Policy and the Structure of Roll Call Voting in the U.S. House

Scott de Marchi
Duke University
demarchi@duke.edu

Spencer Dorsey
Duke University
spencer.dorsey@duke.edu

Michael J. Ensley
Kent State University
mensley@kent.edu

Electronic copy available at: https://ssrn.com/abstract=3262316

Abstract

Competition in the U.S. Congress has been increasingly characterized by competition along a single, left-right dimension. However, we challenge this characterization by showing the content of legislation has far more predictive power, particularly for members of the minority party, than the characteristics of members' constituencies and their own characteristics, most notably their ideological position derived from scaling roll call votes. Using a machine learning approach, we identify a topics model for final passage votes in the 111th through the 113th House of Representatives and conduct out-of-sample tests to evaluate the predictive power of bill topics relative to other factors. We find that bill topics and congressional committees are important for predicting roll call votes but the other variables, including member ideology, lack predictive power. These findings raise serious doubts about the claim that congressional politics can be boiled down to competition along a single left-right continuum and shed new light on the debate about levels of polarization in Congress.

Electronic copy available at: https://ssrn.com/abstract=3262316

The sorting of the Democratic and Republican Parties in the U.S. Congress into two ideologically cohesive teams who are increasingly divided along a single, left-right dimension of political conflict is one of the most recognized, studied, and cited phenomena in the contemporary study of American politics. The traditional account of this sorting and polarization along a single dimension is that staring in the 1960s the Democratic Party became the party of racial liberalism as well as the party of liberalism on economic and social-welfare issues, which created a shift in the composition of the two parties (Carmines and Stimson 1989; Aldrich 1995). The Republicans coalesced into a group of conservatives who grew increasingly distinct from the emerging group of liberal Democrats and the structure of competition between the parties collapsed onto a single, left-right dimension.

The most frequent source of evidence cited for the increased polarization of the parties and the reduced dimensionality of political competition in Congress is the changing pattern of roll call voting in Congress using Poole and Rosenthal's (1991, 1997, 2007) NOMINATE scores. Scaling procedures such as NOMINATE purport to show that since the 1970s the ideological distance between the average Republican and Democrat member in Congress and the amount of variance in roll call voting explained by the first dimension have increased. The implication of these findings is that congressional roll call decisions can be explained by a member's position on the left-right dimension.

The push to explain the causes and consequences of polarization may have been premature given that we have not precisely determined the nature of congressional voting and agenda setting in Congress (Clinton 2012, 2017). Consider that, despite research and ubiquitous journalistic accounts of increasing polarization, bipartisanship still exists and is more widespread than many recognize (Adler and Wilkerson 2013; Curry and Lee 2017; Harbridge 2015). Some

Electronic copy available at: https://ssrn.com/abstract=3262316

scholars have noted that a focus on roll calls tends to exaggerate the level of ideological polarization (Carson et al. 2010; Egar 2016). More generally, it is not clear that just because the parties have sorted into two voting blocs that they are more polarized on the issues. Adler and Wilkerson (2013) argue that the congressional agenda is dominated by compulsory issues and voting patterns are more bipartisan than the literature on congressional roll call voting would lead one to believe. More pertinently, consider the findings of Bateman et al. (2017), who show that the parties have become less polarized on Civil Rights issues if one accounts for the content of legislation in analyzing roll call votes. Finally, the traditional account of ideological polarization assumes that competition occurs largely on a single policy dimension but there is strong evidence that competition is better captured in multiple policy dimensions (Crespin and Rohde 2010; Aldrich et al. 2014; Roberts et al. 2016).

We challenge the proposition that roll call decisions can be adequately explained by a member's position on the left-right dimension of competition. On the contrary, we show that roll call decisions are more accurately predicted by accounting for the policy content of legislation, thus implying the there are many distinct policy dimensions underlying political competition in Congress.[1] We demonstrate this by predicting members' decisions on final passage votes in the contemporary House of Representatives using the characteristics of the individual bills derived from a structural topics model (Roberts et al. 2014). We compare the predictive power of the content of bills against other explanatory variables: (1) the committees that originated the bills; (2) the characteristics of the member's constituency; and (3) the member's first and second dimension NOMINATE scores as a measure of a member's ideological predispositions.

[1] Note that not all members of Congress may care equally (or at all) about all dimensions in a higher dimensional policy space.

Electronic copy available at: https://ssrn.com/abstract=3262316

There are several notable findings. First, we are able to predict roll calls on final passage votes for minority members with 94 percent accuracy on final passage votes, which represents a considerable improvement in the out-of-sample predictive accuracy over alternative models. Second, this accuracy is driven mainly by the substantive content of bills. NOMINATE scores in particular do not improve predictive accuracy, which casts serious doubts on the thesis that a single ideological dimension captures congressional decision making. Relatedly, the relative importance of the bill topics and the committee variables imply that the policy space is highly multidimensional, as demonstrated by Roberts et al. (2016).

Taken together, these results and the literature on bipartisanship in Congress suggest that congressional scholars need to reconsider the meaning and implications of the roll call record for understanding polarization and policy outcomes. In last part of the paper, we take a first step in this direction by examine the substantive areas where the minority party is most likely to oppose the majority party.

Theories of Congressional Roll Call Voting

In his detailed account of congressional decision making, Kingdon claimed that the problem confronting members of Congress in roll call voting is typical of the dilemma confronting decision makers of all kinds: "As a nonexpert who is obliged to make unavoidable decisions about matters in which he is little schooled and to which he is able or willing to devote only limited time and attention, how does he go about making up his mind?" (Kingdon 1989). What Kingdon acutely understood was that members of Congress have neither the time nor the incentive to become fully informed consumers of the legislation they vote on. First, members have too many votes to consider; they have to make choices how to allocate their attention given

Electronic copy available at: https://ssrn.com/abstract=3262316

the complexity of legislation.[2] Second, members have other more valuable activities to invest their time in, such as meeting with lobbyists and constituents, campaigning, and fundraising.

So in lieu of assuming members of Congress are policy experts, Kingdon built a theory of decision making that started with the premise of limited attention and resources. Members gather information and are influenced by other political actors and are guided by their own ideological system of beliefs. But despite its importance, Kingdon cautions against using ideology as the sole determinant in explaining votes:

> Finally, given the obvious importance of the congressman's own policy attitudes in his voting, it would be tempting to adopt an ideological model as one's sole explanation for legislative voting behavior. For a number of reasons, however, I have chosen instead to treat general policy attitudes as one important pre-consensus process which may structure a decisional context, rather than an alternative model. There should be a clear distinction made, in the first place, between general policy orientations and attitudes toward specific pieces of legislation. Congressmen may well have rather well-informed general orientations. But partly because of the sheer volume and complexity of issues upon which they are called to vote, their policy attitudes towards specific amendments and bills are likely to be much less well formed in advance of a given vote (1989, 273).

Thus, for Kingdon, members possess an ideology that helps structure how they approach an issue, who they listen to, and where they gather information. But ideology is not the sole consideration in decision making in his account; the importance and visibility of the issues is considered, how much time members have available is a constraint, and the signals of other actors (e.g., is there consensus or disagreement?) may be influential.

Despite Kingdon's cautions about a purely ideological model of voting, this model now dominates the literature. The spatial model of roll call voting is the foundation of many theories of congressional behavior and debates have been focused on how much ideology matters relative to party (Krehbiel 1992, 1993; Cox and McCubbins 1993, 2005; Smith 2008). These debates have been informed by the measurement of legislator ideology using roll call votes, most notably

[2] Modern Congresses do, of course, have large professional staffs, which likely serves to attenuate the informational problem members face.

Electronic copy available at: https://ssrn.com/abstract=3262316

Poole and Rosenthal's groundbreaking NOMINATE procedure (1991, 1997, 2007). Many scholars, including Poole and Rosenthal, have argued that roll call votes and the resulting summary measures of voting behavior such as NOMINATE scores, are expressions of members' underlying predispositions or ideology.[3] These measures have been used in countless studies that examine legislator behavior and the broader structure of Congress; most recently, Nominate scores have been used to demonstrate increasing polarization in Congress. While there are analyses that consider other influences on roll call voting and examine the limits of ideology as an explanatory factor, Poole's (2007) claim that members of Congress "die with their ideological boots on" represents the core approach of the Congress field.[4]

Recent scholarship, however, has raised doubts about this claim, demonstrating that scaling procedures such as NOMINATE may obscure the dimensionality of roll call voting. Crespin and Rohde (2010) show that by limiting the sample to annual appropriation bills, one finds that member preferences are multidimensional. Anderson (2012) finds that even though domain-specific measures of legislator scores are highly correlated with general measures of ideology such as NOMINATE, there are areas such as agricultural policy where there are important differences that can be consequential for understanding congressional policymaking.

[3] Despite the continued attention to NOMINATE, important theoretical and empirical questions remain about roll call based measures and broader patterns of congressional behavior and organization (c.f. Clinton 2012, 2017). Some do not view NOMINATE scores as pure expressions of legislator ideology. Instead, roll call votes are decisions affected by myriad competing pressures and the summary measures created from these votes are amalgams of these various forces. A similar line of argument centers on whether members change their ideological positions or whether they remain consistent throughout their career (Rothenberg and Sanders 2000; Poole 2007).

[4] While there are periods of American history where there may be another separate and significant dimension that underlies roll call voting, such as when racial issues divided parties from the New Deal through the Civil Rights Era, a single left-right dimension is seen as dominating member behavior.

Electronic copy available at: https://ssrn.com/abstract=3262316

And Jochim and Jones (2013) show that the ideological polarization of parties in Congress has not been a uniform process across 18 issue areas they identify using data from the *Policy Agendas Project* from 1965-2004.

These findings are given further support by Roberts et al. (2016) who showed that multidimensionality in preferences is the typical pattern when roll calls are examined within bill episodes, as well as within major policy areas. What is particularly interesting is that this pattern of multidimensionality begins to disappear when votes are aggregated, suggesting that the typical measures of preferences that aggregate across policy areas may be obscuring the differences across members of Congress.[5]

Finally, scholarship that demonstrates that there is a large degree of bipartisanship in the legislative process (Adler and Wilkerson 2013; Harbridge 2015; Curry and Lee 2017) suggests that the content of legislation cannot be classified by a simple left-right divide. And, without some understanding of the content of bills, it is difficult to disentangle ideological stories about polarization from the (possibly) changing content of legislation.

These lines of research thus imply that the content of legislation may well be a critical component of roll call voting. But given that many scholars have been working with a theoretical model that assumes a unidimensional policy space, combined with the depth and widespread acceptance of Poole and Rosenthal's (1997, 2007) findings, the substantive content of bills has been largely ignored in the literature. If everything maps to a single dimension, there is no need to investigate the content of individual bills.

The leading alternatives in the literature to individual ideology as the central determinant of voting also discount the content of bills. One way for members to make decisions without

[5] See Aldrich et al. (2014) and Dougherty et al. (2014) for possible explanations of this finding.

Electronic copy available at: https://ssrn.com/abstract=3262316

dealing with the full complexity of bill content is to focus on the committee of origin. Shepsle (1979) articulated a theory of congressional organization that placed committees as the central feature of legislator voting behavior. Committees divide legislative activities into a discrete and mutually exclusive set of policy jurisdictions, which award monopoly power to members of the committee. This arrangement helps members solve the social choice problems of legislating and serves members' reelection goals by allowing them to dictate policy in those areas that a member's constituents care most about. Shepsle's structure-induced theory of congressional organization however suggests there should be little opposition to bills on the floor.

A different theory of congressional organization that still places committees in a central role is from Krehbiel (1991). He argues that members face a high level of policy uncertainty and need an efficient means to learn about the content and potential policy consequences of legislation. Members created and maintain a committee system that effectively signals information to legislators on other committees. Ideologically diverse committees can perform this task because one can infer the content of a bill by observing which members of the committee support or oppose it.

We thus have a number of different explanations for why members of Congress vote the way they do and few attempts in the literature to validate these different models. In particular, Nominate is difficult to test out-of-sample and little attention has been paid to building predictive models of roll call votes. In what follows, we build a predictive model of roll call voting in the House and test the competing theories, focusing our attention on whether or not the substantive content of bills is crucial in explaining legislator voting behavior.

Minority Party Roll Call Voting

Electronic copy available at: https://ssrn.com/abstract=3262316

We also focus here on the voting behavior of legislators in the minority party because this is where polarization should have teeth – absent cooperation, members of the minority party should vote "nay" on the meaningful roll calls which constitute our sample.

Much less attention has been focused on the roll call voting of minority members. This is understandable, especially given the emphasis on the role of the majority party in controlling the agenda in Congress, as well as attempts to cajole members to support the leadership (Cox and McCubbins 1993, 2005; Rohde 1991; Aldrich 1995; Evans 2004; Lee 2009). That said, voting decisions of minority party members are important for assessing the level of polarization present in the Congress.

To understand polarization one would have to pin down the reason why the minority party legislators do or do not vote for bills. Consider that the rate at which majority and minority party members support final passage votes on the floor of the House has not decreased in recent terms of Congress, which is contrary to what one would expect if the parties polarized ideologically. Figure 1 shows that percentage of "Yea" votes for the majority and minority parties from the 83rd through the 114th congresses.[6] The percentage of Yea votes has increased slightly over this period, whereas the percentage of Yea votes for the minority party has varied but generally not decreased. If the parties have become more ideologically polarized over this time period, one would expect that the percentage of time the minority party supports final passage votes would decrease. Given that almost all final passage votes are policies supported by a majority of the majority party (Cox and McCubbins 2005), this is a puzzling finding.

Focusing on minority members is also a sterner test for a statistical model of roll call voting. A naïve model of the majority that predicts each member will always vote "yea" will

[6] This data is from Crespin and Rohde (2016).

Electronic copy available at: https://ssrn.com/abstract=3262316

achieve over 90% accuracy. The minority, which has more variance in play, is much more difficult to predict. Additionally, we expect minority party members to be more responsive to constituency pressures as well as other considerations (Clinton 2006; Smith 2007).

Bill Topics and Legislator Voting Behavior

Our main interest is in using the content of bills to explain legislator voting behavior; doing so depends on extracting topics from the full text of bills before Congress. But, it is important to discuss why feature extraction of this kind is important in the context of predicting votes. After all, if one wants to intervene as little as possible between the data and the model, one could instead build a statistical model that uses word counts from each bill directly. While the total dictionary size for a modern term of Congress is nearly 16,000 unique words (thereby producing an equal number of IV's), regularization could trim this space substantially.

Unfortunately, regularized regression models fail in the context of predicting votes in Congress for three reasons. First and unsurprisingly, regularized regression models using the complete dictionary of words as IV's underperform regressions using topic models by a sizable amount (for details, see the Appendix). Second, a high degree of correlation exists between words within the same bill and no solution (e.g., n-gramming) entirely eliminates this problem. In general, the correlation between features blunts the ability of high-dimensional regression methods such as the lasso or random forests to successfully capture the systematic component in the data (Conn and Ramirez, 2016). Last, while we have a large number of observations due to 435 members of Congress voting on the passage of several hundred bills, this is not as plentiful a source of data as it seems. Bill characteristics (whether words or topics) are fixed for each member of Congress and legislator characteristics are fixed across bills; this is close to a

Electronic copy available at: https://ssrn.com/abstract=3262316

situation where there are more columns than rows. The bottom line is this: our "best" models in terms of out-of-sample accuracy depend on aggressively selecting a parsimonious feature space (Bishop, 1995). Feature extraction based on topic models are essential if one wants accurate and stable results of how legislators' vote and they have the welcome side-effect of allowing us to more easily interpret the substantive implications of our models.

For this reason, we follow the machine learning literature and engage in the automated extraction of features (i.e., topics) of bills. The main advantage to an automated approach is that it is fast; the disadvantage is that it embeds assumptions about the data generating process (DGP). The best examples of automated feature extraction and the use of text as data are Gerrish and Blei (2011); Wang, et al (2013); Grimmer (2013), Roberts, et al. (2013), and Kim, Londgregan, and Ratkovic, 2014. Using topics as independent variables thus allows us to build simple regression models of roll call votes.

We build a predictive model of roll call voting using *bill topics* as predictors of final passage votes in the U.S. House. If we are able to predict roll call votes using topics while controlling for other factors that are commonly identified as central in the literature, this will substantiate the importance of policy issues as critical to explaining roll call voting. Further, it will corroborate the findings of Roberts et al. (2016) that the policy space underlying congressional decision making is multidimensional.

The first step in this process is to develop an approach to classifying the content of bills, which is described in detail in the Appendix. Given the length of each bill and the large number of bills, the coding of bills individually "by hand" is not feasible. To classify the issues involved in each bill we employ supervised machine learning methods. This research is possible because new sources of data exist along with new statistical techniques for handling these data. Gerrish

Electronic copy available at: https://ssrn.com/abstract=3262316

and Blei (2011) and Wang, et al. (2013) develop and apply a technique for extracting topics from the full text of bills – Latent Dirichlet Allocation (LDA) – and then use the resultant topics in regression models of roll call votes. LDA, however, has issues with multiple local optima. For this reason, we use an extension of LDA, the structural topic model (STM) developed by Roberts, et al. (2014).[7] For a list of the topics we ultimately estimated, as well as the frequency and exclusivity of terms within these topics, see Tables 1a and 1b.

Data and Model

To investigate roll call voting, we focus on the 111th through the 113th House of Representatives. This is a good period to investigate minority party voting because Democrats were in control of the House in the 111th and Republicans had a ruling majority in the 112th and 113th congresses. In these three congresses, we examine whether members support the final passage of a bill. Supporting the final passage of the bill is tantamount to supporting the position of the majority party given agenda control and the fact that almost all final passage votes are supported by a majority of the majority party; only a few votes of the hundreds of the votes in these three congresses were not supported by a majority of the majority party in the House.

Focusing on final passage votes avoids problems of lumping very different types of votes into a single sample. This is also necessary if one is to be able to use topics as regressors – implicitly, we are leveraging the fact that if a bill makes it to the floor, the majority leadership is in favor of it passing. Without this prior, there is no easy way to interpret the expected direction of the effect of any given topic in a regression. Trivially, it seems obvious that a topic of a bill

[7] In the case of LDA, we found that multiple models provided the same fit to the data albeit with very different topic – word mappings. The overall fit of the regressions predicting roll call votes (reported in the results section of this paper) were the same with LDA. STM, in contrast, provides a solution to the issue of multiple modes, but does not improve the performance of the models over LDA. The main advantage, in our minds, of using STM is ease of replication.

Electronic copy available at: https://ssrn.com/abstract=3262316

focused on "women" might mean something very different if the authors and supporters are Republicans instead of Democrats.

Committees, District Characteristics, and Ideology

As noted in the literature review, the text of bills has not been the main focus of Congress scholars. To account for the prevailing theories of Congressional behavior and test whether bill content matters, we include sets of independent variables that map to the leading theories in this area. We start with the committee system itself. From the work of Shepsle (1979) on structure-induced equilibrium, we have every reason to believe that the committee system structure separates the salient dimensions of choice in the Congress. In the simplest case where there are not overlapping jurisdictions, each committee is primarily responsible for a single dimension. Additionally, "deference to committees, long considered an important norm in Congress, suggests that what the committee does will have a great bearing on what Congress chooses" (Aldrich, 1995, p. 201). We include a set of binary variables that measure the committees of origin for a bill. This is particularly important because it presents a strong test for the value of adding topics derived from the full text of bills – it is entirely possible that once one knows the committee of origin, one does *not* gain any additional information from bill topics.

We also include variables that account for the policy preferences of legislators' constituents and the strategic context of their districts. Since the seminal study by Miller and Stokes (1963), a basic concern in the study of Congress is whether elected members are responsive to the constituents in their district. We account for district preferences in several different ways using survey data from the 2010 and the 2012 Cooperative Congressional Election Studies (CCES). The CCES studies in those years has over fifty-thousand respondents,

Electronic copy available at: https://ssrn.com/abstract=3262316

and thus has a sufficient number of respondents per House district to reliably estimate the average district-level policy and ideological positions of the constituents. First, we estimate the average partisanship of the district by taking the average placement of the citizens in each district on the standard seven-point partisan identification scale. Next, following Ellis and Stimson (2012), we estimated the symbolic ideology of constituents using the five-point, self-identification ideological scale and the operational ideology using a series of policy questions. For the latter, we conducted a principal components analysis of a large group of policy questions and calculated the first factor score. We expect that when Republicans are in the minority, they will be less likely to support the final passage of a bill the more conservative and Republican their district is; we expect that when Democrats are in the minority, they will be less likely to support final passage the more liberal and Democratic their district is.

We also used policy questions to calculate the operational ideology of constituents and estimate the complexity of citizens' preferences in the district. Ensley, et al. (2009) have shown that the complexity of preferences, or the diversity in the structure of citizens' policy preferences, affects how well incumbents perform. Incumbents do better in more complex districts. For similar reasons, we surmise that complex districts may lead minority party members to vote more often with the majority party given the larger amount of uncertainty about the best position in their district. Complexity is measured here as the negative of the first eigenvalue from the principal components analysis used to estimate operational ideology. The eigenvalue captures how much of the variance in the policy questions is captured by the dimension, thus the larger the eigenvalue, the simpler the arrangement of citizens' policy preferences in the district.

Electronic copy available at: https://ssrn.com/abstract=3262316

In addition to the policy characteristics of a member's constituency, we also control for the level of electoral competition in the district. We include the margin of victory for the incumbent in the last election, the margin of victory for the presidential candidate in the most recent election, a binary variable for whether the incumbent's party lost the presidential vote, and the strength of the primary challenge. For the latter, we include the percent of the vote the incumbent received in the last primary, as well as the margin of victory over the next closest competitor. We also include binary variables for whether the incumbent was a freshman and whether the incumbent had announced his or her retirement.

Finally, we test for the effect of a member's ideology by using Poole and Rosenthal's (2007) DW-NOMINATE scores, which are dynamic measure of a members' ideology estimated by scaling roll call votes over the course of a member's career.[8] We use both the first and second dimension scores given there is evidence that the second may capture some important elements of a member's policy preferences. We expect that when Republicans (Democrats) are in the minority, the more conservative (liberal) the member is the less likely he or she will be to support the final passage of a bill.

Given the number of independent variables, it is important to avoid overfitting the sample. Accordingly, we use k-fold cross validation with five folds to focus on out-of-sample performance. The model used in all cases was a simple logit model:

Vote ~ f(topics, committees, constituent characteristics, Nominate scores)

[8] We used scores from the term we were modeling. Thus, for these two variables, we are not engaging in out-of-sample work because the votes of the term were used to create the Nominate scores. A more reasonable approach would have been to use scores from the prior term and not to include information from subsequent votes in members' careers. Our approach will, however, *overestimate* the influence of Nominate and thus represents a strong test.

Electronic copy available at: https://ssrn.com/abstract=3262316

The dependent variable is roll call votes in the House – which by definition are either "yea" or "nay"[9] – but not all votes are created equal. In fact, a large number of votes are procedural or amendments rather than votes that determine the passage of a piece of legislation. It is difficult to argue that these types of votes spring from the same DGP; worse, including procedural votes swamps the more significant votes that determine passage. While there are examples of significant procedural and amendment votes, there is no reason to lump different types of votes together. Accordingly, our sample consists only of final passage votes. In the 111th House there are 297 final passage votes, in the 112th there are 274, and for the 113th there are 293; together, this produces over 300,000 roll call votes.

The standard for evaluating performance is how well the model fits the held-out sample compared to a reasonable baseline. In the sessions of Congress considered here, the baseline model is the percentage of "yea" votes in each session; trivially, a model that predicts "yea" for every observation will do well. The regression model thus includes all of the following variables:

Summary of Independent variables

Type	Source
Topics (50)	Derived from STM topic model (including a binary indicator and a measure of the proportion of the topic for each bill). We also include the squares of each topic.
Committee of origin of bill	Binary variables for each standing committees
Member Characteristics	Legislator's margin of victory in last race, Primary vote share, Primary margin of victory, Binary for Freshman Incumbent, Binary for announced retirement
District Characteristics	Partisanship, Symbolic and Operational ideology, and Complexity of preferences (all from 2010 and 2012 CCES) Margin of victory for president in previous election

[9] There can also be a value of "not voting". One thus has the two main options of either treating these observations as missing or as a "nay" vote – we settle on the latter for the work presented here but results are robust to coding the values as missing instead. For an overview, see Keith Poole's website: http://voteview.com/

Electronic copy available at: https://ssrn.com/abstract=3262316

Member Ideology	First two dimensions of DW-NOMINATE

While this model may have the feel of a "kitchen-sink" regression that includes every variable we can collect, we are validating it by how well the model performs out-of-sample. Finally, given the structure of topic models, we estimate the above regression for each party and term of the House separately.

Results

We begin the analysis by examining how well the model with the complete set of variables predicts roll call votes in the House with a focus on the minority party. As indicated in Table 2, that probability that a minority party member votes in favor of final passage is between 62 and 67 percent and the appropriate baseline model would be that all minority party legislators vote "yea" on every bill. The percent predicted correctly, or accuracy, of our model is approximately 94% out-of-sample for the two congresses when the Democrats are in the minority. The model is less accurate for the 111th Congress when the Republicans are in the minority with an accuracy of 86%. Overall, our models provide a large improvement in fit but with a clear difference in accuracy between the Republican and Democrat minorities.

Given that model is doing significant work in predicting roll call votes out-of-sample, the next question is which variables are most important to the model. To explore this question, we reran the models with regularization to push marginally useful coefficients to zero. In short, the answer is that the topic variables, followed by the committee variables, are doing the most work in predicting roll calls. The constituency variables are doing considerably less, as are the variables related to the members' characteristics. Of particular note, is the observation that the

Electronic copy available at: https://ssrn.com/abstract=3262316

ideology of the members as measured by the DW-NOMINATE scores is not a significant factor in predicting roll call votes.

In Tables 3 through 5 we present the top 15 predictors from each of the logistic regressions with regularization from the three House terms. The variables are ranked according to the absolute size of the standardized coefficient. These results show that for each congressional term, the top predictors are almost exclusively topic variables or committee variables, with topics dominating. In fact, the only other variable that is in the top 15 predictors is the first dimension DW-NOMINATE score in the 111th Congress.

Among the top 15 variables in Tables 3, 4, and 5, the relationship between topics and roll call votes also makes sense substantively. For example, in the 112th Congress, some of the important topics are veterans' affairs, research and science, manufacturing, and memorials and commemorations, which are areas where we may expect some bipartisan agreement. However, other areas where we might expect more conflict, such as business, children and family law, and election law, disincline members of the minority party to vote "yea" on final passage. There are also seems to be diversity in terms of which committees are significant. For example, in the 111th Congress, the results indicate that, if bills originated from Appropriation, Rules, Ways and Means, and Education and Labor, members of the minority are less likely to support final passage. However, there are also committees such as the Intelligence committee where the connections with minority party voting behavior are less obvious.

In Tables 3, 4, and 5, we took advantage of the lasso to select the top 15 coefficients from the logistic regression that have the highest magnitude.[10] Figures 2, 3, and 4 shows the plot of all

[10] This adds a penalty term to the likelihood function and results are generated with α=1: $\sim\lambda[(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1]$. We also used an elastic net penalty of α = .5; only the lasso is

Electronic copy available at: https://ssrn.com/abstract=3262316

of the standardized coefficients against the parameter lambda, which is the penalty the lasso assesses for variables with non-zero coefficients in the model. The figures illustrate that for reasonable values of lambda, that only a handful of the variables maintain their predictive power, which reinforces the power of bill topics as the driving factors in legislator voting behavior.

The last comparison we make in Table 2 is to examine model performance with different subsets of independent variables. The important observation to make from Table 2 is that the predictive power of the models for each congressional term is driven by topics and committee variables. If we run the models with just those variables, the percent correctly predicted is approximately the same as the full models reported in Table 2. Or alternatively, we can see in Table 2 that if we add the constituency variables or the NOMINATE scores, they do not improve the predictive power of the models.

The Minority Party and Polarization

Finally, we examine the topics that are most likely to incline the minority party to attempt to thwart the majority. As we noted in the introduction, modern terms of the House do not have high levels of "nay" votes, indicating that polarization may not be as prevalent as is commonly thought. Given the strength of our model in predicting votes and the dominant role topics play in this model, we have an opportunity to investigate the substantive areas of legislation that most incline the minority party to either support or try to defeat legislation.

Our approach is to examine party behavior by comparing the marginal effects of each topic by party, which indicates whether the party supports (or not) legislation in that particular

reported here and results are invariant to these choices. See https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html.

Electronic copy available at: https://ssrn.com/abstract=3262316

substantive area. We focus on the ten topics that have the largest influence in our models.[11] As can be seen in Figure 5, when the Republicans are in the minority during the 111th Congress, there are several topics that consistently drive them to vote against the majority including bills dealing with veterans, financial regulations, and the environment. There are also several topics where they are more unified with the majority leadership and vote "yea" including defense, family law, and tariffs.

Figures 6 and 7 show the 112th and 113th Congresses, when Democrats are in the minority. Democrats most often split with the majority on bills focused on family law, financial regulations, sex / crime bills, nutrition, and US territories. They support the majority on bills concerning veterans, foreign aid, election laws, memorials, foreign policy, and homeland security.

The above differences in party support for legislation make good sense based on our understanding of the policy agreements and disagreements between the parties. Democrats and Republicans most often disagree on issue areas involving crime, the environment, finance, and health, and the and they most often agree on issue areas involving defense, foreign policy, and homeland security. There are also two highly salient areas where agreement is conditioned on which party holds the majority: family law and veterans.

Overall, though, there is more agreement than disagreement and this is in contrast to the prevailing narrative both in journalistic accounts as well as much of the research on polarization. Among media sources at the time, for example, there was a widespread conception of

[11] Given that we are using both the base topics and squared terms, the figures in this section present the marginal effects when topics are set to a high value (i.e., equal to when the topics are in the top 10% of their range). Given the power of the naïve model in the case of predicting majority votes, the main emphasis is on the marginal values for the minority party.

Electronic copy available at: https://ssrn.com/abstract=3262316

Republicans as the party of "no" who were willing to go to any length to stop the Democrats (most especially President Obama) from securing political victories.[12] This dynamic was reinforced by the rise of the Tea Party. Likewise, Republican leadership in the house in the 112th and 113th Congresses frequently blamed the minority party and split control of Congress for obstructing its legislative agenda.[13] Yet, in all of the terms studied here, the minority party (whether it is Republican or Democrat) most often voted with the majority party, though with varying support based on the substantive topics of the bills in question (as noted above).

There are, of course, two elephants in the room. The first is that the agenda is endogenous and the absolute amount of legislation decreased dramatically from the 111th to the 113th Congresses. The other is that by focusing on final passage votes, we miss out on important intra-party strife and bargaining, which was particularly prevalent in the GOP controlled terms in our study. These problems are relevant to all studies that rely on roll call votes. But, topic models, of the sort presented here, at least offer the hope that we can more directly study the content of legislation. And that means that it is possible to examine which topics are excluded from consideration – either due to intra-party strife or lack of support by the minority party. We will take up these problems in future work.

Discussion

Minority party roll call voting in the U.S. Congress has received scant attention in the literature, in part because much of the focus has been on the role of the majority party in

[12] There are numerous news stories to this effect; e.g., https://www.economist.com/node/15393693.
[13] See, for example, Roll Call's synopsis: https://www.rollcall.com/news/113th-congress-legacy-shutdowns-shoutdowns-gridlock.

Electronic copy available at: https://ssrn.com/abstract=3262316

controlling the agenda. In this paper we have examined the roll call behavior of the minority party across three congresses, the 111th to the 113th. We sought to the test the predictive power of several sets of variables. We have included traditional variables found in the literature ranging from ideology scores from Poole and Rosenthal's NOMINATE procedure to measures of partisanship, policy preferences, and ideological characteristics of members' constituents.

Building off of recent research, we add measures of bill content based on a structural topic model. We also include measures of the committees that originated the bills. The out-of-sample predictive power of the model with the topics and committees does an excellent job of predicting roll call voting, whereas the constituency and other member characteristics variables does not contribute significantly to the overall predictive power of the model.

There are at least two additional implications that we can draw from this analysis. First, after controlling for the content of the bills being voted on, NOMINATE scores do not meaningfully contribute to the overall predictive power of the models. This calls into question whether or not stories of polarization, based largely on Nominate scores, should be trusted. This is especially important given the historical record of minority party support for roll call votes does not provide evidence that polarization in the modern House is at an apex.

Second, given the diversity and number of topics and committee variables that are significant in predicting roll calls, we believe that the policy space in Congress is highly multidimensional. Monte Carlo simulations and the results presented here show that the best models based on out-of-sample performance require at least 50 dimensions to fit roll call data. Overall, a single policy dimension (or any small number) does not do a good job of explaining roll call voting for minority members on final passage votes. If a single dimension underlies roll call voting, then we should be able to predict minority roll call voting using simple measures and

Electronic copy available at: https://ssrn.com/abstract=3262316

we should see growing resistance to the majority party agenda in Congress. Accordingly, simple stories of divergence in a unidimensional policy space (which underlies most of the polarization literature) are likely misleading.

Thus, it seems traditional measures of legislator ideology, which are then used to measure polarization, obscure the deeper reality of congressional politics (Roberts et al. 2016; Aldrich et al. 2014; Crepsin and Rohde 2010). Congress may only appear to be unidimensional *ex post*, when aggregating across bill types, aggregating the majority party with the minority, and when one includes party-line votes like procedural votes. However, more than likely, the unidimensionality of roll call voting found by others may more likely reflect how well the parties have solved the collective action and social problems they confront (Aldrich 1995; Cox and McCubbins 1993, 2005; Harbridge 2015), as opposed to revealing growing differences policies that can fit neatly into a single, left-right ideological space.

Electronic copy available at: https://ssrn.com/abstract=3262316

References

Adler, E.S., 2002. *Why congressional reforms fail: Reelection and the House committee system*. University of Chicago Press.

Adler, E.S. and Wilkerson, J.D., 2013. *Congress and the politics of problem solving*. Cambridge University Press.

Aldrich, John H. 1995. *Why Parties? The Origin and Transformation of Political Parties in America*. Chicago, IL: The University of Chicago Press.

Aldrich, John H., and David W. Rohde. 2000. "The Republican Revolution and the House Appropriations Committee." *Journal of Politics* 62: 1–33.

Aldrich J. H., Montgomery J. M., Sparks D. B. 2014) Polarization and ideology: Partisan sources of low dimensionality in scaled roll call analyses. *Political Analysis*, 22, 435-456.

Carson, Jamie, Charles Finocchiaro, and David Rohde. 2010. "Consensus, Conflict, and Partisanship in House Decision Making: A Bill-Level Examination of Committee and Floor Behavior." *Congress and the Presidency* 37(3): 231-253.

Clinton, J.D., 2006. Representation in Congress: constituents and roll calls in the 106th House. *Journal of Politics*, *68*(2), pp.397-409.

Cox, Gary W., and Mathew D. McCubbins. 1993. *Legislative Leviathan*. Berkeley, CA: University of California Press.

Cox, Gary W., and Mathew D. McCubbins. 2005. *Setting the Agenda: Responsible Party Government in the U.S. House of Representatives*. New York, NY: Cambridge University Press.

Crespin M., Rohde D. W. (2010). Dimensions, issues, and bills: Appropriations voting on the House floor. *The Journal of Politics*, 72, 976-989.

Crespin, Michael H. and David Rohde. 2016. Political Institutions and Public Choice Roll-Call Database. Retrieved on March 30, 2017 from http://cacexplore.org/pipcvotes/

Deering, Christopher J., and Steven S. Smith. 1997. *Committees in Congress*, 3rd ed. Washington, DC: CQ Press.

Denny, Matthew J., and Arthur Spirling. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* (2018): 1-22.

Dougherty, K.L., Lynch, M.S. and Madonna, A.J., 2014. Partisan agenda control and the dimensionality of congress. *American Politics Research*, *42*(4), pp.600-627.

Egar, W.T., 2016. Tarnishing Opponents, Polarizing Congress: The House Minority Party and the Construction of the Roll-Call Record. *Legislative Studies Quarterly*, *41*(4), pp.935-964.

Ellis, C. and Stimson, J.A., 2012. *Ideology in America*. Cambridge University Press.

Ensley, Michael J., Michael W. Tofias, and Scott De Marchi. "District complexity as an advantage in congressional elections." *American Journal of Political Science* 53.4 (2009): 990-1005.

Evans, D., 2004. *Greasing the wheels: Using pork barrel projects to build majority coalitions in Congress*. Cambridge University Press.

Fenno, Richard F., Jr. 1973. *Congressmen in Committees*. Boston, MA: Little, Brown.

Harbridge, L., 2015. *Is Bipartisanship Dead?: Policy Agreement and Agenda-setting in the House of Representatives*. Cambridge University Press.

Hurwitz M. S., Moiles R. J., Rohde D. W. (2001). Distributive and partisan issues in agriculture policy in the 104th House. *American Political Science Review*, 95, 911-922.

Electronic copy available at: https://ssrn.com/abstract=3262316

Jochim, A.E. and Jones, B.D., 2013. Issue politics in a polarized Congress. *Political Research Quarterly*, *66*(2), pp.352-369.
Krehbiel, Keith. 1991. *Information and Legislative Organization*. Ann Arbor: University of Michigan Press.
Krehbiel, Keith. 1998. *Pivotal Politics: A Theory of U.S. Lawmaking*. Chicago, IL: University of Chicago Press.
Lee, Frances E. 2009. *Beyond Ideology: Politics, Principles, and Partisanship in the U.S. Senate*. Chicago, IL: The University of Chicago Press.
Maltzman, Forrest. 1995. "Meeting Competing Demands: Committee Performance in the Postreform House." *American Journal of Political Science* 39: 653–82.
Miller W. E., Stokes D. E. (1963). Constituency influence in Congress. *American Political Science Review*, 57, 45-56.
Poole, Keith T. 2007. "Changing Minds? Not in Congress!" *Public Choice* 131:435–451.
Poole, Keith T. and Rosenthal, Howard. 1991. ''Patterns of Congressional Voting.'' *American Journal of Political Science* 35: 228–78.
Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.
Poole, Keith T., and Howard Rosenthal. 2007. *Ideology and Congress*. New Brunswick, NJ: Transaction Publishers.
Roberts, Jason M. 2007. "The Statistical Analysis of Roll Call Data: A Cautionary Tale." *Legislative Studies Quarterly* 32: 341–60.
Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. "Navigating the local modes of big data: The case of topic models." (2014).
Rohde, David W. 1991. *Parties and Leaders in the Postreform House*. Chicago, IL: University of Chicago Press.
Rohde, David W. 1994. "Parties and Committees in the House: Member Motivations, Issues, and Institutional Arrangements." *Legislative Studies Quarterly* 19: 341–60.
Rothenberg, Lawrence S., and Mitchell S. Sanders. 2000. "Severing the Electoral Connection: Shirking in the Contemporary Congress." *American Journal of Political Science* 44: 316–25.
Shepsle, Kenneth A. "Institutional arrangements and equilibrium in multidimensional voting models." *American Journal of Political Science* (1979): 27-59.
Shepsle, K.A. and Weingast, B.R., 1981. Structure-induced equilibrium and legislative choice. *Public choice*, *37*(3), pp.503-519.
Shepsle, K.A. and Weingast, B.R., 1984. Uncovered sets and sophisticated voting outcomes with implications for agenda institutions. *American Journal of Political Science*, pp.49-74.
Smith, Steven S. 2007. *Party Influence in Congress*. New York, NY: Cambridge University Press.

Electronic copy available at: https://ssrn.com/abstract=3262316

Figure 1: Final Passage Roll Call Votes by Party Status

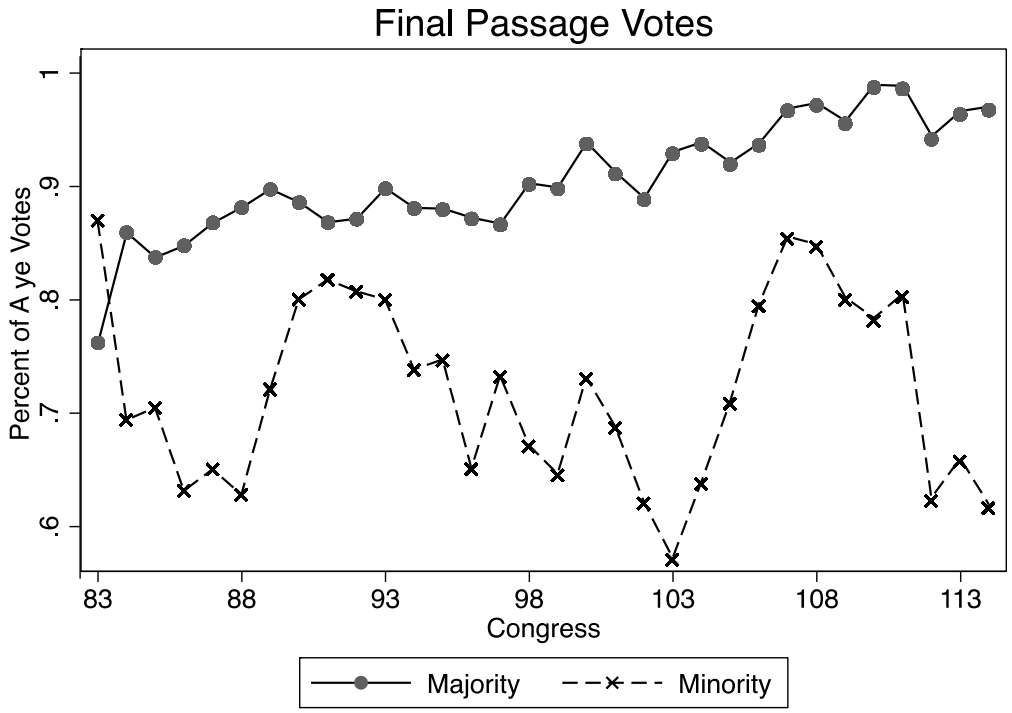
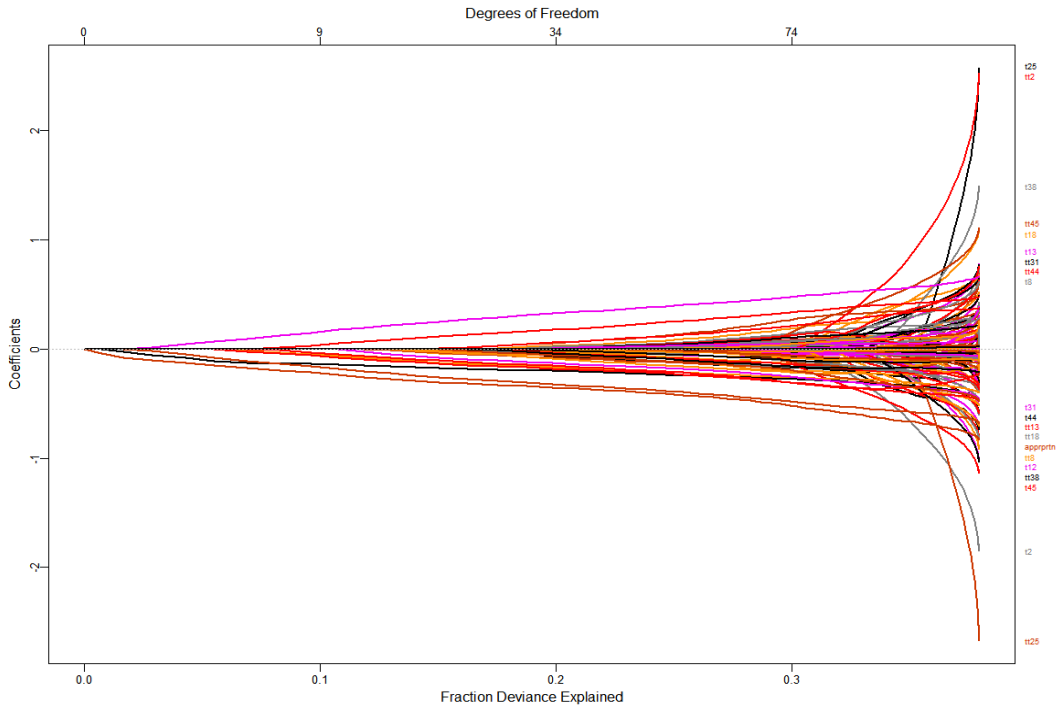


Figure 2: 111th Congress



Note: in all graphs that follow, t# indicates a topic variable and tt# indicates the square of the topic variable. See Table 1 for variable ID's.

Electronic copy available at: https://ssrn.com/abstract=3262316

Figure 3: 112th Congress

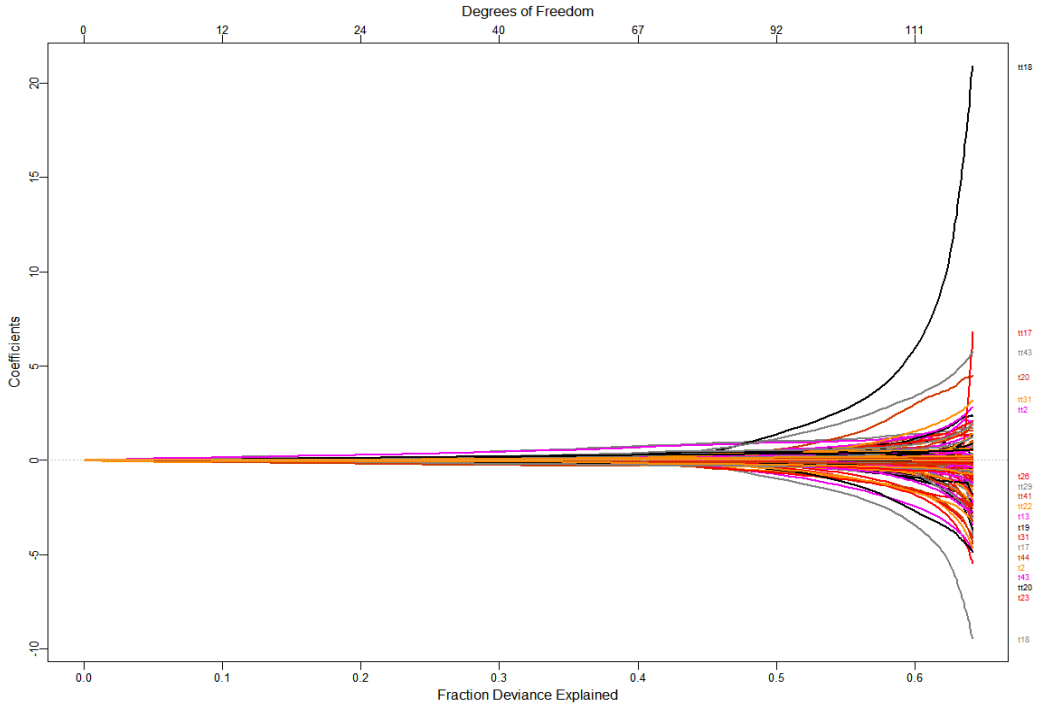
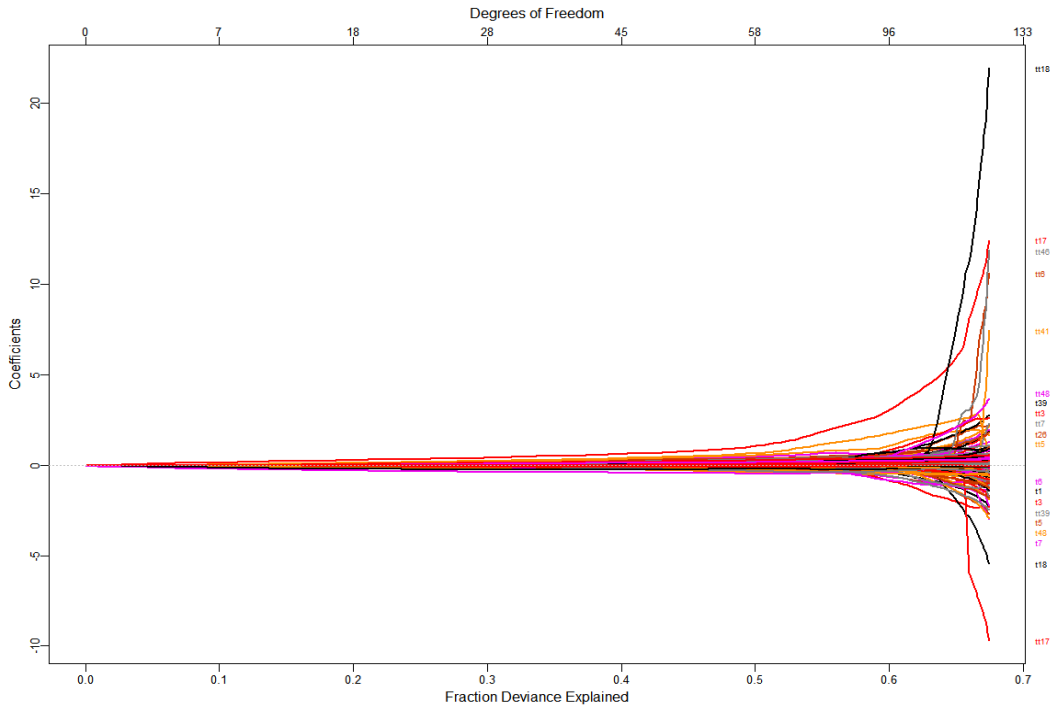


Figure 4: 113th Congress



Electronic copy available at: https://ssrn.com/abstract=3262316

Figure 5: Support for Legislation, 111th Congress (Democrat majority)

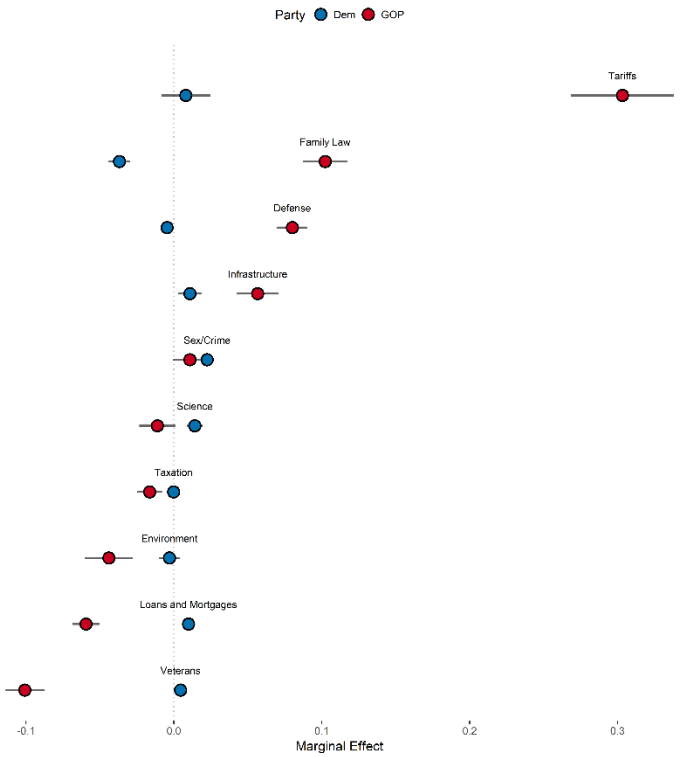
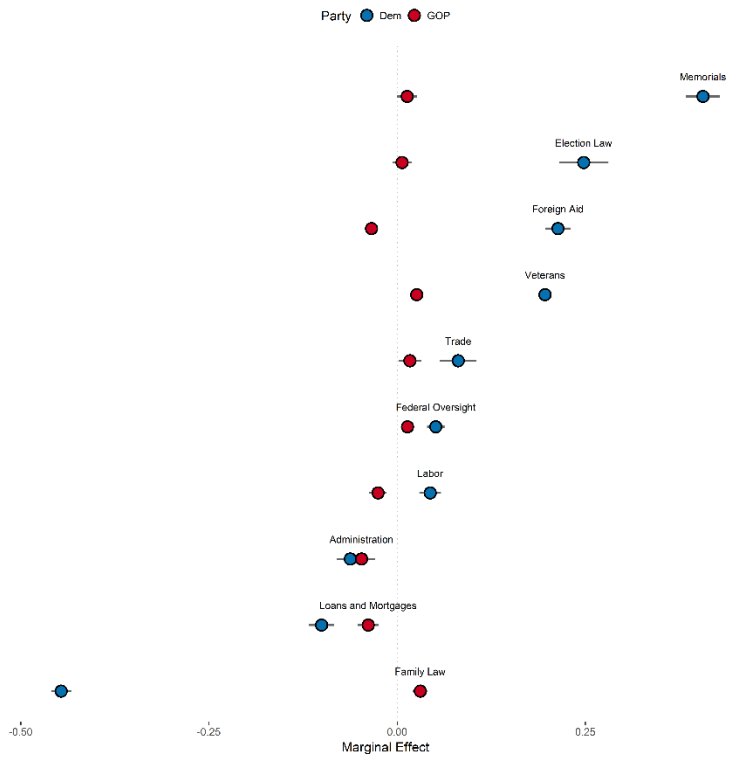
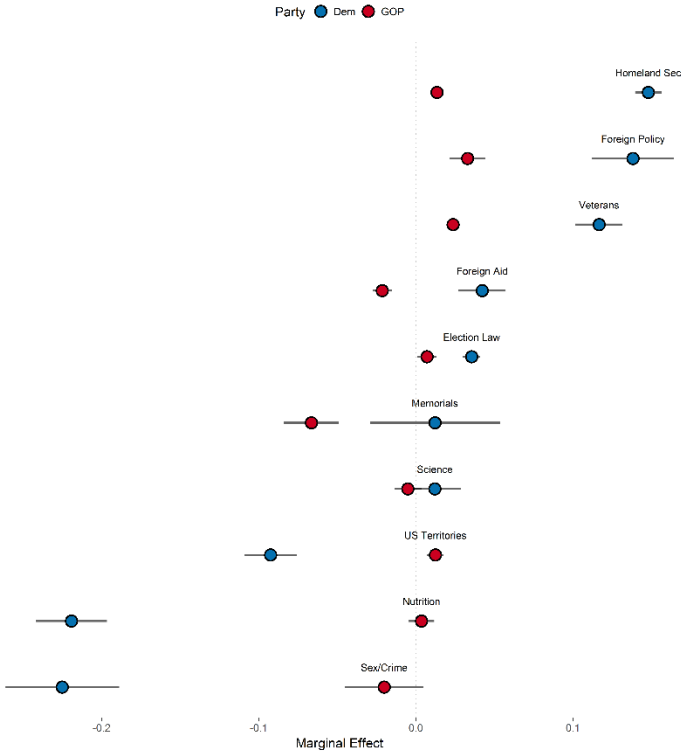


Figure 6: Support for Legislation, 112th Congress (Republican majority)



Electronic copy available at: https://ssrn.com/abstract=3262316

Figure 7: Support for Legislation, 113th Congress (Republican majority)



Electronic copy available at: https://ssrn.com/abstract=3262316

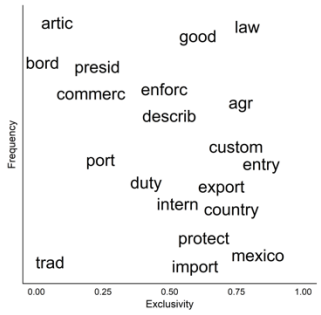
Table 1a: Top 50 Topics from the STM

Var Name	Topic Substance	Var Name	Topic Substance
t 0	North American trade	t 25	Tariffs
t 1	Medicine	t 26	Homeland security
t 2	Lawsuits	t 27	Defense
t 3	Conservation and Federal lands	t 28	Taxation
t 4	Taxes	t 29	Veteran affairs
t 5	Immigration	t 30	Energy
t 6	Territories	t 31	Federal lands
t 7	Diet and nutrition	t 32	Environment
t 8	Defense	t 33	Energy and fuel efficiency
t 9	Bill terminology	t 34	Manufacturing
t 10	Pharmaceuticals	t 35	Loans and mortgages
t 11	Administration	t 36	Disasters
t 12	Bill section terminology	t 37	Unemployment
t 13	Budget	t 38	Administration
t 14	Education	t 39	Labor
t 15	Agriculture	t 40	Foreign aid
t 16	Business	t 41	Memorials and commemorations
t 17	Election law	t 42	Banking
t 18	Children and family law	t 43	Research and Science
t 19	Climate and environment	t 44	Administration
t 20	Legal	t 45	Names
t 21	Medicare and Medicaid	t 46	Criminal law and sex crimes
t 22	Internet	t 47	Federal oversight
t 23	Foreign policy and sanctions	t 48	Infrastructure
t 24	Investment	t 49	Tax law

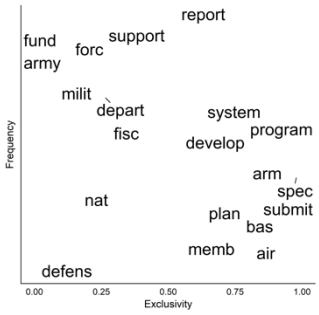
Electronic copy available at: https://ssrn.com/abstract=3262316

Table 1b: Top Nine Topics, Frequency and Exclusivity of Terms

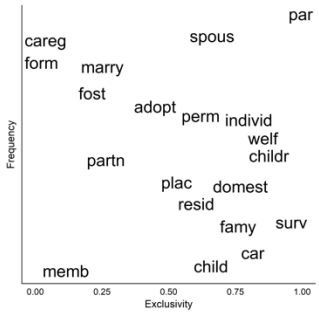
Trade



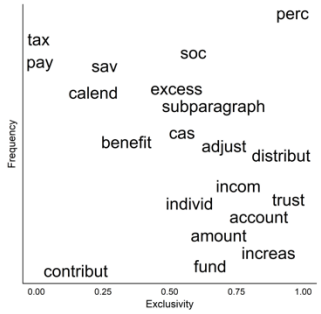
Defense



Election Law



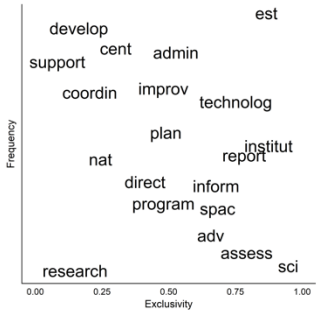
Taxation



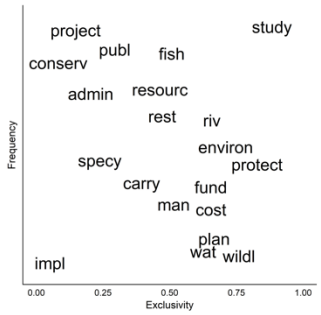
Foreign Aid



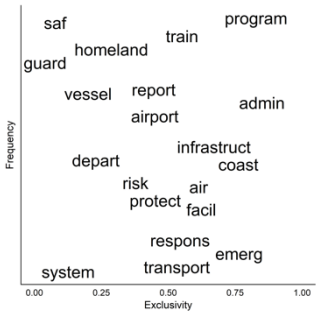
Science



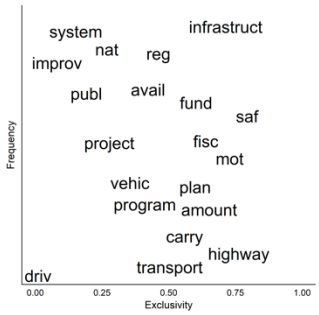
Environment



Homeland Security



Infrastructure



Electronic copy available at: https://ssrn.com/abstract=3262316

Table 2: Out-of-sample performance[14]

111th Congress	Minority	Majority
Full Topics + Covariates	0.857	0.977
Topics only	0.701	0.973
Nominate 2d only	0.661	0.973
Baseline, Pr(yea) = 1	0.666	0.973

112th Congress	Minority	Majority
Full Topics + Covariates	0.93	0.955
Topics only	0.776	0.938
Nominate 2d only	0.63	0.935
Baseline, Pr(yea) = 1	0.623	0.935

113th Congress	Minority	Majority
Full Topics + Covariates	0.949	0.971
Topics only	0.857	0.963
Nominate 2d only	0.649	0.963
Baseline, Pr(yea) = 1	0.657	0.963

[14] Note: Gerrish and Blei's models of roll call votes using LDA to produce bill topics achieve an overall performance of 89.7% in modern terms.

Electronic copy available at: https://ssrn.com/abstract=3262316

Table 3: Top 15 predictors using lasso, 111th Congress

ID	Variable	coef	FREX terms / Description
52	tt2	1.95	claim, court, dam, brought, lawsuit, discrimin, liabl, alleg, complaint, party, disput, civil, remedy, injury, settl, punit, injunct, assert, decid, plaintiff
75	tt25	-1.89	subhead, 15th, amino, hydroxy, methyl, harmon, sequ, leath, footwear, fila, knit, acryl, resin, wareh, chloro, polym, bis, column, withdrawn, suspend
25	t25	1.76	subhead, 15th, amino, hydroxy, methyl, harmon, sequ, leath, footwear, fila, knit, acryl, resin, wareh, chloro, polym, bis, column, withdrawn, suspend
3	t2	-1.45	claim, court, dam, brought, lawsuit, discrimin, liabl, alleg, complaint, party, disput, civil, remedy, injury, settl, punit, injunct, assert, decid, plaintiff
38	t38	1.13	secret, fed, requir, term, determin, auth, appropry, paragraph, box, rel, submit, approv, review, publ, licens, respect, respons, lat, request, reg
45	t45	-0.95	hast, florid, mrs, texa, brown, blackburn, johnson, bishop, westmoreland, hinchey, broun, connol, gingrey, conaway, sanchez, bachman, lamborn, brady, latt, posey
95	tt45	0.93	hast, florid, mrs, texa, brown, blackburn, johnson, bishop, westmoreland, hinchey, broun, connol, gingrey, conaway, sanchez, bachman, lamborn, brady, latt, posey
18	t18	0.90	par, child, spous, marry, childr, fath, widow, fost, famy, welf, daught, perm, surv, bir, partn, son, adopt, domest, vis, immigr
88	tt38	-0.81	secret, fed, requir, term, determin, auth, appropry, paragraph, box, rel, submit, approv, review, publ, licens, respect, respons, lat, request, reg
12	t12	-0.78	serv, year, end, paragraph, term, determin, requir, rel, cod, respect, appropry, auth, fed, nonpost, fisc, secret, part, begin, tot, defin
100	appropriations	-0.78	committee
58	tt8	-0.71	unspec, defens, navy, missil, command, army, fort, milit, tact, afb, worldwid, aircraft, comb, def, subtot, afgh, weapon, forc, air, nuclear
68	tt18	-0.66	par, child, spous, marry, childr, fath, widow, fost, famy, welf, daught, perm, surv, bir, partn, son, adopt, domest, vis, immigr
63	tt13	-0.64	budget, spend, resolv, deficit, joint, legisl, outlay, bienn, fisc, estim, conf, discret, deb, mot, sequest, pass, bien, bal, concur, omb
130	dwnom1	-0.64	1st Dimension Nominate

Note: in all tables that follow, t# indicates a topic variable and tt# indicates the square of the topic variable. See Table 1 for variable ID's.

Table 4: Top 15 predictors using lasso, 112th Congress

ID	Variable	coef	FREX terms
68	tt18	17.44	par, child, spous, marry, childr, fath, widow, fost, famy, welf, daught, perm, surv, bir, partn, son, adopt, domest, vis, immigr
18	t18	-8.07	par, child, spous, marry, childr, fath, widow, fost, famy, welf, daught, perm, surv, bir, partn, son, adopt, domest, vis, immigr
93	tt43	5.30	sci, fishery, research, technolog, spac, innov, collab, council, strategic, atmosph, coordin, consort, marin, engin, goal, adv, fish, develop, strategy, stakehold
23	t23	-4.50	syr, kore, sanct, iae, israel, religy, chin, russ, freedom, taiw, palestin, sud, cub, vietnam, ter, egypt, ukrain, prol, republ, arab
70	tt20	-4.48	firearm, whoev, imprison, offens, know, forfeit, pun, gambl, fals, licens, attempt, interst, unlaw, gun, possess, interceiv, ammunit, explod, handgun, wir
20	t20	4.37	firearm, whoev, imprison, offens, know, forfeit, pun, gambl, fals, licens, attempt, interst, unlaw, gun, possess, interceiv, ammunit, explod, handgun, wir
43	t43	-4.26	sci, fishery, research, technolog, spac, innov, collab, council, strategic, atmosph, coordin, consort, marin, engin, goal, adv, fish, develop, strategy, stakehold
3	t2	-3.92	claim, court, dam, brought, lawsuit, discrimin, liabl, alleg, complaint, party, disput, civil, remedy, injury, settl, punit, injunct, assert, decid, plaintiff
31	t31	-3.50	wild, creek, scen, approxim, depict, herit, fork, forest, mountain, headw, conflu, map, trail, canyon, acr, idaho, seg, riv, salmon, bound
44	t44	-3.49	expend, expens, remain, avail, fund, septemb, oblig, non, unoblig, necess, transfer, excess, repair, head, reprogram, phs, herein, reimburs, proviso, hir
72	tt22	-2.89	internet, electron, disclos, inform, websit, record, network, telephon, access, databas, onlin, televid, broadband, search, commun, transmit, digit, nam, not, mobl
81	tt31	2.84	wild, creek, scen, approxim, depict, herit, fork, forest, mountain, headw, conflu, map, trail, canyon, acr, idaho, seg, riv, salmon, bound
19	t19	-2.83	emit, auct, greenh, clim, spectr, carbon, reb, offset, wast, mercury, dioxid, build, retrofit, vint, gas, combust, cle, efficy, lamp, allow
13	t13	-2.58	budget, spend, resolv, deficit, joint, legisl, outlay, bienn, fisc, estim, conf, discret, deb, mot, sequest, pass, bien, bal, concur, omb
96	tt46	-2.58	victim, sex, stalk, assault, prison, juvenil, gang, crim, offend, crimin, traffick, inm, abus, incarc, just, viol, attorney, prosecut, harass, rap

Table 5: Top 15 predictors using lasso, 113th Congress

ID	Variable	coef	FREX terms
67	tt18	9.12	par, child, spous, marry, childr, fath, widow, fost, famy, welf, daught, perm, surv, bir, partn, son, adopt, domest, vis, immigr
17	t17	6.36	candid, ballot, vot, post, campaign, redistrict, mail, lobby, elect, pap, abs, polit, cast, pol, vic, ref, begun, count, postmast, hundr
95	tt46	2.81	victim, sex, stalk, assault, prison, juvenil, gang, crim, offend, crimin, traffick, inm, abus, incarc, just, viol, attorney, prosecut, harass, rap
41	t41	2.55	coin, mint, gold, commem, muse, mem, concess, battlefield, hist, athlet, struck, smithson, silv, surcharg, landmark, park, world, hon, centen, inscrib
18	t18	-2.32	par, child, spous, marry, childr, fath, widow, fost, famy, welf, daught, perm, surv, bir, partn, son, adopt, domest, vis, immigr
55	tt6	2.29	island, columb, commonweal, district, puerto, rico, marian, samo, guam, northern, virgin, circuit, marijuan, judg, northwest, thent, maryland, territ, decen, cens
97	tt48	2.26	highway, rail, freight, transport, metropolit, driv, apport, traff, vehic, saf, mot, passeng, motorcoach, infrastruct, bridg, safete, surfac, corrid, bus, project
52	tt3	2.14	convey, trib, parcel, haway, ind, navajo, land, county, alask, pueblo, compact, recr, intery, hunt, monu, sealask, nevad, band, appra, vil
4	t3	-2.09	convey, trib, parcel, haway, ind, navajo, land, county, alask, pueblo, compact, recr, intery, hunt, monu, sealask, nevad, band, appra, vil
29	t29	1.89	vet, affair, arm, dis, task, forc, servicememb, cemetery, fun, milit, dog, memb, traum, wound, deceas, depart, comb, tric, uniform, bur
39	t39	1.82	grant, workforc, train, you, commun, recipy, elig, program, award, org, demonst, adult, ent, skil, assist, job, loc, techn, coordin, allot
26	t26	1.75	vessel, airport, cybersec, avy, flight, homeland, threat, aircraft, maritim, alert, detect, guard, cyb, incid, fir, emerg, respond, coast, vuln, air
6	t5	-1.65	aly, nonimmigr, immigr, petit, detain, det, admit, homeland, ver, vis, remov, visa, deport, passport, work, custody, asyl, apprehend, refug, recruit
88	tt39	-1.61	grant, workforc, train, you, commun, recipy, elig, program, award, org, demonst, adult, ent, skil, assist, job, loc, techn, coordin, allot
48	t48	-1.59	highway, rail, freight, transport, metropolit, driv, apport, traff, vehic, saf, mot, passeng, motorcoach, infrastruct, bridg, safete, surfac, corrid, bus, project

Electronic copy available at: https://ssrn.com/abstract=3262316

Table 6: Net effects of top 10 topics by Congress

111th: Topics	effect
Children and family law	+
Taxes	+
Education	+
Medicare and Medicaid	+
Administration	+
Infrastructure	+
Medicine	+
Internet	+
Conservation and Federal lands	+
Election law	+

112th: Topics	effect
Election law	+
Memorials and commemorations	+
Veteran affairs	+
Foreign policy and sanctions	+
Administration	-
Federal oversight	+
Medicare and Medicaid	-
Climate and environment	-
Names	-
Disasters	+

113th: Topics	effect
Election law	+
Energy	-
Memorials and commemorations	+
Veteran affairs	+
Lawsuits	-
Homeland security	+
Foreign policy and sanctions	+
Budget	-
Medicare and Medicaid	-
Taxes	-

Electronic copy available at: https://ssrn.com/abstract=3262316

Online Appendix

This appendix describes the process and various decisions that were made to derive the topics from the content of bills. For a good overview of many of these topics, see Denny and Spirling (2018).

Modeling Issues

We want to be confident that we have a reasonable representation of the data generating process (DGP) and that our observations are independent and identically distributed (IID). In the case of modeling the Congress, there are several issues to consider.

First, one must consider what the appropriate sample is. The dependent variable is roll call votes in the House – which by definition are either "yea" or "nay"[15] – but not all votes are created equal. In fact, a large number of votes are procedural or amendments rather than votes that determine the passage of a piece of legislation. It is difficult to argue that these types of votes spring from the same DGP; worse, including procedural votes swamps the more salient votes that determine passage. While there are examples of salient procedural and amendment votes, there is no reason to lump different types of votes together. Accordingly, our sample consists only of final passage votes. In the 111th House there are 297 final passage votes, in the 112th there are 274, and for the 113th there are 293; together, this produces over 300,000 roll call votes.

Second, there is the issue of whether or not one should combine votes by members of the different political parties into one sample (controlling for party) or to treat them as two different

[15] There can also be a value of "not voting". One thus has the two main options of either treating these observations as missing or as a "nay" vote – we settle on the latter for the work presented here but results are robust to coding the values as missing instead.

Electronic copy available at: https://ssrn.com/abstract=3262316

samples. Given our restriction to final passage votes, there is an obvious temptation to increase the sample size by including both parties together. But, this flies in the face of what we know about politics. Parties in power have different considerations than opposition parties and centrist parties tend to have more complex ideologies than extreme parties (Tetlock 1984). Further, different topics may hold very different meanings depending on the party -- accordingly, the models here disaggregate by party.

Third, it is not obvious that one should include different terms of Congress in the same sample if one wants minimally IID observations. Agendas exist and they vary over time (and not in some nicely continuous fashion). Votes are always about something and there is no reason to believe that these votes are drawn from the same DGP in different terms, especially if there is a change in party control of the agenda.[16] Indeed, there is every reason to think the agenda matters:

> "It is something of an open question whether any session of Congress would prove to be typical. All would have entirely usual aspects and some very unusual ones. As one congressman remarked, 'There hasn't been a year since I've been here that has been a usual year.'" (Kingdon, p. 15).

For the terms considered here, there is also abundant evidence that strategies and the substance of legislation changed markedly from term to term. Republican Speaker John Boehner's attempts to enforce the Hastert rule was in contrast to the strategy adopted by his Democratic predecessor, Nancy Pelosi. The solution to this problem is to simply fit each term of Congress separately. This is the also the approach taken by Gerrish and Blei. Treating terms independently substantially improves results.

[16] There are other ways to deal with time – see Wang, et al. (2013) for an alternative approach. Note also that there is a distinction between the regression stage in our modeling process (which should be disaggregated by term) and generating the topic model (which should not be disaggregated given the data requirements of LDA / STM).

Electronic copy available at: https://ssrn.com/abstract=3262316

Why use topics as features?

Using text as data has become much more important in the social sciences.[17] Our approach depends on extracting topics from the full text of bills before Congress. But, it is important to discuss why feature extraction of this kind is important in the context of predicting votes. After all, if one wants to intervene as little as possible between the data and the model, one could instead build a statistical model that uses word counts from each bill directly. While the total dictionary size for a modern term of Congress is nearly 16,000 unique words (thereby producing an equal number of IVs), regularization approaches like the lasso could trim this space substantially.

Unfortunately, regularized regression models fail in the context of predicting votes in Congress for three reasons. First and unsurprisingly, regularized regression models using the complete dictionary of words as IVs underperform regressions using topic models by a sizable amount. Second, a high degree of correlation exists between words within the same bill and no solution such as n-gramming entirely eliminates this problem. In general, the correlation between features blunts the ability of high-dimensional regression methods such as the elastic net penalty or random forests to successfully capture the systematic component in the data (Conn and Ramirez 2016). Last, while we have a large number of observations due to 435 members of Congress voting on the passage of several hundred bills, this is not as plentiful a source of data as it seems. Bill characteristics (whether words or topics) are fixed for each member of Congress and legislator characteristics are fixed across bills; this is close to a situation where there are

[17] A partial list of research using text as data is Laver, et al. (2003), Lauderdale and Clark (2012); Grimmer and Stewart (2013), Roberts, et al. (2014), and Kim, et al. (2014).

Electronic copy available at: https://ssrn.com/abstract=3262316

more columns than rows. The bottom line is this: our "best" models in terms of out-of-sample accuracy depend on aggressively selecting a parsimonious feature space (Bishop 1995). Feature extraction based on topic models are essential if one wants accurate and stable results of how legislators' vote and they have the welcome side-effect of allowing us to more easily interpret the substantive implications of our models.

Regularization and Raw Dictionary Data

To see why topic models (LDA or STM) perform so well, we consider one alternative which is to instead use the raw word counts in each bill directly. Accordingly, we created a dataset based on the complete dictionary of each term of Congress, resulting in nearly 16,000 IVs (one column per token). To push coefficients which contributed little to the predictive power of our model to zero, we relied on regularized regression approaches like the lasso.[18] The lasso is a simple amendment to a regression model that penalizes model complexity. In addition to minimizing the residual sum of squares, an extra term is added to the minimization that is the sum of the absolute value of the coefficients (Tibshirani 1996). One issue, however, immediately presents itself with using word counts directly. Regularized regression produced stable models with this dataset when we generated out-of-sample results for an entire term using k-fold cross-validation. But, for the genuinely predictive task of forecasting votes in the 113th Congress as each bill occurs, there are too few rows (a maximum of 435) available.

[18] We also used the ridge penalty and an elastic-net penalty of 0.5. None of these models performed as well as topics-based regression models.

Electronic copy available at: https://ssrn.com/abstract=3262316

For both the 111th and 112th terms, regularized regression models produced accuracies just over 0.8.[19] This is substantially worse than regressions based on topic models and reinforces the need for pre-processing and feature extraction in problems of this kind.

Parameters for Topic Models

Building topic models from the text of bills is a complicated process given the diversity and frequency of idiosyncratic and technical terms in legislation. Following Gerrish and Blei (2011), we used TreeTagger (Schmid 1994) to stem and lemmatize bills, called from the koRpus package in R.[20] Unlike Gerrish and Blei, we did, however, improve the results by hand-coding unknown tokens that represent specialist vocabulary commonly found in bills. For example, the word "Medicaid" occurred 3077 times in the 111th House alone; "cybersecurity" appeared 556 times. The resultant topics were substantially improved by including these terms.[21]

Our corpus is the set of all bills before each term of Congress. Bills consist of a vector of words **w** of length N, where each word is the result of pre-processing. It is worth noting that LDA and STM employ a bag of words (exchangeability) assumption; the order of the words in **w** is unimportant in the statistical model.

[19] Random forest models did better, but did not exceed 0.88.
[20] http://r.reaktanz.de/pckg/koRpus/koRpus.pdf
[21] TreeTagger provided us with lemmas which we used to build the dictionary for bills before Congress – this helped reduce the size of the dictionary. It also provided parts of speech, which we did not use in our model. TreeTagger's default dictionary lacked technical terms, salient proper names, and modern words though which were commonly found in Congressional bills. Additional examples include "Cherokee", "subprime", "Dodd-Frank" (after forming bigrams), and "permittee". In these cases, TreeTagger marked the word as unknown and we manually added in the proper lemma afterwards. There is a large literature on improving lemmatizing for technical contexts, but the bottom line is that one should probably not use "off-the-shelf" tools without modification (Liu, et al. 2012).

Electronic copy available at: https://ssrn.com/abstract=3262316

Given this, LDA estimates a number of topics based on the corpus; this number is a parameter chosen by the modeler and here is set at 100. Each bill d is expressed as a vector θ_d consisting of a proportion for each of the one hundred topics that could make up a bill.

Following Blei, et al. (2003), first choose a number of words N in a bill:

$$N \sim Poisson(\xi).$$

Second, the proportions of topics are drawn from a common prior with shape parameters α:

$$\theta \sim Dirichlet(\alpha^{\theta}).$$

Last, for each of N words in the dictionary built from the bills in a term, we iterate the process of choosing a topic and then within the topic the word that is represented. For each word w_n from a bill d, a topic z is drawn as $z \sim Multinomial(\theta_d)$; words are drawn from $p(w_n|z,\beta)$, a multinomial distribution of word probabilities for that topic**.** Topics are nearly independent (since they sum to unity) in LDA and the number of topics is fixed prior to estimating the model.

The parameter space implied by this model is huge – each of the N topics used is a vector of weights over the entire dictionary of words, which even after aggressive pre-processing has between 5 thousand and 10 thousand entries for any given term of Congress. Unsurprisingly, multiple local optima exist for the likelihood function implied by this model. Thus, to the extent that researchers are presenting word clouds or lists of the most salient words by topic, often these results are non-unique and as such should be taken with a grain of salt.

The solution we employ, outlined by Roberts, Stewart, Tingley, and Airoldi (2014) extends topic models in two important ways. First, the STM model allows for spectral learning, which provides outcomes consistent with the global optimum. This requires an additional assumption of separability (identified by Arora, et al. 2013), where a unique anchor word is identified for each topic in θ.

Electronic copy available at: https://ssrn.com/abstract=3262316

Second, the STM model allows for covariates at two levels: topic prevalence and the content of topics. If, for example, multiple terms of Congress are aggregated into one dataset and party control of the agenda changes, it is necessary to consider that a Republican authored bill using the word "women" might be entirely different in meaning than a Democrat authored bill using "women". Though we do not focus on results generated by aggregating terms here, we relied on STM's ability to use a covariate for topic prevalence – in this case, party control – to deal with this issue.[22] For topic prevalence covariates **x**,

$\theta_d \sim \text{LogisticNormal}(\Gamma' \mathbf{x}'_d, \text{topics})$.

If no covariates are used, STM is identical to the correlated topic model (Blei and Lafferty 2007), which relaxes LDA's assumption of independent topics but is otherwise the same. STM is a meaningful generalization of topic models that allows known attributes of each bill to influence topics. For these reasons, we use the STM package in R to estimate topics. For redundancy, we also generated results with LDA implemented by the python **gensim** library.[23]

Another concern with topic models is how one should choose parameters, including the number of topics and the lower and upper frequency threshold for including tokens in the model. We ran simulations using two training sets, one from the 111th term and another from the 112th. We evaluated performance for the following combinations:

Number of topics $\in \{10, 25, 50, 100\}$

Lower threshold (count) $\in \{5, 10, 25, 50\}$

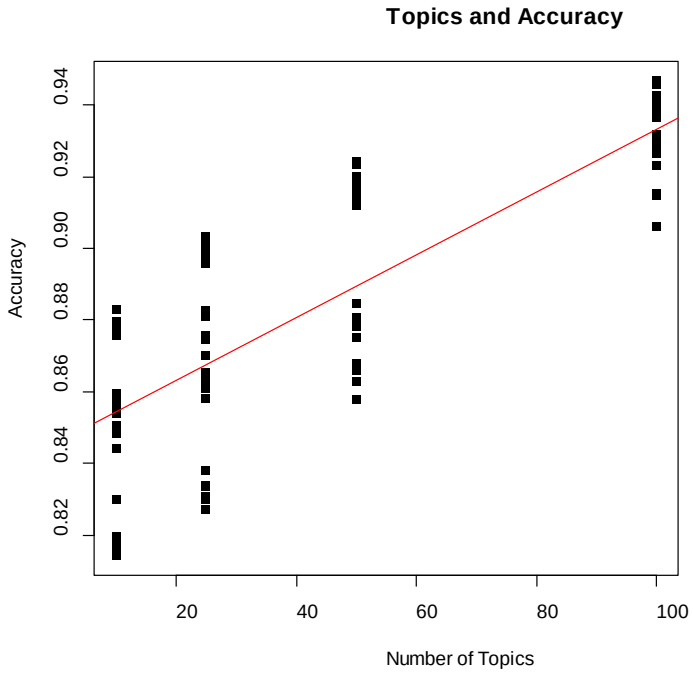
[22] Aggregating terms produces less good results during periods when party control changes. For a simple comparison with the results here, our model that aggregated the 111th and 112th terms of the House had an out-of-sample accuracy rate of ~90% using the STM model. Using the simpler LDA model did less well -- ~88%.

[23] In the case of LDA, we found that multiple models of single terms provided the same fit to the data albeit with very different topic – word mappings. The overall fit of the regressions predicting roll call votes were the same with LDA when we disaggregated by Congress. STM does, however, have a significant advantage when it comes to replicability.

Electronic copy available at: https://ssrn.com/abstract=3262316

Upper threshold (frequency) $\in$ {.2, .3, .5}

This produces a total of 96 possible topic models and we evaluated each in terms of out-of-sample accuracy. Results in both terms were nearly identical. Weakly, the higher the lower count threshold the better; the higher frequency threshold had no effect. The number of topics, though, mattered a great deal and overall, more topics produced higher accuracy (as represented in the figure below):



We settled on 100 topics, a lower count threshold of 25, and an upper frequency threshold of .5 for the models presented in the main text in the out-of-sample tests using the 111th and 112th terms. For the predictive model using the 113th term, we used 25 topics instead which was necessary for convergence.

Electronic copy available at: https://ssrn.com/abstract=3262316

Finally, we experimented with multiple different pre-processing schemes for the corpus. In general, the only choice that had a substantial impact on our models was replacing “unknown” tokens (e.g., we would normally exclude “Dodd-Frank” as proper names, but the usage in this case demanded inclusion in the dictionary) by hand. We also found that bigrams gave the biggest bang for the buck in terms of the tradeoff between accuracy and increasing the size of the parameter space.

Electronic copy available at: https://ssrn.com/abstract=3262316