Predicting Bill Votes in the House of Representatives

Tom Henighan* and Scott Kravitz†
Physics Department, Stanford University, Stanford, California, USA
(Dated: December 12, 2015)

We develop a generic model of voting behavior in the House of Representatives (election cycles 2004-2014) without the use of individual vote histories, allowing for generalization to future Congresses with new members. We find that these Representatives nearly always vote in line with their political party's collective decision, and that greater than 95% prediction accuracy is possible when the minority party's collective vote can be perfectly determined. Using logistic regression with party information for bill sponsors, cosponsors, and voters, and further information about how controversial a bill is from a list of topics relevant to it achieves a prediction accuracy of 88%, comparable to state-of-the-art methods. More complex models which try to account for party-specific preferences by topic showed no improvement.

INTRODUCTION

Voting on bills in the House of Representatives is one of the primary steps required for creation of a new federal law, but what influences an individual Representative's vote is not always transparent. High-profile bills such as the Patient Protection and Affordable Care Act [1] highlight partisanship in the voting process, suggesting that political party is a primary driver of voting behavior, particularly in recent years. Moreover, movements such as "Occupy Wall Street" [2] have brought the idea of a wealthy few buying a disproportionate amount of influence into the minds of Americans, raising the question of how influential election campaign contributions and lobbying efforts are in affecting a Representative's votes. This work attempts to both accurately predict individual votes in the House of Representatives and elucidate the primary influences on voting behavior. In addition, it does so in a way that is easy to programmatically evaluate and to generalize to a future Congress with new members who may not have a voting history to train on.

PRIOR WORK

Much of the past work in this area involves training a separate classifier for each Representative, and hence requires voting history information, making it difficult to generalize to future Congresses. Some attempts have been made to determine whether a bill will survive being referred to a congressional committee, using both unsupervised clustering of bills by textual content [3] and logistic regression on features like sponsor party, committee chairman party, and state of origin [4]. There is ample evidence that a Representative's party is relevant for predicting voting behavior [5, 6], as one might expect. The evidence for campaign finance data being relevant is mixed, with greater influence seen for votes on special interest issues, where contributions from interest groups are most likely to be salient and accountability to constituents may be lower [7]. There is more evidence that lobbying efforts can affect voting results [8]. However, gathering lobbying information can be costly (such as through directly reaching out to Representatives), limited (only the number of lobbying appeals for a given bill is public), and unreliable (some lobbying is done off-the-books), making such data outside the scope of this work.

A meta-analysis of older studies suggests that the best models reach roughly 90% accuracy in predicting individual votes, though all methods reviewed rely on using individual past voting records [6]. A more recent model takes a baseline ideology score for each Representative (on the conservative-liberal spectrum) and adjusts it according to past voting record on topics relevant to the bill, as determined by a programmatic study of the most common associated keywords (assigned to bills by an outside group) [9].

This work is unique (to our knowledge) in that it does not use the past voting records of Representatives, instead creating a single generic model which can be applied to any future Representative, and hence which can be used to glean information about general voting behavior patterns. State of the art models seem to cap out at roughly 90% prediction accuracy for a wide class of models, even with voting records, so we hope to achieve roughly similar performance.

DATASET CHARACTERIZATION

This work focuses on votes in the House of Representatives for Congresses 108-113 (election cycles 2004-2014). Congresses 108-112 were used as training data, while Congress 113 was set aside for testing. Each example consists of a bill-Representative pair, with the Representative's vote on that bill as the result we wish to predict. We restricted the data to roll call votes, the results of which are publicly available. Further, votes were only considered if they concerned the passage of House Bills (which, if also passed in the Senate, would become law and are hence of most interest to the general public). If

there were multiple votes for the passage of one bill, we considered only the first such vote. For simplicity, we excluded abstentions and votes by members of a third party, so that both the vote result and each Representative's political party are binary.

Several different input feature lists were considered, and can be separated into features associated with Representatives and with bills. For Representatives, the features considered were the Representative's political party and campaign contributions, obtained from opensecrets.org as bulk text files containing a list of all contributions from both political action committees (PACs) and individuals for each election cycle. Each contribution is assigned to a "sector" identifying the donor's industry or ideology. More information on all thirteen sectors (such as Agriculture, Health, and Finance) can be found on the OpenSecrets website [10, 11]. We processed the contribution files to obtain the total contributions to each Representative for each sector. For bills, the features considered were the bills's sponsor (party and campaign contributions), the number of cosponsors from each party, the congressional committees the bill was referred to, and a list of "tags" describing the topics related to the bill, each no more than a few words. These bill features, as well as the roll call votes for each bill, were obtained from govtrack.us [12]. The result for each example is the vote ("yes" or "no"). The vote of each of the 435 Representatives on ∼300 bills provides ∼130,000 examples for each Congress, totaling to ∼780,000.

In order to better understand the structure of campaign contributions, we tried several ways of clustering the campaign contribution data by sector for Representatives from all Congresses. For this step, we normalized the campaign contributions for each Representative to sum to 1, since otherwise the clustering was dominated by a few Representatives with very large contributions (such as John Kerry, when running for President). However, the results were not very informative, aside from indicating that the three largest sectors in terms of campaign finance are Finance, Health, and Labor Unions in that order. Because of this, we chose instead to visualize the campaign finance data using principal component analysis, or PCA. This algorithm (and all others described, unless otherwise mentioned) was implemented using the Python machine learning package [13].

PCA reduces the dimensionality of the data by finding the linear combinations of the initial feature axes along which the variance is maximized. This is done by finding the eigenvectors of the matrix $\sum_{i=1}^{m} x^{(i)} x^{(i)T}$, where $x^{(i)}$ is the feature vector for example i and m is the total number of examples, and choosing the k eigenvectors with the largest eigenvalues, where k is the desired number of reduced dimensions. To maximize the variance of the data while still allowing it to be easily visualized, we chose k=3 (Fig. 1).

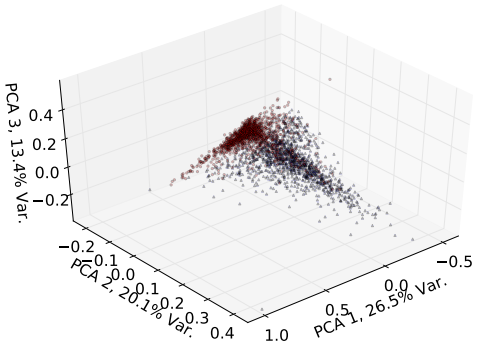

FIG. 1. Campaign finance data, reduced to three axes using PCA, with Republicans in red and Democrats in blue. The fraction of the total variance of the data explained by each axis is given in the axis label. Note that the two parties are fairly clearly separated by the second PCA axis, which is dominated by contributions from labor unions.

These three axes accounted for 65% of the total variance of the campaign finance data. The PCA results suggest that Representatives do not naturally separate into distinct clusters, which could be related to the lack of insights from clustering. However, they do show a clear separation by party according to the second PCA axis, which largely (anti-)aligned with contributions from labor unions. This accords with the expectation that labor unions contribute primarily to Democrats, while other sectors are less ideologically aligned with a particular party. The principal axis was largely aligned with the Candidate Committees sector as well as the Ideological/Single Issue PACs sector, suggesting that much of the variance may come from separating out high-profile politicians who get a substantial fraction of their campaign funds from dedicated election committees. The third principal axis was primarily aligned with the Finance sector, though it also was a significant component of the other axes. Hence, while both Finance and Health contribute substantially, the Finance sector seems to be more politicized, while Representatives across the board receive donations from the Health sector.

Significant effort was also devoted to understanding the average voting behavior of Representatives, with particular attention paid to their political party affiliations. An analysis of bill votes for a sample Congress is given in Fig. 2. For all Congresses in this dataset, the majority party sponsors most of the bills and its members vote "yes" nearly unanimously on most bills. In contrast, minority party members vote "yes" nearly unanimously on bills it sponsored, but votes bimodally for bills sponsored by the opposition. This is true regardless of whether the majority party is Republican or Democrat. This can be distilled into two main features of this dataset: 1) Rep-

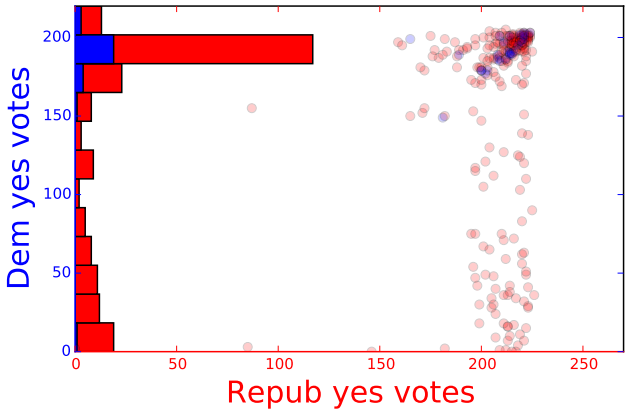

FIG. 2. Scatter plot of bill votes by party for Congress 108, with each bill colored according to the party of the bill's sponsor. The results are projected onto a histogram on the minority party vote axis, again with bars colored by the sponsor party. The majority party sponsors most of the bills and votes "yes" nearly unanimously, while the minority party votes "yes" nearly unanimously on bills it sponsored, but votes bimodally for bills sponsored by the opposition.

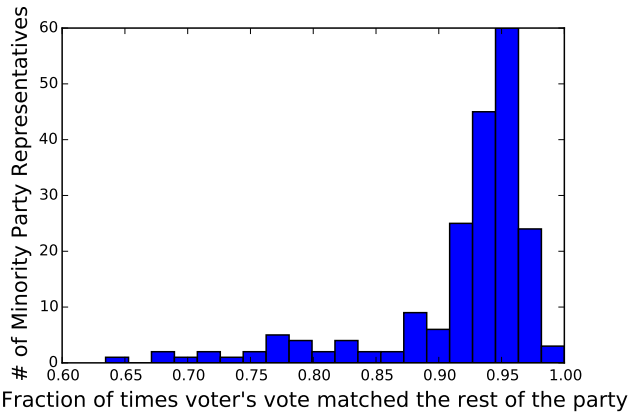

FIG. 3. Histogram of the fraction of votes which agree with the collective vote of the party, by minority party Representative.

resentatives largely vote according to party (herd mentality), and 2) bills can be separated into a large class of uncontroversial bills (which pass with an overwhelming margin) and the remainder which are controversial. In fact, roughly 80% of votes in this dataset were "yes" votes, indicating that the result classes are quite unbalanced, and further establishing a baseline accuracy of 80% with which to compare the performance of any prediction method.

This is further illustrated in Fig. 3. This demonstrates that most members of the minority party vote the same way as their party (as determined by the majority of votes) greater than 90% of the time, with an average agreement of about 95%. Hence, since Representatives cast roughly equal numbers of votes, very high prediction accuracy (>95%) can be achieved by determining how the minority party will vote on any given bill, without any further distinguishing information about any Representative than their party.

ANALYSIS

Given the characterization of this dataset, most of our prediction efforts went toward determining whether a bill would be "controversial," meaning that the minority party as a whole would vote against it. Our primary algorithm of choice for doing this was logistic regression. Logistic regression consists of finding a coefficient vector θ of the same length as the input feature vectors $x^{(i)}$, and returning the quantity $h_\theta(x^{(i)}) = \frac{1}{1+e^{-\theta^T x^{(i)}}}$ as the probability that example $x^{(i)}$ will be a "yes" vote. The coefficient vector θ is determined so as to maximize the likelihood of θ given the training data, using gradient descent. Specifically, the gradient descent rule for updating θ is given by $\theta := \theta + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x^{(i)}$ where α is a parameter (the "learning rate") which determines the speed of convergence. This step is repeated (either for one $x^{(i)}$ at a time, or with a sum over all $x^{(i)}$ in the training set) until θ converges to within some tolerance.

Initial tests of this method used the following features:

1. Voter and sponsor party
2. Features from 1 plus bill features (described below)
3. Features from 2 plus campaign contributions for both voter and sponsor

Originally, party was encoded as 0 for Republicans and 1 for Democrats, but we found that the generalization error for training on Congresses 108-112 and testing on Congress 113 was worse than when training and testing on any single training Congress, which we interpreted to be due to changes to the party in power from one Congress to the next. When the party was instead encoded as 0 for the minority party and 1 for the majority, this discrepancy went away, so this encoding was used for all results shown, unless otherwise noted. The bill features include the number of cosponsors from the majority and minority parties, as well as "vote fractions" for the bill's tags and congressional committees. For any given tag (committee), the vote fraction is calculated by taking the fraction of all training votes on bills containing that tag (referred to that committee) which were "yes" votes. Hence, this is an indicator of how generically uncontroversial (to either party) bills with that tag (referred to that committee) are. This attempts to capture the intuition that there are many generically uncontroversial bills related to topics such as naming a new post office, and

that some tags (or committees) could identify such a bill. Note that while most tags present in bills for Congress 113 (test data) were present in the training data, none of the actual votes or bills were, so that there is no way for the algorithm to cheat by knowing in advance how a given bill will be voted on.

Given the list of tags (committees) for a bill, the feature used as input to logistic regression was the simple mean of the vote fraction for all the tags (committees) present; hence, there are only two additional features involving vote fractions. Inspection of the tag vote fractions by eye suggests that they do capture generic uncontroversiality: tags with low vote fractions include controversial topics such as terminal illness and the draft, while tags with high vote fractions include topics all Representatives are likely to agree on such as nature, radioactivity, and Nazism. Campaign contributions were not normalized by individual (as they were for PCA), but were left as absolute contributions, in thousands of dollars, from each of the thirteen funding sectors.

The results of performing logistic regression on these feature subsets are shown in Fig. 4. We choose as our performance metric the overall accuracy of classification, as there is no particular reason to penalize false negatives over false positives, though we also report the precision-recall curve for "yes" votes for completeness, where precision is the fraction of predicted "yes" votes which are correct predictions, and recall is the fraction of true "yes" votes which are correct predictions. Note that due to the high precision for much of the graph, the area under the curve is near 1 for all classifiers, making it less sensitive as a metric than accuracy. The results indicate that inclusion of the bill features does improve the accuracy, from 82% to 88%, comparable to methods of other researchers. However, such a comparison is at best suggestive, as the datasets are different (it seems plausible that more recent polarization of the House would make voting behavior more easily predicted, for example).

The further inclusion of campaign finance data smoothed out the curves, but did not provide any substantial improvement, suggesting that any relevant information it might provide was already captured by the other features. This agrees with the expectation (from dataset characterization) that a Representative's party is the only distinguishing feature needed to predict her or his voting behavior. In addition, the training and test errors were very similar in all cases (88.1% training accuracy versus 87.8% test accuracy for the best-performing classifier, *i.e.* using the bill features). This is to be expected, as the models are all fairly simple (<50 features) and the dataset is reasonably large (nearly 1 million examples).

Because the models above did not show evidence of overfitting, we chose to include additional features that might better capture the content of a bill, beyond whether it is generically controversial to both parties. To

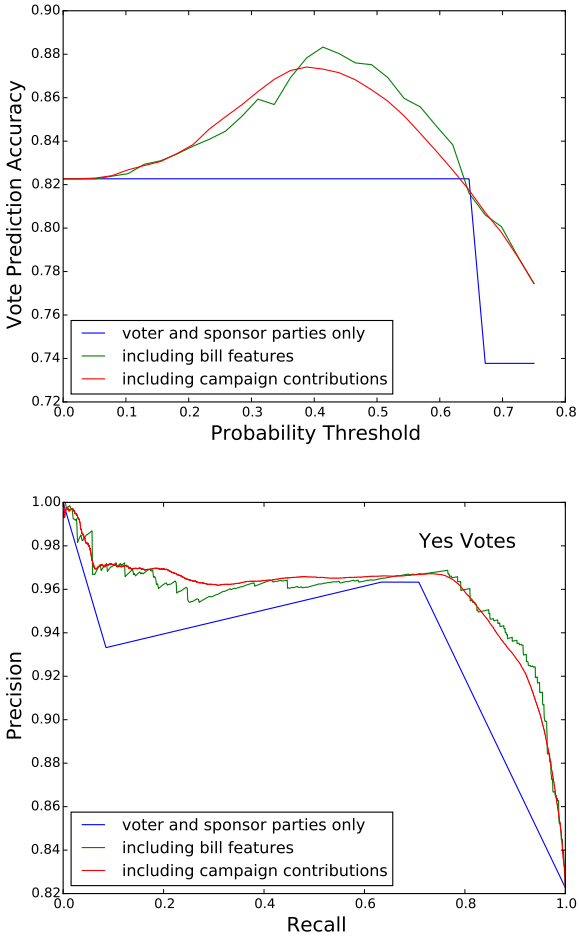

FIG. 4. Above: accuracy of the logistic regression classifiers with different input features versus threshold for predicting a "yes" vote. Below: precision-recall curve for the logistic regression classifiers.

do this, we included as features the presence of tags, encoded as 0 when the relevant tag was absent, and 1 when present. Since there are over 3000 unique tags in this dataset, too many to include as features, we reduced the tags to the 100 most useful, as determined by their mutual information with the vote outcomes from the training dataset. The mutual information between tag t_i and the vote result y (both binary) is given by

$$MI(t_i, y) = \sum_{t_i \epsilon \{0,1\}} \sum_{y \epsilon \{0,1\}} p(t_i, y) \log \frac{p(t_i, y)}{p(t_i)p(y)}$$

Including these top 100 tags (or fewer) in the logistic regression model as additional features, however, did not improve the accuracy. Switching the party encoding back to Republicans and Democrats (rather than majority and minority), to try and better account for party-specific differences, did not improve the model.

Another attempt to determine whether a bill is contro-

versial or not was to use the Naive Bayes method on the tags: for tag t_i, the probability of the tag being present in a controversial (uncontroversial) bill, ϕ_i^c (ϕ_i^u), is given by the number of controversial (uncontroversial) bills containing that tag divided by the total number of controversial (uncontroversial) bills. This was modified slightly by the Laplace smoothing method to account for tags which appeared very rarely, by adding 1 to the controversial (uncontroversial) count for each tag and also adding the total number of unique tags to the denominator (number of controversial or uncontroversial bills). The probability of a future bill with tag list T being controversial is then calculated as:

$$\left(1+\frac{\phi^u}{\phi^c}\prod_{t_i \epsilon T}\frac{\phi_i^u}{\phi_i^c}\right)^{-1}$$

where ϕ^c (ϕ^u) is the fraction of all bills which are controversial (uncontroversial). In this context, a bill was determined to be controversial if >50% of minority party Representatives who cast a vote on it voted "no." Any tag which was not present in the training dataset was ignored for this calculation (a small fraction of total tags). To try and better determine party preferences (since some topics may be more likely to be voted down by Republicans than Democrats or vice-versa), the training was done only on Republican-majority Congresses (108, 109, and 112), and tested on Congress 113, which also had a Republican majority. This method achieved 75.9% accuracy in predicting which bills would be controversial in Congress 113. However, including the Naive Bayes probability of being controversial as an additional input to logistic regression did not improve the model's accuracy. This suggests that all of the information in the Naive Bayes output which might help classify bills as controversial was already present in the tag vote fractions.

As a final step, the same features were used to train a support vector machine (SVM), with the thought that its increased complexity might reduce the bias of our model. Because the SVM method relies on defining a distance metric between examples, and the inputs are measured in different units, we first scaled the features so that each feature from the training data had a mean of 0 and a variance of 1, then applied the same scaling to the testing data. The SVM algorithm works by finding a coefficient vector w (and bias term b) so as to maximize the minimum distance (over all examples) to the decision boundary defined by $w^T x+b$; hence, its goal is to choose a decision boundary using the examples nearest to the true boundary between classes ("support vectors") as a guide. Because real data is rarely perfectly separable by such a linear decision boundary, the constraints are loosened to allow examples to be on the wrong side of the decision boundary, with a corresponding penalty in the maximization term determined by the regularization parameter C (larger values of C mean a higher penalty for such misclassified examples). This problem can be mapped onto the following constrained maximization problem:

$$\max_\alpha W(\alpha)=\sum_{i=1}^m \alpha_i-\frac{1}{2}\sum_{i,j=1}^m \alpha_i\alpha_j y^{(i)}y^{(j)}x^{(i)T}x^{(i)}$$

$$\text{s.t. } 0\le\alpha_i\le C,\ i=1,...,m \quad \text{and} \quad \sum_{i=1}^m \alpha_i y^{(i)}=0$$

The results of this approach, with different feature subsets and with $C=1$, are essentially identical to the logistic regression results, with the highest accuracy of 88% achieved with the simple list of bill features; adding the top 100 tags or the Naive Bayes controversiality probability did not improve the classifier, and in the case of the raw tags, led to overfitting (a training accuracy of 90% was achieved, while the testing accuracy was 85%). Adjustments to C did not improve upon the best result of 88% testing accuracy.

CONCLUSIONS

Campaign contributions to members of the House of Representatives were analyzed via PCA, finding that the Health sector gave to Representatives across the board, Labor Unions donated primarily to Democrats, and contributions from Finance and Candidate Committees showed the highest variation independent of party; hence, donations from these last two sectors are ideal for distinguishing between Representatives. However, it was determined that votes are primarily the result of a collective party decision, so individual features (including campaign finance data) have limited value. Furthermore, many bills pass with a high margin, and members of the majority party almost always vote "yes" (as they sponsor the bills being considered), so most of the improvement in accuracy over the 80% baseline (from predicting all "yes" votes) comes from correctly predicting which bills will be controversial to the minority party.

Using logistic regression with party information for bill sponsors, cosponsors, and voters, and further information about the bill's controversiality from its tags improves the accuracy to 88%, comparable to the accuracy of state-of-the-art methods which use Representatives' vote histories (albeit on a different dataset, making such direct comparison difficult). More complex models which try to account for party-specific preferences showed no improvement. This technique could be improved by a more complete topical model (sorting bills into broader categories) or the addition of lobbying information. It would also be interesting to learn more about the 5% of Representatives who do not vote with their party, by looking more into individual features, such as by analyzing outliers in the campaign contribution data.

The authors acknowledge Nick Sher for useful discussions.

* henighan@stanford.edu
† skravitz@stanford.edu

[1] "Patient protection and affordable care act, 42 u.s.c. § 18001," (2010).
[2] "Occupy wall street," http://occupywallst.org/, accessed: 2015-11-15.
[3] V. Eidelman, in *COLING (Posters)* (2012) pp. 275–286.
[4] T. Yano, N. A. Smith, and J. D. Wilkerson, in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, 2012) pp. 793–802.
[5] T. G. James M. Snyder, American Journal of Political Science **44**, 193 (2000).
[6] H. F. Weisberg, American Journal of Political Science **22**, 554 (1978).
[7] T. Stratmann, Southern Economic Journal **57**, 606 (1991).
[8] J. R. Wright, American Political Science Review **84**, 417 (1990).
[9] S. Gerrish and D. M. Blei, in *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Curran Associates, Inc., 2012) pp. 2753–2761.
[10] "Follow the money, a handbook," https://www.opensecrets.org/resources/ftm/ch12p1.php, accessed: 2015-11-15.
[11] "Crp categories," https://www.opensecrets.org/downloads/crp/CRP_Categories.txt, accessed: 2015-11-15.
[12] "Govtrack," https://www.govtrack.us/data/us/, accessed: 2015-11-15.
[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).