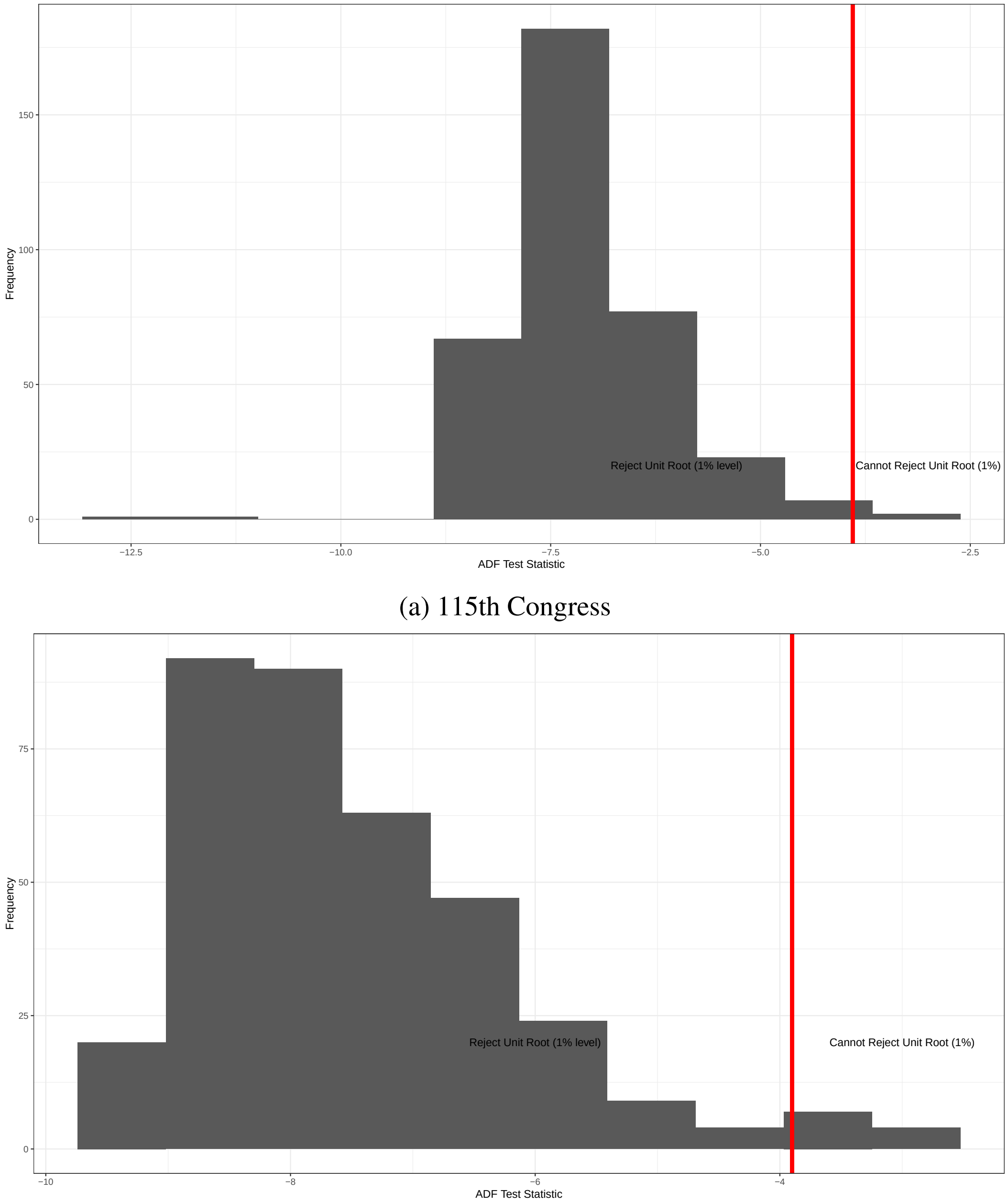
This section provides details for our dynamic analysis, in particular our vector autoregression methodology.

First, we sample some key time series to show the stationarity assumption – which is key to the validity of the VAR – holds across a variety of topics. We also show the full histogram of Augmented Dickey Fuller statistics, which tests for non-stationarity. The vast majority of our time series are consistent with the stationarity assumption, rejecting the unit root at the 1% level for over 95% of topics for the Democratic and Republican Parties in both the 115th and 116th Congresses.

Finally, we show a robustness check and that institutional leadership influence is substantively large. In Table A.6.5 shows that institutional leaders exert on average more influence than the most followed accounts in each party and the leadership of the other party. On average, leaders exert double the influence as leaders from the other party on their members, as well nearly double the influence as the most followed accounts from within the same party. This latter finding highlights the relative strength of institutional leadership within the party caucus relative to the influence of members of the party who are popular with the public social media.

Table A.6.5: IRFs Robustness

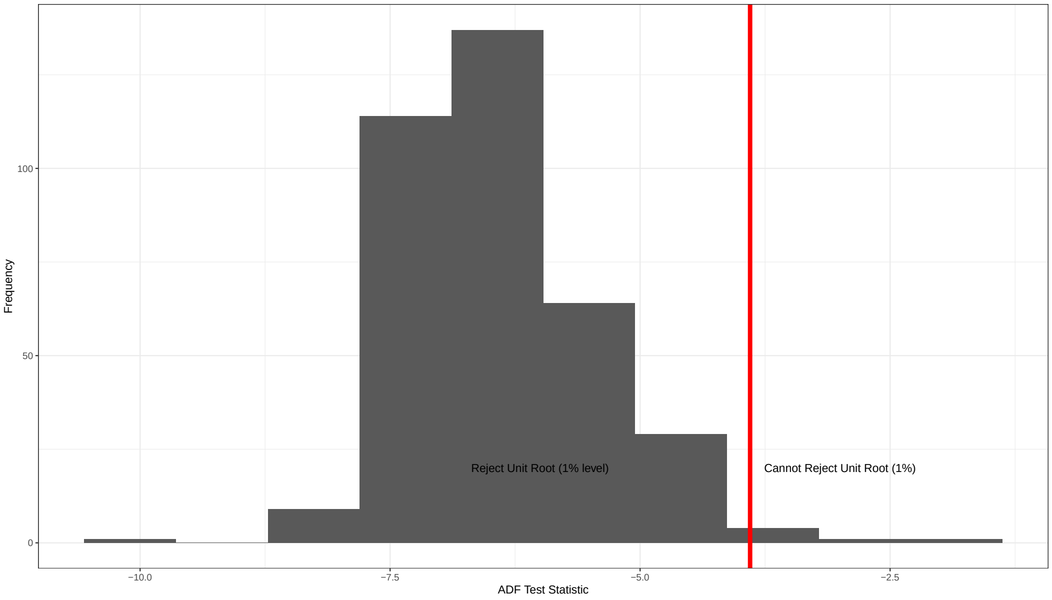
Leaders	Most Followed	Cross-Party Leaders	Total Percent Contribution	Congress	Party
0.202	0.160	0.042	57.245	115	Democra
0.297	0.063	0.134	57.245	115	Republic
0.184	0.050	0.135	64.409	116	Democra
0.408	0.364	0.257	64.409	116	Republic



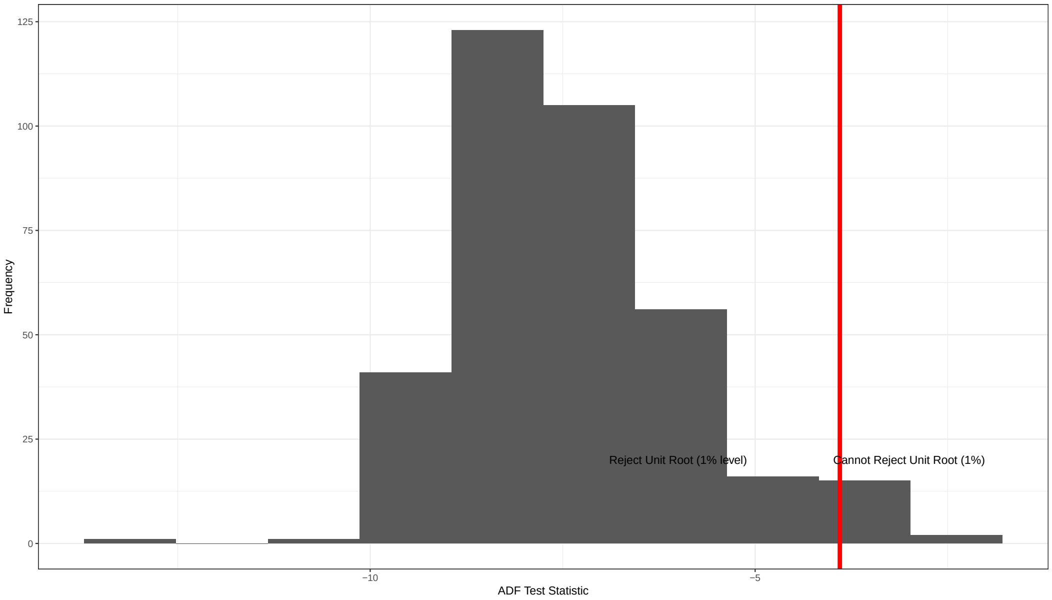
(a) 115th Congress

(b) 116th Congress

Figure A.6.5: ADF Unit Root Test Statistics for all Topics: **Republican Party**. This Figure shows the distribution of tests ADF statistics for unit roots. All statistics to the left of the line represent topics for which we reject the null of a unit root at the 1% level, implying the stationarity assumption is satisfied.

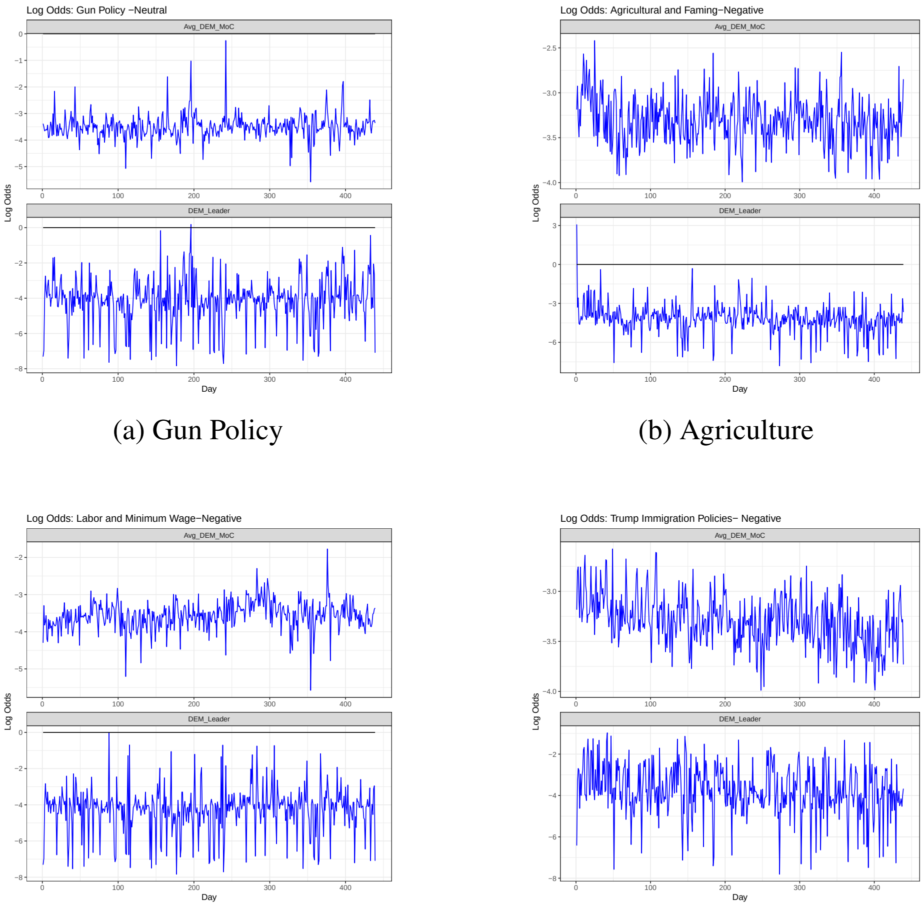


(a) 115th Congress



(b) 116th Congress.

Figure A.6.6: ADF Unit Root Test Statistics for all Topics: **Democratic Party**. This Figure shows the distribution of tests ADF statistics for unit roots. All statistics to the left of the line represent topics for which we reject the null of a unit root at the 1% level, implying the stationarity assumption is satisfied.



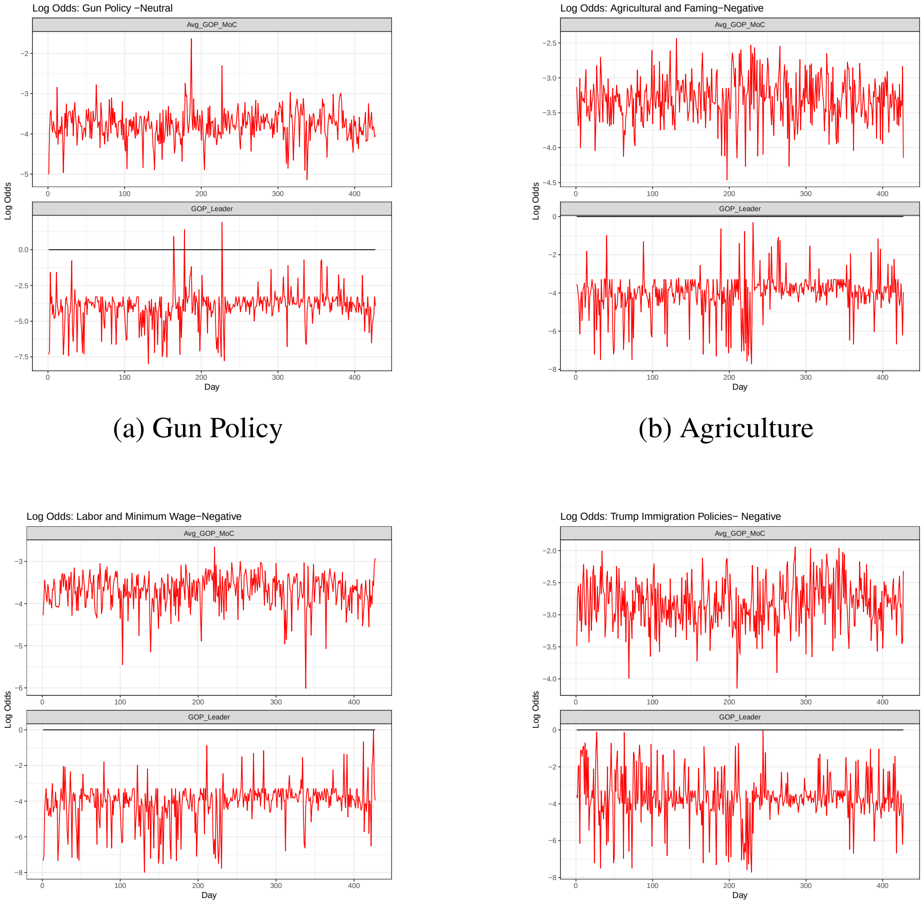
(a) Gun Policy

(b) Agriculture

(c) Minimum Wage

(d) Immigration Policies

Figure A.6.7: Stationarity in Log Odds of Daily Propensity of Discussion- Democratic Party.



(a) Gun Policy

(b) Agriculture

(c) Minimum Wage

(d) Immigration Policies

Figure A.6.8: Stationarity in Log Odds of Daily Propensity of Discussion- Republican Party.

Appendix B

MATHEMATICAL APPENDIX FOR CHAPTER 2

B.1 Statistical Details

This appendix provides the full likelihood function for our model, including all the features described in Section 2.2, as well as situations where the effective vote is both included in the model as a lagged covariate and unobserved (because previous election was uncontested).

To write the full likelihood function, define an uncontestedness indicator U_{it} as 1 if the Democrat runs uncontested, 0 if contested, and -1 if the Republican runs uncontested in district i and time t. Then partition elections into four sets depending on whether the current election i, t and its lag $i, t-1$ are contested or uncontested. Denote CC as the set of all elections for which $U_{it} = 0$ and $U_{i,t-1} = 0$; UC as the set of elections for which $U_{it} \neq 0$ and $U_{i,t-1} = 0$; CU as the set of elections where $U_{i,t} = 0$ and $U_{i,t-1} \neq 0$; and UU as the set of elections for which $U_{it} \neq 0$ and $U_{i,t-1} \neq 0$. Then the likelihood function factors into four parts corresponding to these sets:

$$L = \left(\prod_{i,t \in \{\text{CC}\}} L_{it}^{\text{CC}} \right) \left(\prod_{i,t \in \{\text{UC}\}} L_{it}^{\text{UC}} \right) \left(\prod_{i,t \in \{\text{CU}\}} L_{it}^{\text{CU}} \right) \left(\prod_{i,t \in \{\text{UU}\}} L_{it}^{\text{UU}} \right) \tag{B.1}$$

each of which we now define.

The first component of the likelihood, for when election i, t and $i, t-1$ are both contested, is by far the most prevalent for the US congress. The likelihood for observation i, t is then simply

$$L_{it}^{\text{CC}} = \text{ALT}(v_{it} \mid \mu_{it}, \phi_t^2, \nu_t). \tag{B.2}$$

The second component of the likelihood accounts for which party is running uncontested at time t:

$$L_{it}^{\text{UC}} = \mathit{1}(U_{it} = 1)\psi_{it} + \mathit{1}(U_{it} = -1)(1 - \psi_{it}), \tag{B.3}$$

where our censoring assumption from Section 2.2 implies that $\psi_{it} \equiv \int_0^{0.5} \text{ALT}(v^* \mid \mu_{it}, \phi_t^2, \nu_t) dv^*$, given the indicator function defined as $\mathit{1}(a) = 1$ if a is true and 0 otherwise, for any statement a.

To write the third component, where the lagged value of the effective vote is unobserved (because it is uncontested), we require a prior distribution for how this variable is distributed. The posterior will be computed from the entire model, but to begin we need an assumption about this prior. One option is to let $v^*_{i,t-1}$ be a censored ALT when unobserved (and equal to v_{it} when observed) but this creates a substantial computational burden with little substantive benefit. Instead, we find we can represent almost all relevant information by assuming that, when unobserved, $v^*_{i,t-1} \sim \mathcal{N}(Z_{i,t-1}\alpha_t, \sigma^2_v)$, with $Z_{i,t-1}$ a vector of covariates such as lagged presidential vote in a congressional district and incumbency status. Then this component of the likelihood is

$$L^{\mathrm{CU}}_{it} = \int_{-\infty}^{\infty} \mathrm{ALT}(v_{it} \mid \mu_{i,t}, \phi^2_t, \nu_t) \cdot \mathcal{N}(v^* \mid Z_{i,t-1}\alpha_t, \sigma^2_v) dv^*, \tag{B.4}$$

where the unobserved lagged effective vote v^* is included in X and so contributes to μ_{it}.

For the final component of the likelihood, we use features of all three previous components, so that

$$L^{\mathrm{UU}}_{it} = \mathit{1}(U_{it} = 1)\psi'_{it} + \mathit{1}(U_{it} = -1)(1 - \psi'_{it}), \tag{B.5}$$

where

$$\psi' = \int_{-\infty}^{\infty} \int_{0}^{0.5} \mathrm{ALT}(v \mid \mu_{i,t}, \phi^2_t, \nu_t) dv \cdot \mathcal{N}(v^* \mid Z_{i,t-1}\alpha_t, \sigma^2_v) dv^*.$$

Appendix C

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

C.1 Comparison of Nominal Confidence Interval Lengths

To quantify the magnitude of uncertainty differences between the Normal and LogisTiCC models for district- and legislature-level statistics, we compute the ratios of the credible interval (CI) widths from these two models. To compute the ratio of CI widths for district-level results, we take each of the elections for which we make a prediction and compute the width of the 95% credibility interval for both the Normal and LogisTiCC models. We then calculate the ratio of the widths of the LogisTiCC CI's to the Normal. To compute the ratio of the credibility intervals for the legislative median, we compute a 95% credibility interval for the median seat in the House for each year under each model, again out-of-sample. We take the ratio for each of the 27 years for which we make a prediction, and report the density of these ratios.

Figure C.1.1 reports distributions of these ratios, with summaries in Table C.1.1. The table shows that, at the individual level, the LogisTiCC forecast credible intervals are only 42 percent larger than those of Gelman-King model on average, with a mode at 25 percent, which we can see from the figure. At the same time, because of the correlations between different districts represented in the LogisTiCC, its CIs for the legislative median are 500 percent larger, on average. Given the results in Figures 1–3, it is clear that these larger CIs are needed for accurate calibration due to dependence across districts.

	Mean	Standard Deviation
District Level Results	1.42	0.246
Legislature Level Results	5.06	1.19

Table C.1.1: Numerical Summaries of Figure C.1.1

C.2 Ablation Studies

We make four modeling innovations to achieve generatively accurate model predictions: a national trend, coefficient stability, local uniqueness, and electoral surprises. In this section, we conduct “ablation studies,” where each model component is se-

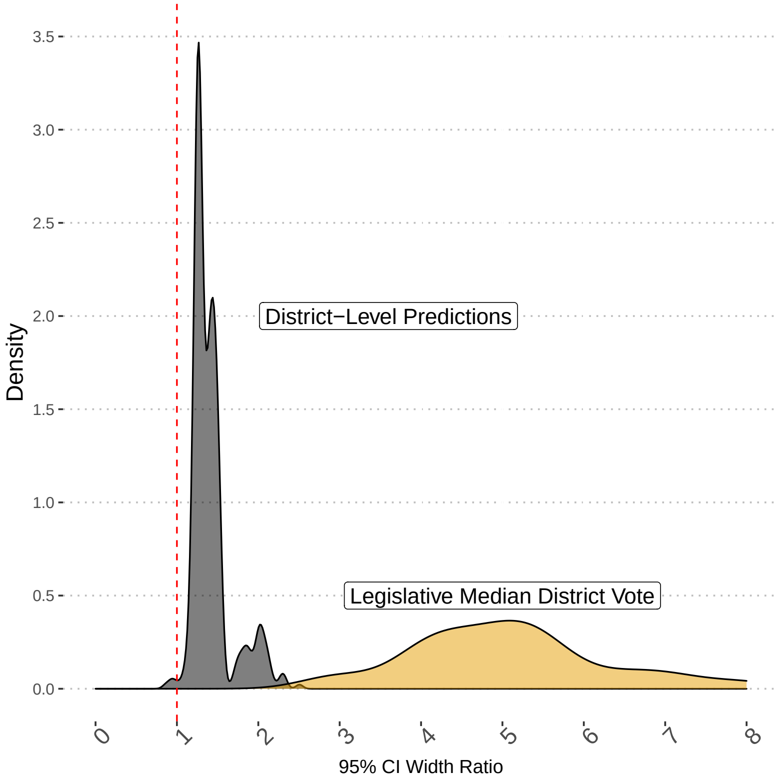


Figure C.1.1: LogisTiCC-to-Normal Ratios of 95% Credibility Interval Widths.

quentially removed to show how the model degrades. The conclusion of this section is that all model components are essential to achieve the performance we report.

The linear-normal model treats the data as having 435 independent district-level observations for each election year. In reality, congressional elections data have high levels and sophisticated patterns of dependence among voting outcomes across districts. In Figure C.2.2, we replicate the calibration exercise from Figure 2.3, which reports the model predictions and observed values for the median congressional seat in the given election year. We report results for three ablated models. We give the normal model with none of the modeling innovations (in gold); a model with neither a National trend assumption nor coefficient stability, but with an additive logistic student-T (ALT) assumption on the error term (in yellow); and a model with normal errors, but with a national trend and coefficient stability (in green).

We would expect a well-calibrated model to contain the true value of the median seat's vote share about ∼ 95 percent of the time. To that end, we see that the normal with none of our innovations fares poorly, correctly containing the true value for the median seat only 25 percent of the elections. If we switch to the ALT specification,

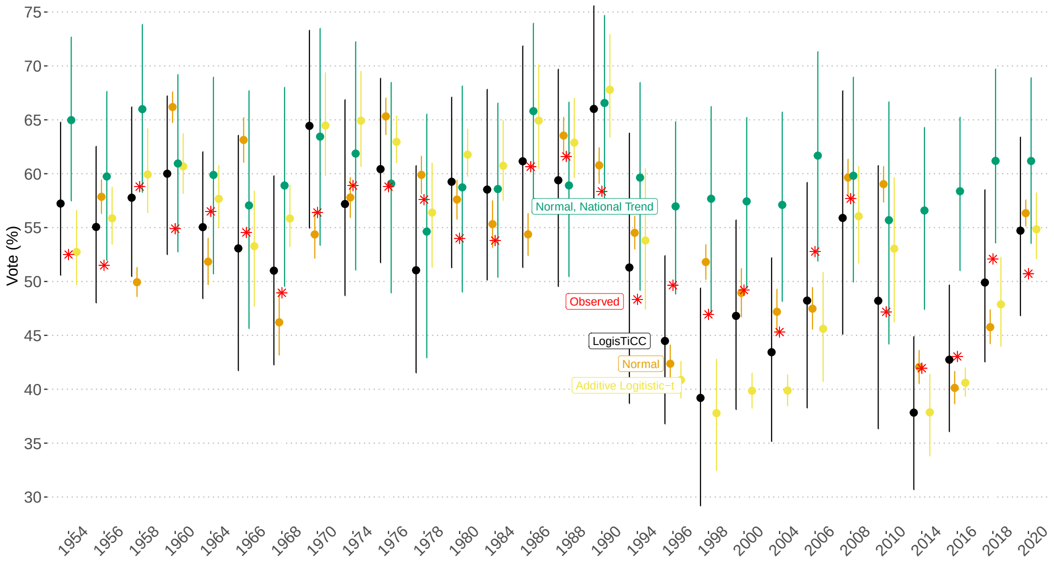


Figure C.2.2: Comparison of Model Calibration as under Ablation.

we achieve a 40 percent accuracy rate, which is still inadequate, but better than Normal alone. When we assume normal errors with a national trend and coefficient stability, we achieve 64 percent accuracy. Under the ablated models, we find that the coefficient stability and national trend alone allow the model to achieve about 60 percent accuracy in our calibration calibration, while the ALT error assumption achieves 40 percent accuracy. Only the inclusion of all our modeling assumptions allowed us to achieve 100 percent accuracy.

In Figure C.2.3, we reproduce Figure 2.3 from the paper with additional information. As in the original, the linear-normal model (in gold), which assumes independence, has confidence intervals that are extremely overconfident, and the LogisTiCC (in black) has accurately calibrated intervals. To these results, we add a version of our LogisTiCC that zeros out the parameters that model dependence. These include the national swing parameter σ_η and also our covariate stability parameter $\sigma_\beta > 0$ which, after transforming to the vote scale, also allows for some dependence across districts. In this model, we retain local uniquenesss.

Thus, we add to Figure C.2.3, in green, estimates from the LogisTiCC model constrained to give predictions with zero cross-district independence, while retaining local uniqueness. While this set of assumptions reduces overconfidence of the model relative to the normal somewhat, the model is still highly overconfident. Only when we allow our full ALT error structure with cross-district correlations are the out-of-

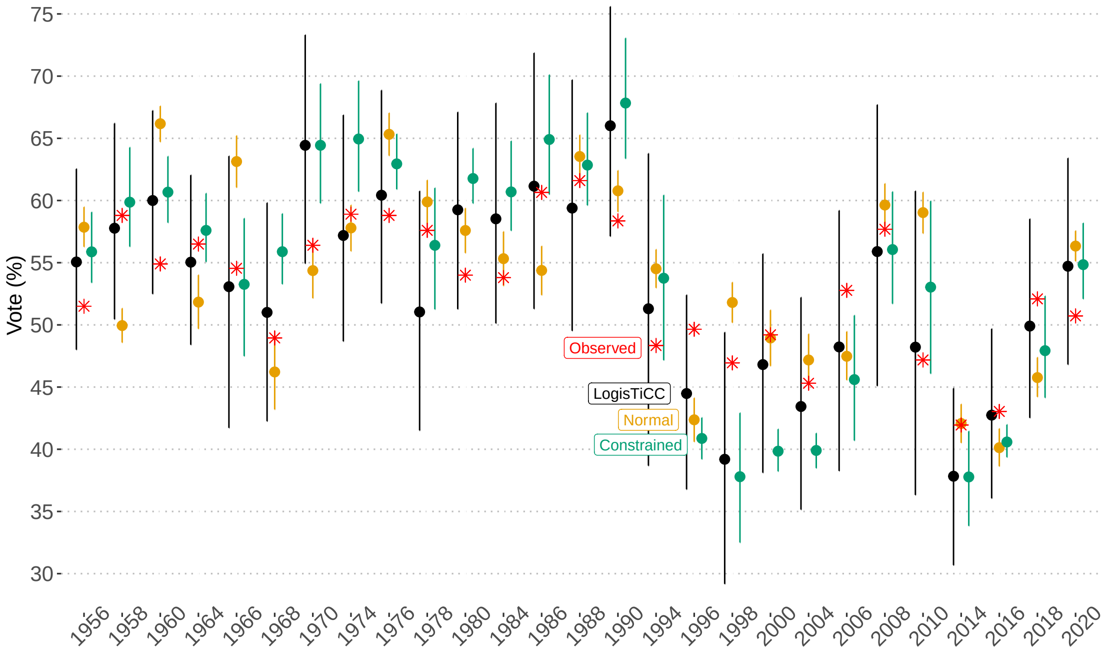


Figure C.2.3: Expected Vote Share of the Median House Seat (95 Percent Credible Interval).

sample model predictions from the LogisTiCC well-calibrated to the historical data (in black). Under the linear-normal error structure, the incumbent party will never lose control of the House of Representative. Under the ALT without cross-district correlation, the uncertainty gets larger so that the incumbent party is sometimes forecast to lose an election, but clearly not often enough. By introducing cross-district correlation, our forecasts are well-calibrated.

C.3 Imputation for Uncontested Seats

Missingingess due to uncontestedness is an important feature of historical congressional election data. In Figure C.3.4, we show the historical rate of uncontestedness in U.S. Congressional elections, which ranges from 21 percent in 1954 to 4 percent in 1996. Rather than drop these estimates which compose a nontrivial share of the data in any given election year, we impute predictive vote shares within our wholesale model framework.

To account for missing data due to uncontestedness, we jointly estimate a multivariate model which predicts the uncontested vote share and missing lagged uncontested vote share. To this end, we assume that missing vote share is a censored variable where an uncontested incumbent is constrained to always win. That is, we know

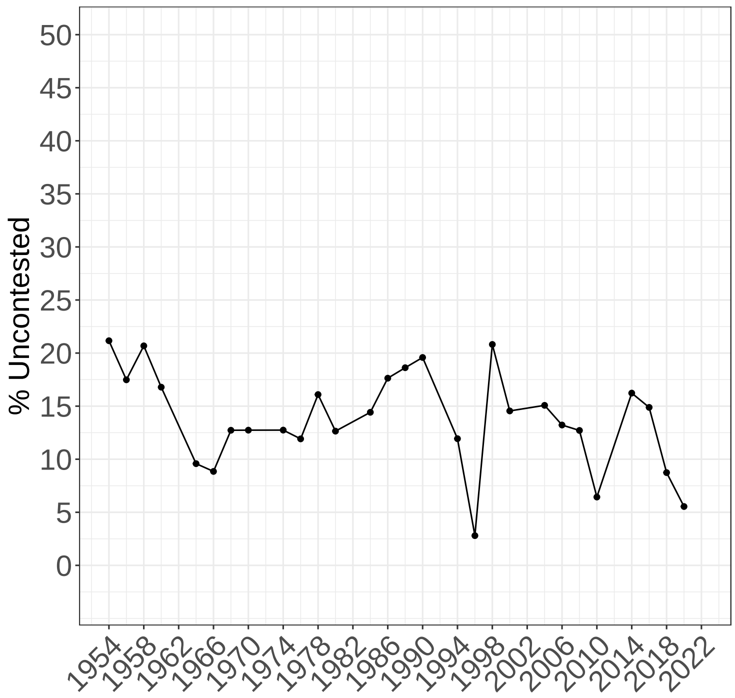


Figure C.3.4: Uncontested Elections over Time.

uncontested vote share data are not missing at random.

In Figure C.3.5, we show that our predictions are bimodal around modes centered at 25 and 75 percent vote shares. These predictions are in line for historical estimates of uncontested vote shares.

C.4 Computational Details

The standard approach is usually estimated with a linear regression for forecasting (i.e., dropping γ_i) or, for other quantities of interest, via an approximate two-step procedure designed to avoid computational challenges that were difficult in the 1990s (see Gelman and King, 1994).

Because of improvements in computation and Bayesian modeling, we estimate our LogisTiCC model via a fully Bayesian specification of Equation 2.2, beginning with the likelihood in Equation B.1. We implement the model in "brms," open-source software that uses Hamiltonian Markov Chains (HMC) sampling to draw from the posterior distribution of a mixed-effects model (Bürkner, 2018). In practice, we draw 50,000 samples of the posterior distribution from the Bayesian mixed-effects representation. When lagged congressional vote share is a covariate, we drop the first election of each redistricting decade to fit the model. Our Bayesian

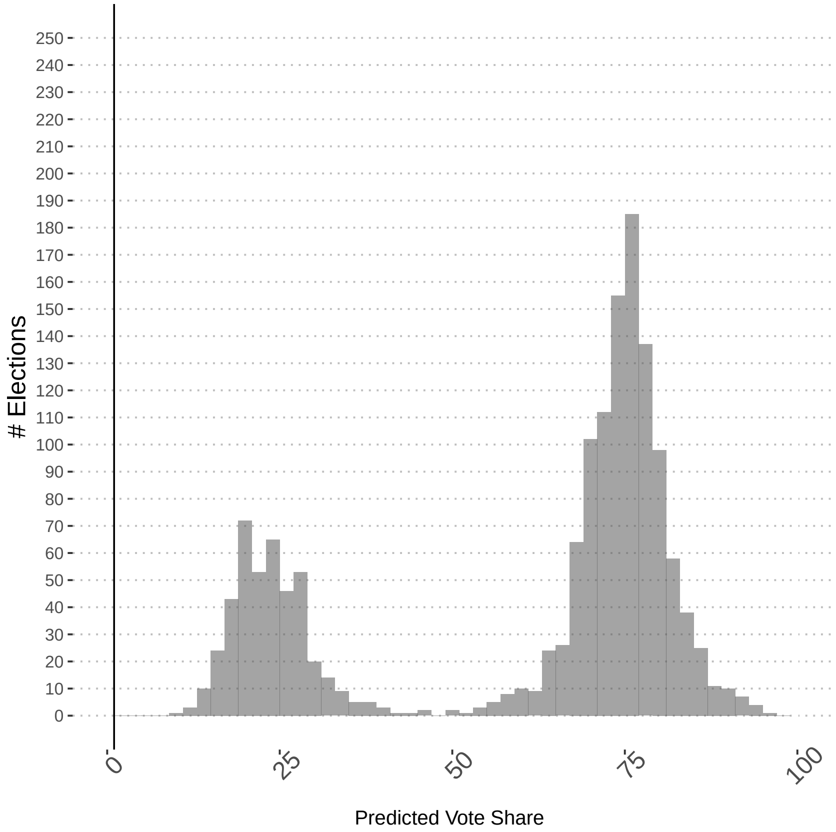


Figure C.3.5: Histogram of Predicted Values for Uncontested Elections.

methods are computationally demanding but efficient, which enables us to analyze large legislatures, and does not require asymptotic assumptions, which is especially important for legislatures like the small U.S. Senate class up for election in any one year, small national legislatures, or the many small state houses. We are also able to simulate quantities of interest directly from the full joint posterior distribution of the predicted values and parameters, which means researchers can easily calculate any relevant quantity of interest, along with accurate and calibrated uncertainty estimates.

In order to achieve valid calibrated uncertainty estimates, we use conservative search parameters for Stan's HMC sampler. We set a delta step of 0.99, set a maximum tree depth of 10, draw $50,000$ samples with a warm up of $5,000$ iterations on 5 chains run in parallel. All Markov Chains successfully converged, with no divergent transitions, Rhats of 1 across all parameters, well-mixed chains, and no breaches of maximum tree depth.

We employ weakly informative priors for estimation convenience. In our case,

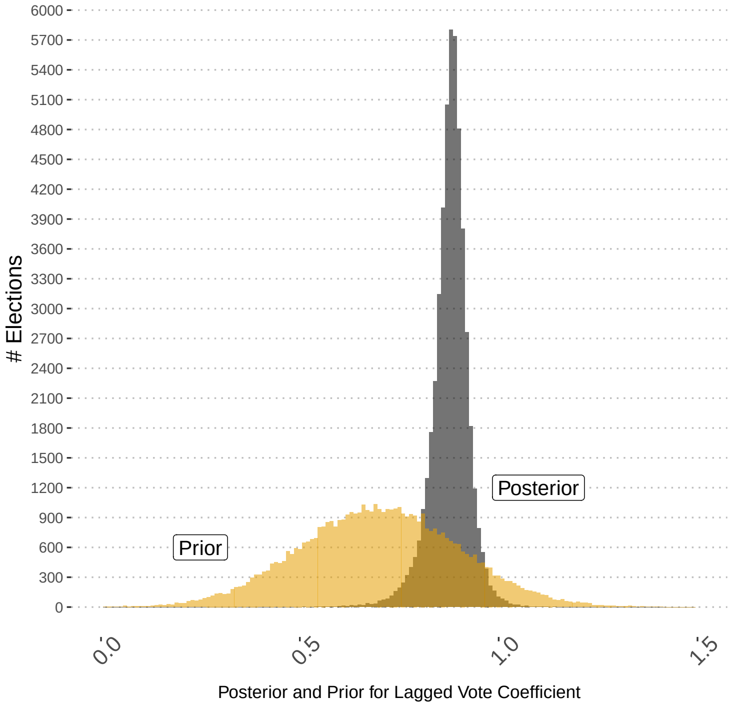


Figure C.4.6: Posterior vs. Prior Densities.

because we have an average of about $1,500$ elections per decade, we do not require regularization to identify model parameters, although our weakly informative priors reduce computational time for HMC convergence. Priors are useful for speeding computation but, in our data, the choice of hyperprior parameter values does not have much effect on empirical results. The specific values we use are $\sigma_\beta, \sigma_\omega, \sigma_{tk}, \sigma_i, \sim$ Exponential(0.2) and $\nu \sim \Gamma(3, 0.5)$.

In Figure C.4.6, we show the prior and posterior histograms for the coefficient on our predictor of the "normal" vote. This figure shows that our weakly informative prior is diffuse, while the coefficient posterior is tightly estimated around its mean, confirming that our model estimates are mostly a function of the data rather than priors. We have also found that small changes in the priors have little substantive consequences for our estimates.

Statistical results are likely less robust to the choice of the these parameters in smaller legislators. In applications with small legislatures, researchers should carefully consider the impacts of both prior specification and sampler behavior to guarantee statistically valid inference of the HMC chains.

C.5 Alternative Modeling Assumptions

We tried to eliminate any feature of our model not required for accurate out-of-sample validation and accurate uncertainty intervals, to include additional features that would improve performance, and to consider alternative specifications that might be easier to understand.

As we have shown in the main text, the linear-normal model is poorly calibrated for congressional elections. Additionally, we fit a linear-normal Student-t, which failed because it lacked the flexibility and asymmetry in the tails provided by the additive logistic *t* (ALT). The Additive Logistic Normal failed because it could not properly capture the levels of concentration (nearly 60 percent in the 1980s) exhibited in Figure 2.5a, nor did it accurately capture surprises with appropriate tails. Fitting an IID ALT, that is without contemporaneous correlations, is not well-calibrated because it misses the correlations due to year-to-year swings in the national trend or dependence due to the stability of coefficient estimates, as we showed in Figure C.2.3.

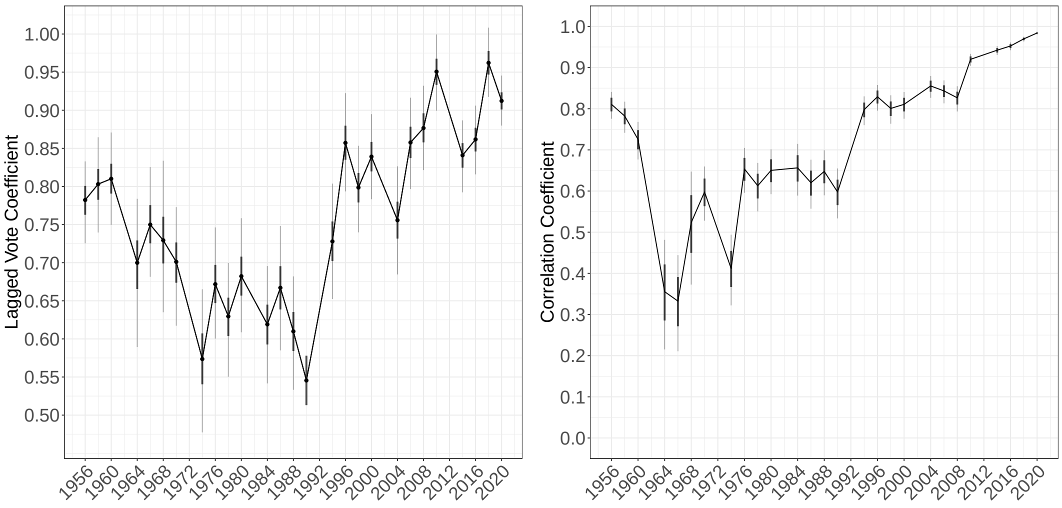
We also tried other flexible distributions. We tried the Beta distribution, which models the unit interval directly, but produces poorly calibrated results because it, like the IID normal, does not capture appropriate levels of concentration or tail behavior. We also tried mixture distributions and errors which, while flexible, wound up being highly model dependent, poorly identified, and computationally fragile.

We also attempted to find alternative correlation structures, besides time mixed effects and district random effects on the logit scale, such as regional mixed effects. Besides districts in the south and outside the south, there was little predictable inter-regional variation. Districts in the North, West, and Southwest do not seem to systematically vary, conditional on the covariates. Our covariates includes an indicator for districts in the South that varies over time to capture what appear to be the most important systematic effects. In terms of covariate selection, we made choices for easy comparison to the literature. Our general model structure, like the normal, can easily accommodate other indicators when desired.

C.6 Additional Information about The Three Regimes

We now give additional ways of distinguishing the three regimes described in Section 2.4. These regimes are also characterized by high levels of continuity, which we convey by a plot of the coefficient on the lagged vote from our model in Figure

C.6.7a ranging in 0.8–0.95 in the early and later periods, and as low as 0.3 in the middle period. We can also see high levels of partisan alignment during the same periods outside of our model by observing the correlation between the congressional and presidential vote. We construct a time series plot of these correlations in Figure C.6.7b, and they again reveal a now familiar U-shaped pattern.



(a) Election 1–Election 2 (Lagged Vote Coefficient)

(b) Congress–President Vote (correlation)

Figure C.6.7: Partisan Voter Alignment.

C.7 The History of Generative Modeling

To calculate generatively accurate descriptive summaries, the statistical model generating these summaries should (a) pass extensive, rigorous out-of-sample tests that validate its generative abilities and (b) reflect available prior information from the literature. In our efforts to meet these conditions, we benefit from developments in three major fields of statistics, each of which has engaged with these same conditions. We now situate the ideas described in this paper (particularly Section 2.5) in the history of statistical analyses by briefly describing these three research traditions.

First, direct attempts to build generative models in the social sciences have a long history, from path analysis originating in 1920s sociology, to linear structural equation modeling in econometrics and psychometrics in the 1960s and 70s, and, more recently, to hierarchical Bayesian models in statistics. At one point, econometricians had built many structural equation models of the economy, sometimes with hundreds of equations and each finely tuned to their past theoretical knowledge. However, rigorous out-of-sample forecasting were surprisingly embarrassed by a comparison with “atheoretical” univariate ARIMA models, leading many to reassess the value of their prior information. These attempts failed because researchers lacked the requisite computational resources to build models that reflected prior knowledge and sufficient data to make extensive validation possible. Now, model checking has become a more common part of Bayesian best practices (e.g., Gelman, Meng, and Stern, 1996).

Second, when estimating accurate generative models was not feasible, or required too many unjustified assumptions, social scientists turned to other research frameworks, often changing their quantities of interest in the process. Most notably, the literature on causal inference, especially since the 1980s, has made tremendous progress by developing ways of estimating causal effects without modeling assumptions. Although numerous articles had previously attempted to make causal inferences, Leamer (1983) and others pointed out that high levels of (what came to be known as) model dependence meant that most of these inferences were not right, and maybe not even wrong, but instead mostly reflected researchers’ priors. The “credibility crisis” that resulted from this skepticism and from rigorous tests of observational estimates compared with out-of-sample randomized experiments (Lalonde, 1986), lit a fire under the methodological community, resulting in remarkable progress that continues until today (Imbens, 2022). The theories and descriptive stories that emerge from generative models, including ours, often include many causal effects, and so the ability of these methods to proceed without modeling assumptions has been valuable for everyone. At the same time, even if we had exact knowledge of all causal effects ever estimated and a vast number of others, we would not come close to the range of descriptive knowledge social scientists seek and which can be gained by generatively accurate descriptive summaries. Descriptive quantities such as partisan bias, responsiveness, forecasts, farcasts, and many others are not causal effects but of course remain of interest to political scientists and policymakers.

Finally, machine learning methods of classification and prediction have made continual progress by their single-minded focus on out-of-sample validation. By taking their task as engineering better algorithms and downplaying constraints suggested by prior theoretical “knowledge,” they make themselves continually vulnerable to being proven wrong. Although one can often do as well with simpler models that explicitly code more prior knowledge, this literature’s focus on validation helps them avoid being fooled by elegant theories that do not have empirical support.

As has been true throughout the history of quantitative social science methodology, political scientists have a comparative advantage when they employ their knowledge of the political world, but do best when subjecting their statistical claims to the possibility of being proven wrong.

Appendix D

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

D.1 Derivation of Gradient for Stochastic Gradient Descent

- k number of sentiment-topics ($S * T$)
- h - sentiment-topic mixture
- $w_1, ..w_v$ - words in a document, $w_i \in \mathbf{R}^d$
- d vocabularly size
- $\mu = [\mu_1, .., , \mu_k]$, $E[w_i|h] = \mu h$
- $c_t = (c_{1,t}, ...c_{d,t}) \in \mathbf{R}^d$, centered frequency vector for documents for the t-th document
- N number of documents
- $y_t = c_t W, \quad W \in \mathbf{R}^{d\times k}, \quad y_t \in \mathbf{R}^k$

We have whitened tensor of centered data,

- $v : k \times k$
- $Y : N \times k$ (Whitened counts, centered)
- k : number of sentiment-topics
- N: batch size

We want to solve for

$$L(\mathbf{v}) = \frac{1}{n_x}\sum_{t=1}^{n_x} \frac{1+\theta}{2}||\sum_{i=1}^{k} \otimes^3 v_i||_F^2 - \left\langle \sum_{i=1}^{k} \otimes^3 v_i, \mathcal{T} \right\rangle$$

This gives loss function

$$L(v) = \frac{1+\theta}{2}||\sum_{i=1}^{k} v_i \otimes v_i \otimes v_i||_F^2 - \left\langle \sum_{i=1}^{k} v_i^{\otimes^3}, \frac{(\alpha_0+1)(\alpha_0+2)}{2N}\sum_{t=1}^{N} y_t^{\otimes^3} \right\rangle$$

Now, notice we can write

$$\begin{aligned}
& ||\sum_{i=1}^{k} v_i \otimes v_i \otimes v_i||_F^2 \text{ as} \\
& ||v_i(v_i \circ v_i)^T|| \text{where } \circ \text{ is the column-wise Kronecker product} \\
= & \text{ Trace}\Big((v_i \circ v_i)v_i^T v_i(v_i^T \circ v_i^T) \text{ from the definition of the Frobenius norm}\Big) \\
= & \text{ Trace}\Big((v_i^T \circ v_i^T)(v_i \circ v_i)v_i^T v_i\Big) \text{ from permutation invariance of the Trace operator} \\
= & \text{ Trace}\Big([(v_i^T v_i) * (v_i^T v_i)]v_i^T v_i\Big) \text{ property of } \circ
\end{aligned}$$

where $*$ is the Hadamard product (element-wise product). Similarly,

$$\begin{aligned}
& \left\langle \sum_{i=1}^{k} v_i^{\otimes^3}, \sum_{t=1}^{N} y_t^{\otimes^3} \right\rangle \\
= & \left\langle v(v \circ v)^T, Y^T(Y^T \circ Y^T)^T \right\rangle \\
= & \text{ Trace}\Big((v \circ v)v^T Y^T (Y \circ Y)\Big) \\
= & \text{ Trace}\Big((Y \circ Y)(v \circ v)v^T Y^T\Big) \\
= & \text{ Trace}\Big((Yv) * (Yv)v^T Y^T\Big)
\end{aligned}$$

Finally, taking the derivative of L with respect to v, we have

$$\frac{\partial L}{\partial v} = 3(1+\theta)v\left(\hat{v}^T\hat{v} * v^T v\right) + \frac{3(\alpha_0+1)(\alpha_0+2)}{2n_x} y^T(yv * yv)$$

D.2 Moment Derivations

$$E[w_1|\theta,\pi] = \sum_l \sum_z \pi_l \theta_{l,z} \phi_z^l \quad \text{(D.1)}$$

$$= \left[\phi_1^1, \phi_2^1, ..., \phi_T^1, \phi_T^2, ..., \phi_T^S\right] \begin{bmatrix} \pi_1\theta_1 \\ \pi_2\theta_2 \\ \vdots \\ \pi_S\theta_S \end{bmatrix} \quad \text{(D.2)}$$

Now, taking the expectation,

$$\begin{aligned} E\left[\sum_l \sum_z \pi_l \theta_{l,z} \phi_z^l\right] &= OE[h] & \text{(D.3)} \\ &= O \begin{bmatrix} E[\pi_1\theta_1] \\ E[\pi_2\theta_2] \\ \vdots \\ E[\pi_S\theta_S] \end{bmatrix} & \text{(D.4)} \end{aligned}$$

And we have by distributional assumption

$$E[\pi_l \theta_{l,z}] = \gamma_l \alpha_{l,z}$$

Which means

$$\begin{aligned} E[h] &= \begin{bmatrix} \gamma_1\alpha \\ \gamma_2\alpha \\ \vdots \\ \gamma_S\alpha \end{bmatrix} & \text{(D.5)} \\ &= \begin{bmatrix} \gamma_1\alpha_1 \\ \vdots \\ \gamma_1\alpha_T \\ \gamma_2\alpha_1 \\ \vdots \\ \gamma_S\alpha_T \end{bmatrix} & \text{(D.6)} \end{aligned}$$