Institute for New Economic Thinking

Working Papers
ineteconomics.org/workingpapers

Measuring the Impact of Campaign Finance on Congressional Voting: A Machine Learning Approach

Matthias Lalisse*

Working Paper No. 178

February 22nd, 2022

ABSTRACT

How much does money drive legislative outcomes in the United States? In this article, we use aggregated campaign finance data as well as a Transformer based text embedding model to predict roll call votes for legislation in the US Congress with more than 90% accuracy. In a series of model comparisons in which the input feature sets are varied, we investigate the extent to which campaign finance is predictive of voting behavior in comparison with variables like partisan affiliation. We find that the financial interests backing a legislator's campaigns are independently predictive in both chambers of Congress, but also uncover a sizable asymmetry between the Senate and the House of Representatives. These findings are cross-referenced with a Representational Similarity Analysis (RSA) linking legislators' financial and voting records, in which we show that "legislators who vote together get paid together", again discovering an asymmetry between the House and the Senate in the additional predictive power of campaign finance once party is accounted for. We suggest an explanation of these facts in terms of Thomas Ferguson's Investment Theory of Party Competition: due to a number of structural differences between the House and Senate, but chiefly the lower amortized cost of obtaining individuated influence with Senators, political investors prefer operating on the House using the party as a proxy.

https://doi.org/10.36687/inetwp178

JEL Codes: H10, D72, P16, C45

Keywords: campaign finance, congressional voting, investment theory of party competition, machine learning, Representational Similarity Analysis, political money

* Department of Cognitive Science, Johns Hopkins University (lalisse@jhu.edu)

1 Introduction [1]

The empirical case that campaign finance plays a defining role in American politics is not difficult to make, to the point that voters essentially take for granted money's dominant role in defining the parameters of the political agenda in Washington [11]. The cost of presidential and congressional elections hit a record high in 2020, with more than $14 billion spent in direct and indirect ("dark money") contributions [14]. The sharp escalation of the role of money in politics over the last decade is understood at a broad level by both the public and by politicians. Some of the latter have advanced popular agendas to curb its influence through various measures, ranging from constitutional repeal of the Supreme Court's *Citizens United* decision [54] to the establishment of public election financing instruments that would match or supplant the private purchase of political influence [53].

The amount of money at play in federal elections is well documented at the aggregate level [6], and there is no shortage of expert warnings about the dangers that the corresponding economic capture of political institutions poses for democracy [50, 59]. Money's effects on the probability of campaign success is well studied [17, 61, 30, 31], and reporters covering political dealings in D.C. frequently propose highly granular causal links between contributions and legislative behavior in relation to live political issues—e.g. [27, 39]. While there exist quantitative case studies making similar arguments for small or narrow samples of legislation [18, 34, 26, 21, 22], as well as work identifying other significant vote predictors such as gender [60, 32, 23, 55] and party [57], few studies investigate (1) the effect of money in politics in a manner that (2) is based on aggregate campaign finance data and (3) establishes quantitative links between aggregate campaign finance flows and particular legislators' political decisions [3, 35].

Machine learning can provide a bridge between these levels of analysis. By modeling legislators and bills, this article produces accurate predictions of individual legislators' votes across several Congresses. By exploring the feature space for the legislator and bill models, we demonstrate the independent predictive power of campaign finance in comparison to indicators like party and state. A central feature of our approach is the use of a

[1] All the data and code required to replicate this paper's main analyses are publicly available on the Harvard Dataverse `https://doi.org/10.7910/DVN/DHQQHX`.

Transformer-based document embedding model to provide representations of bills [9]. In contrast to some prior approaches, this allows us to generate predictions for new legislation with no prior observations of votes on a given bill [3, 56]. This computational methodology, which leverages well-validated feature-extraction (vectorization) techniques for text data, facilitates the analysis congressional voting patterns at scale.

1.1 Overview of the article

Section 2 describes and validates the feature sets used in our analyses, which are a combination of text embeddings from a language model and financial features summarizing the source of a legislator's campaign funds. In Section 3, we describe the model evaluation task: legislative roll-call prediction. Section 4 presents the results from the vote prediction test. We find that individual votes are predictable with more than 90% accuracy when campaign finance features are included, and that these features are more than a proxy for political party. In addition, we find that the independent effect of financial features is substantially larger in the Senate than in the House. After discussion in Section 5, we present a Representational Similarity Analysis (RSA) that corroborates both results from a different perspective: "legislators who vote together get paid together", with a stronger effect in the Senate than in the House. In Section 7, we develop an explanation of these results in terms of the Investor Theory of Party Competition [16]. Section 8 concludes.

2 Data

2.1 Bill text model

To produce bill features, we use the Longformer model [2], a Transformer optimized for embedding medium-length documents. In contrast to BERT's standard global attention, each Longformer layer uses self-attention in a fixed window around each linear text position, reducing the computational complexity from quadratic to linear in the input length—$O(\ell \times w)$, with ℓ the input length and w the window size. Transformers in general

Variable	n-features	Description
LEGFIN	25	Summary vector for a legislator's campaign finance profile. PCA on federal election filings.
LEGPRTY	3	Binary indicator variables for the legislator's party (Democrat, Republican, or Independent).
BILLVEC	25	Document embedding of each bill summary. CRS summary text embedded to 768d using the LongFormer, projected to 25d via PCA.
SPONPRTY	3	Binary indicator variables for the party of the bill's main sponsor.

Table 1: The variable sets used in vote prediction.

and the LongFormer in particular are similar in principle to convolutional neural networks (CNNs) in that they iteratively aggregate local information across successive layers. Information can thus propagate across long positional distances in a way that is sensitive both to linear adjacency, hierarchical, and also larger-scale structure. In contrast to uncontextualized word embeddings like GLOVE and WORD2VEC [49, 37, 43, 47], models within this architectural class have been demonstrated to carry sophisticated linguistic information, such as token-level syntax and semantics [41].

Using the 12-head, 12-layer `longformer-base` model pre-trained on masked language modeling, we obtained embeddings of the bill summaries generated by the Congressional Research Service (CRS) for each bill before Congress. The bill's official title was concatenated with its CRS description, both retrieved using the ProPublica Congress API [51] and truncated to the embedding model's maximum input length (4,096 tokens). The document summary vector is the last hidden state of the first token (`CLS`) in the input. The set of 5,083 summary vectors—one for each bill brought to a roll-call vote from the start of the 110th Congress (Jan. 3, 2007) to the 117th Congress (up to Dec. 2, 2021)—was then projected to a 25-dimensional subspace using PCA. The bill sets for the 110th to 112th Congress were not included in the roll call prediction evaluation since we did not have campaign finance data going back far enough to model legislators from that period, but they were included in the PCA fit in order to get better coverage of the semantic space. We

also included the party of the bill sponsor as a feature in the model comparisons of Section 4.

2.1.1 Party of sponsor classification

To verify that the bill vectors are capturing political/policy content, we evaluated whether it was possible to predict the party of the bill's main sponsor using the Transformer embeddings. This is an imperfect measure since 49.5% of the bills in our dataset have bipartisan cosponsors, and for this reason and others, partisan sponsorship prediction is not an interesting outcome variable in and of itself. Thus, we use it exclusively for the purpose of validating the embedding method. We find sponsor party classification based on the `BILLVEC` features is at 78.2% accuracy ($N = 2,274$) in ten-fold cross-validation, which is 145% of the chance level.[2] Thus, LongFormer summary embeddings are generally sufficient to detect partisan lean.

2.1.2 Policy subject classification

Do the embeddings also carry information about policy content? We checked whether the representations allowed us to predict each bill's primary subject as identified by the CRS. This too is an imperfect measure since bills typically span a broad range of policy topics, and so we use it solely for validation of the input representations. CRS policy topics were detected with 79.2% accuracy ($N = 2,274$), with 29 classes (3.4% from uniform guessing and 41.6% frequency for the most frequent class, "Congress").

2.2 Legislator model

To model legislators, we used Federal Election Commission (FEC) data pre-processed by MapLight [42]. The MapLight data tables record individual contributions to candidate-

[2]Democrat and GOP bill sponsorship is approximately balanced, and Independents are rare. The most frequent class (Republican sponsor) for bills brought to a floor vote occurs 54% of the time in this dataset.

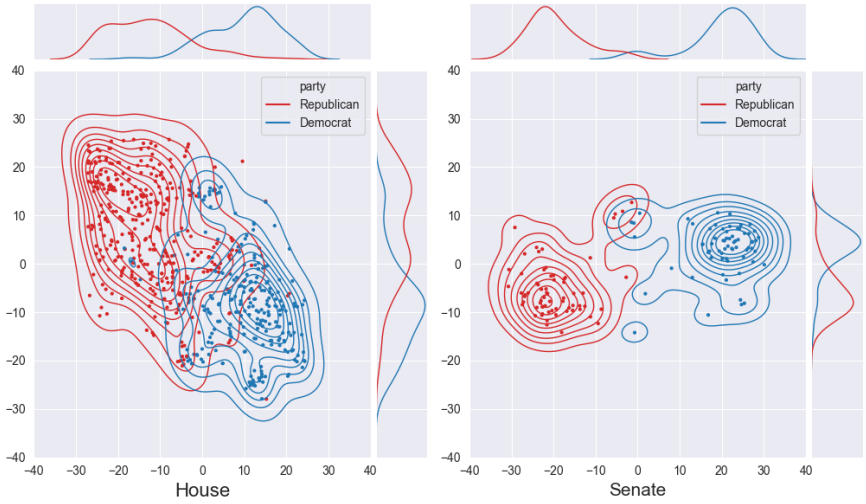

Figure 1: t-SNE visualization of the campaign finance "embeddings" (LEGFIN features) for the House (left) and Senate (right), with Gaussian kernel density estimates of the distributions for Republicans and Democrats superposed. Axes are shared across chambers. A projection annotated with legislator names is provided in the Appendix (Fig. 15).

affiliated Political Action Committees (PACs). Unitemized contributions, i.e. those from donors who contribute $200 or less in a given cycle, and whose names are not required to be reported to the FEC, were removed. We modeled every legislator who served in any of the 110th through 117th Congresses and whose campaign committee(s) logged at least $100,000 in contributions across that period. To disentangle effects of financial contributions from the influence of party, contributions from party organizations as well as transfers from other candidates' committees were removed, retaining all payments from candidates into their own committees and from institutional donors (e.g. corporate, labor, ideological, or trade group PACs such as those affiliated with the NRA and Chamber of Commerce) who had made at least one contribution of $1,000 to a legislator who served. This resulted in an inventory of 17,002 unique donors.[3]

These donors' transactions were arranged into a sparse matrix with legislators as rows and donors as columns, each cell containing the cumulative campaign contributions from

[3]As a check on the robustness of the results to small changes in the data annotation, we replicated all analyses in this article using data pre-processed by the Center for Responsive Politics (CRP). Full results tables are give in the Appendix (Section 9.4). All results (roll call prediction and RSA) successfully replicated across this change in source data.

the donor to the candidate during the 12-year period covered by the data. This sample accounted for about $3 billion in campaign contributions. The columns of this matrix were normalized (featurewise z-scoring), and from it were extracted 25 principal components. Each legislator was projected onto the subspace spanned by these components, yielding a dense vector that is the input to vote prediction.

2.2.1 Legislator party classification

To validate the resulting representations, we performed a simple three-way classification analysis to predict a legislator's party affiliation from their legislator finance (`LEGFIN`) vector using a random forest classifier with 100 estimators. The dataset was approximately balanced between Democrats and Republicans, but Libertarians and self-identified Independents were merged into a single class due to their small number. **Results.** A legislator's party was predictable from their `LEGFIN` embedding with 97.3% accuracy overall, with the results marginally higher for the House than for the Senate (Table 2). Independents were always misclassified, though this is not surprising given their low prior.

Table 2: Summary statistics and party-classification results for the legislator model (`LEGFIN`).

Chamber	*N* **Dem**	*N* **Rep**	*N* **Ind**	*N* **Total**	**Accuracy**
House	418	439	2	859	.977
Senate	81	82	3	166	.952
All	499	521	5	1025	.973
Accuracy	.968	.987	0.0	.973	

In the model comparisons of Section 3, we additionally include an indicator variable for the legislator's party affiliation. Table 5 in the Appendix reports expanded results with features for the legislator state, but the change to the results from adding this 50d indicator was minimal.

2.3 Roll Calls

Not every congressional vote is comprehensively tallied with individual legislators' positions recorded. In both chambers (House and Senate), those that are are identified with a roll call. Roll is taken for a variety of floor votes, including passage of a bill, amendments, votes on non-binding resolutions, joint resolutions (legislation before both chambers), and cloture (vote to close debate, e.g. to end a filibuster).

To evaluate our featural specifications, we obtained complete records of all floor votes for the 113th to the 117th Congress (Jan. 2013 to Nov 2021) using the VoteView roll call database [40]. Individual legislators were matched to FEC, partisan, and other identifiers using an interstitial database [52]. We retained only votes on the passage of bills, resolutions, or joint resolutions, omitting votes on amendments, procedural motions, overriding a veto, and other categories that have an unclear relation between the bill content and the individual vote.

3 Evaluation

House and Senate votes were grouped together in training, and a single random forest model (200 estimators) was fit to the data array obtained by concatenating the elements of the feature inventory of Table 1 for each individual vote. We evaluated each feature set using a cross-validation procedure in which 10% of bills are held out in each fold, holding the folds themselves constant across model runs.

In addition to this bill-held-out CV scheme, we add **cross-validation by Congress** in which the model is trained on the data from four of the five congressional roll call sets (the 113th to 117th), and predicts roll call for the remaining session. This is more difficult not only because the held-out partition is larger, but because the held out set is guaranteed to include unseen *legislators* as well as unseen *bills*, meaning that the model has to successfully generalize from features of both [47, 36].

3.1 Baselines

The notional chance accuracy for the three-class vote prediction problem is .33. However, votes of `Present` are infrequent and floor votes on bills are substantially weighted towards `Yea` (around 60%), likely due to party leaders' reluctance to move for a recorded vote unless a bill is likely to succeed. Thus, the unconditioned majority vote of `Yea` provides a first baseline. We also include prediction on the basis of the `BILLVEC` vector alone, which roughly corresponds to predicting whether the bill will pass (predicting the majority vote conditioned on the Transformer embedding).

As an addition to the naive baselines, the `LEGPRTY + SPONPRTY` specification introduces interactions between the sponsor's and voter's party affiliations. With these two variables, the model can easily learn the partisan heuristic that Republicans vote for bills sponsored by Republicans but not those sponsored by Democrats, and mutatis mutandis for Democrats vis á vis Republicans. The model runs with this specification provide a more stringent baseline of 91.6% accuracy when combining House and Senate, but less so in the Senate (64.5%), where cross-partisan votes are more common.

		Accuracy					
		Cross-val by bill			**Cross-val by Congress**		
	Model specification	**House**	**Senate**	**Total**	**House**	**Senate**	**Total**
Baselines	MAJORITY_CLASS	.599	.772	.605	.599	.772	.605
	BILLVEC	.599	.767	.605	.599	.772	.605
Excludes financials	LEGPRTY+BILLVEC	.816	.700	.812	.714	.703	.713
	LEGPRTY+SPONPRTY	.926	.646	.916	.925	.648	.916
	LEGPRTY+BILLVEC+SPONPRTY	.924	.719	.917	.921	.689	.913
Includes financials	LEGFIN+BILLVEC	.845	.832	.845	.734	.769	.735
	LEGFIN+SPONPRTY	.927	.777	.922	.896	.772	.892
	LEGFIN+BILLVEC+SPONPRTY	**.932**	**.863**	**.930**	**.927**	**.816**	.923
	FULL_MODEL	**.932**	.859	**.930**	**.927**	.812	**.924**

Table 3: **Vote prediction results**. Results on roll call prediction in the two cross-validation regimes. The `FULL_MODEL` refers to the specification with all variable sets included (`LEGFIN + LEGPRTY + BILLVEC + SPONPRTY`). The best results excluding financials are highlighted in red, and the best overall in blue.

4 Results

Accuracy statistics for the two cross-validation regimes are reported in Table 3, where they are also split by chamber. Models that included the LEGFIN features reliably performed better than those without, including the strong partisan baseline of LEGPRTY + SPONPRTY, though the differences are small in the House. The LEGFIN features are an effective substitute for the legislator's party in all but one case (the House in the CV by Congress regime), indicating that the campaign finance features *at minimum* serve as a proxy for party. Apart from that exception, swapping the party indicator with the LEGFIN features[4] increases performance, and drastically in the Senate (always more than 6 and up to 15 points in LEGPRTY + BILLVEC + SPONPRTY). The bill summary embeddings tended to improve performance, and in a manner that is not reducible to the BILLVEC representation serving as a proxy for the SPONPRTY indicator. The best performing model overall includes the legislator financials, the bill embedding, as well as the partisan signal SPONPRTY. Partisan information about the voter was not included in the best model, suggesting that campaign finance provides an exhaustive proxy for that variable while also carrying additional predictive information.

We also observed a dramatic asymmetry between the House and the Senate in (1) overall performance and (2) the degree to which campaign finance features affected performance. Top vote prediction accuracy in the Senate is 6.9%/11.1% lower than in the House (CV by bill/CV by Congress), but while the difference between LEGPRTY + BILLVEC + SPONPRTY and LEGPRTY + BILLVEC + SPONPRTY is .6%/.2% in the House, it is 14.4%/11.3% in the Senate.

[4] The relevant comparisons are the corresponding rows of the "Includes financials"/"Excludes financials" super-rows, i.e. LEGPRTY + BILLVEC $\longrightarrow$ LEGFIN + BILLVEC, LEGPRTY + SPONPRTY $\longrightarrow$ LEGFIN + SPONPRTY, etc.

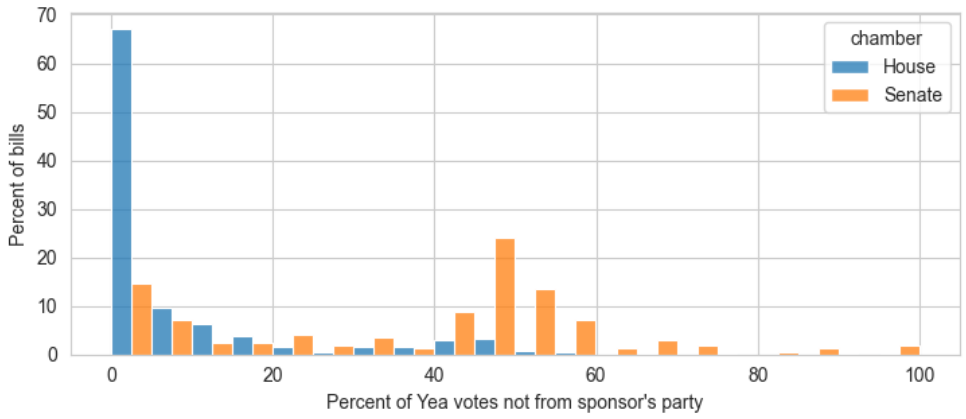

Figure 2: **Bipartisan voting in the House and Senate**. 45 percent of House votes in the 113th to 117th Congresses had less than 5 `Yea` votes from outside the sponsor's party. Normalizing for the size of each body, the corresponding figure for the Senate is less than 10 percent. The median House bill has 6 non-partisan votes. In the Senate, 44.

5 Interim Discussion

5.1 Money's effect on predicting political outcomes

The best performance across both evaluation settings required the inclusion of (1) the Transformer bill embeddings, which capture both partisan lean and policy content, and (2) information about who financed the legislator's campaign. Importantly, the financial features did not just substitute for partisan affiliation, but almost always *improved* upon it in pairwise model comparisons—the empirical signature of money's independent effect.

Moreover, these results are the best reported in the literature on legislative vote prediction (see the Appendix for a review, esp. Table 6). Likely for the good, roll call prediction has not annealed into a machine learning task with publicly traded train-test partitions. Also, that literature is sparse, and evaluation procedures are not always reported in sufficient detail to make direct comparison possible. Thus, caution is advised. However, it would appear that a legislator's campaign finance record is the feature that best predicts how that legislator will vote [47, 37, 56, 36].

5.2 The influence of LEGFIN in the House vs. Senate

Above, we remarked that marginal effect of the LEGFIN features is small in the House relative to the Senate (where it is 15 percent). To verify that the asymmetry was not due to an imbalanced dataset—a consequence of the House's larger size—we ran the best models with (a) an indicator variable for the chamber—allowing the model to distinguish between the House and Senate—and, separately, (b) with distinct model runs for each chamber. Performance decreased in both cases, confirming that the asymmetry was not due to training the House/Senate models from a common pool of votes.

In part, this result can be attributed to the fact that partisan voting is dominant in the House (Fig. 2), meaning that partisan affiliation is usually sufficient to predict a House member's vote [8, 1, 56]. Bipartisan voting is much more common in the Senate. That difference has clear structural origins in Senate mechanisms like the filibuster, which forces bills to clear a supermajority threshold, thus coercing them into a form palatable to both parties, generally at the expense of majoritarian policy [20, 13]. Nevertheless, while the comparative rate of bipartisan voting can explain why the partisan indicators are less predictive in the Senate, it does not explain why *financial* features in particular would fill the gap.

6 Representational Similarity Analysis

To further investigate the differing roles of money in the two chambers of Congress, we carried out a Representational Similarity Analysis (RSA) to relate legislators' voting behavior to their voting profiles in a way that allows us to visualize and quantify the extent to which voting blocs correspond to blocs of legislators with common funding sources. In RSA [38, 10], multivariate series are modeled as collections of points with defined spatial relationships (similarities and distances). We then ask whether the spatial structure observed in one set of variables is replicated in the other set.

In our case, the multivariate series are the legislators' campaign finance features (LEGFIN) and their corresponding voting profiles (how they voted on a common inventory of bills). If legislators are backed by similar economic interests (i.e. are close together with reference

	House				Senate			
Congress	ρ	p	ρ_{part}	p_{part}	ρ	p	ρ_{part}	p_{part}
113th	$.476^{*}$	$< 10^{-5}$	$.047*$	$< 10^{-5}$	$.548^{*}$	$< 10^{-5}$	$.291^{*}$	$< 10^{-5}$
114th	$.529^{*}$	$< 10^{-5}$	$.057^{*}$	$< 10^{-5}$	$.679^{*}$	$< 10^{-5}$	$.347^{*}$	$< 10^{-5}$
115th	$.534^{*}$	$< 10^{-5}$	$.076^{*}$	$< 10^{-5}$	$.681^{*}$	$< 10^{-5}$	$.179^{*}$	$< 10^{-5}$
116th	$.522^{*}$	$< 10^{-5}$	$.056^{*}$	$< 10^{-5}$	$.670^{*}$	$< 10^{-5}$	$.274^{*}$	$< 10^{-5}$
117th	$.418^{*}$	$< 10^{-5}$	$-.009^{n.s.}$	.2171	$.635^{*}$	$< 10^{-5}$	$.269^{*}$	$< 10^{-5}$

Table 4: **Representational similarity analysis** of the voting profile and financial profile (`LEGFIN`) RDMs (e.g. Fig. 3). ρ_{part} is the Spearman correlation between the vote/financial profiles when party and state are controlled for. Congress-wise p-values are based on a permutation test in which the rows of the vote profile vectors, the `LEGFIN` vectors, and the party and state indicators were shuffled 100,000 times, the pairwise distances and correlations then recomputed, including the partial correlations. $*$: significant ($\alpha = .01$) after Bonferroni correction for the 20 tests.

to their `LEGFIN` vectors), then we expect that they will have similar voting patterns (i.e. are close together in the space spanned by their votes). We represented legislators' voting profiles as a matrix with legislators as rows and bills as columns, with a 1 if the legislator voted `Yea` on the bill, -1 if they voted `Nay`, and a 0 if they voted `Present` or abstained from the vote. Pairwise distances between voting profiles were measured using the Manhattan metric, which is the sum of the absolute differences between corresponding votes. In the `LEGFIN` space, we used the cosine distance between the legislators' 25-dimensional PCA vectors. Since legislators have different vote inventories if they served in different chambers/terms, we performed the RSA separately in each chamber and in each of the five Congresses. These distances were then assembled into pairs of Representational Dissimilarity Matrices (one for the votes and one for the financial features) which are the inputs to the RSA. Fig. 3 shows the RDMs for the 116th Congress.

6.1 RSA results

To test whether the structure of congressional voting is predicted by a legislator's mix of campaign finance sources, we correlated the upper diagonal entries of the vote profile

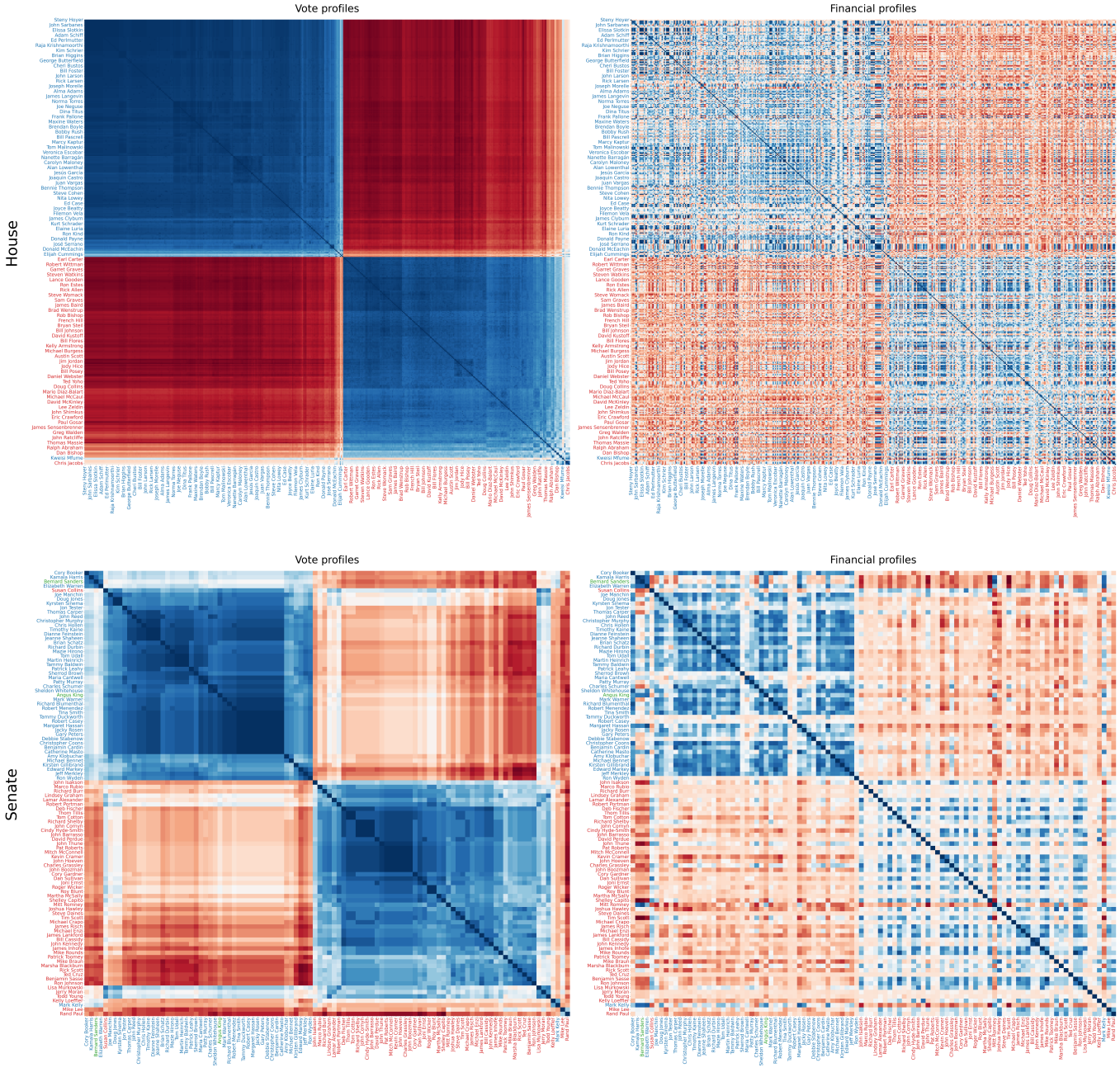

Figure 3: **Representational Distance Matrices (RDMs)** for Congressmembers (top) and Senators (bottom) in the 116th Congress (2019-2021) with legislators represented using their vote profiles (left) and financial profiles (`LEGFIN` features, right). Each cell represents the pairwise distance between the legislators in the corresponding column and row. Vote profile distances are measured using the Manhattan metric with `Yea` coded as 1, `Nay` as -1, and both `Present` and `Abstain` as 0. Financial profile distances are computed using the cosine measure applied pairwise to the 25d PCA vectors representing each legislator's funding sources. **Blue** denotes **high similarity**/low distance, and **Red** denotes **low similarity**/high distance. Only one fifth of the 435 House members are labeled with their names. The axes are sorted to preserve similarities between legislators' vote profiles. For the same data sorted according to similarity in funding sources, see the Appendix (Fig. 12).

and financial profile RDMs. If it is the case that "legislators who vote together get paid together", we expect a strong positive correlation between these matrices.

RDM cells were highly correlated across vote and financial profiles, with a mean Spearman's ρ of .495 in the House and .643 in the Senate. In large part, this reflects partisan structure: partisan voting blocs align well with the financial blocs reflected in the similarities among legislators' campaign finance profiles, a fact visualized in Fig. 3.

6.2 Partial correlation results

Does financial similarity explain voting behavior *beyond* party affiliation? We computed partial Spearman correlations between the vote/financial RDMs after controlling for the influence of party in both variables. We also controlled for the legislator's state as another potential source of variation, but including this variable had a negligible effect on the partial correlations in each case (see Table 7 in the Appendices). Table 4 reports the partial correlations controlling for both.[5]

We found weak but significant partial correlations (mean $\rho_{\text{part}} = .059$) in 4 of the 5 House sessions, all except for the 117th Congress (2021-2022) for which the vote record has half the data as well as a higher proportion of members (all the non-incumbents) with only one cycle of campaign finance data. In the Senate, we found strong correlations in each of the five sessions (mean $\rho_{\text{part}} = .272$), with campaign finance explaining on average 26 times the amount of variance in the behavioral measure (Senators' votes) as in the House after controlling for party. This converges with the asymmetry observed in vote prediction.

[5] In the partial correlation, the party-distance RDM—with a 0 (distance) in cell ij if legislator i is of the same party as legislator j and 1 otherwise—was entered as a covariate for both voting and financial distances, removing the variance in the Spearman rank series that can be explained by party affiliation (likewise for the state indicator).

7 Discussion

Cross-referencing two sources of evidence, we have found that (1) legislators' campaign finances are strongly predictive of how they vote on legislation, (2) that financial data are predictive of voting patterns *beyond* what is predictable from party identification, and (3) that financial data are more predictive in the Senate than in the House.

7.1 Explaining the House-Senate asymmetry

Traditional theories of political competition propose that political parties compete for vote shares from a population with defined policy preferences, and that they do so by modulating their policy platform so as to increase that share, thereby gravitating towards the position of the median voter[6] [7, 59, 12]. Theories of this sort have difficulty explaining the fact, which is not difficult to observe, that policies which have strong majoritarian support—such as single-payer healthcare—are viewed as politically unfeasible by politicians/media,[7] or why popular opinion is not very predictive of policy outcomes [25, 5].

In his Investment Theory of Party Competition, Ferguson [16, 19] proposes instead that political parties compete in a marketplace of *investors* whose financial contributions are necessary to secure electoral victories. When designing platforms, parties must balance semi-static voter preferences with the interests of moneyed and industrial interests who can help them meet the costs of mounting a campaign [46]. To meet the fixed costs of competing in an election, candidates must attain a minimum threshold of support from organized investors, which can take the form of mass support through institutions like unions, but more commonly is configured around corporate and industrial interests, since this requires less spontaneous social coordination. Mobilizing a base of investors in order to overcome the financial barriers to entry will circumscribe the set of policy positions a

[6] Stiglitz [59] refers to the median voter perspective—critically—as "the standard theory".

[7] In one of the more interesting moments of the 2020 presidential election, conservative Fox News Network found itself reporting that its own exit polls showed that Joe Biden's plan for a public option of government-provided health insurance had majority support not only generically, but also *among Republicans*, including in the "median-state" swing states [44]. Facts like this are difficult to account for if party platforms are the result of party adaptation to voter preferences rather than the other way around.

candidate can take, and for countries like the United States where the influence of money in elections is little regulated, the costs of meeting the financial threshold will tend to override voter preferences [58].

Rather than explaining voting patterns at the level of individual legislators and their individual votes, the Investment Theory is intended to account for party alignment and the content of party platforms on aggregate and at decade-long timescales. As such, it can explain the convergence of partisan voting blocs that are mirrored in campaign financing blocs observed in the RSA. The same basic framework can also help explain the House-Senate asymmetry.

The gist of the account is as follows: the brevity of House terms in comparison with Senate terms, coupled with the fact that House members tend to serve for fewer cumulative years than Senators, means that while campaign contributors may develop individuated investment strategies in relation to Senators, in the House they will prefer to operate using the party as an institutional proxy. We can understand this dynamic from the perspective of the two parties involved.

7.2 From the perspective of the investor

Obtaining individualized sway with a legislator is costly not just in terms of the sums of money involved, but also in terms of the more indefinite costs of developing a relationship that translates into influence. To a large extent, these are fixed costs. A relationship of influence requires a certain amount of direct contact either with the investor or with an intermediary (e.g. a lobbyist) so as to solidify the legislator's impression of a potential patronage, as well as an understanding of the investor interests on which that patronage is conditioned. Once a patronage is secured, it can be maintained automatically via regular transactions, but those initial fixed costs must be met. At a superficial quantitative level, the amortized cost of developing that relationship is much higher for a House member than for a Senator, since the typical House member will only serve half as long (Fig. 4), with nearly half (47 percent) of the current Senators having "graduated" from the House. In spite of presumably comparable (unamortized) fixed costs, the marginal value of a House

vote is less than that of a Senate vote due to the House's larger size, and an investment in sway with a House member is always riskier because the seat can be lost every two years.

Finally, while the size of constituencies for House members—and, therefore, the number of voters that must be reached in a campaign—is essentially uniform (~700,000), Senate constituencies range from 600,000 (Wyoming) to 40 million (California). The size of a constituency is related to the expected cost of the campaign to win it, and this heterogeneity can be exploited by investors who seek to capture congressional clients who can be bought for less as legislators set about meeting the costs of their election. Opportunistic poaching of this sort again favors targeted investment in the Senate over the House.

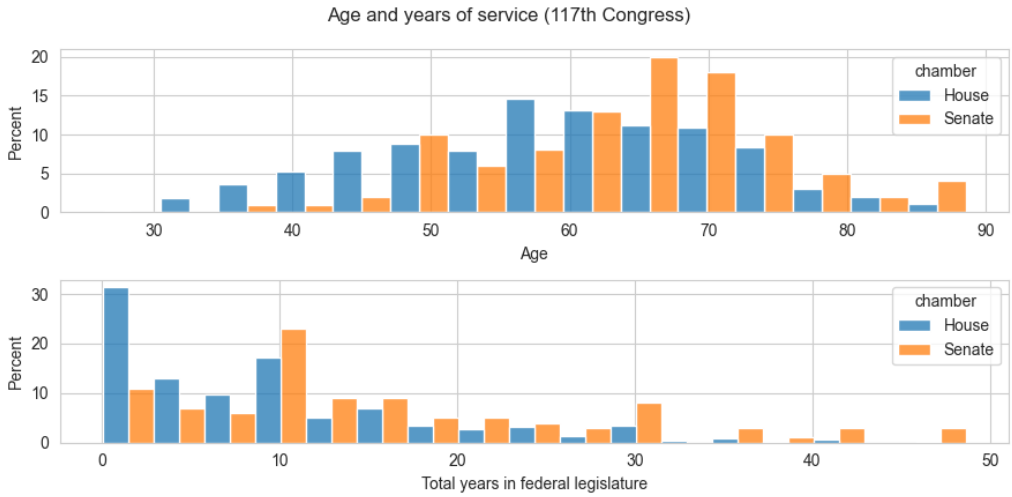

Figure 4: Distribution of age and years of service in the federal legislature (117th Congress). The median House representative has served 6.9 years, compared with 12.9 years for the Senate. Analysis of data from the `congress-legislators` database [52].

The alternative to retaining individuated patronage is the use of the party as a political proxy. Operating through institutional proxy comes with a loss of flexibility. The investor's role becomes primarily to maintain an aggregate party consensus by communicating with party leadership, who act as agents for investor interests via established mechanisms of party discipline (party whips, endorsements from leadership, apportionment of committee assignments, solidaristic fundraising within sub-partisan blocs, etc. [28, 48]). These mechanisms would appear to be very effective, as evinced by the rate of partisan voting

in the House (Fig. 2) and the crispness of voting blocs as visualized in Fig. 3. However, proxy control does not allow the investor to easily intervene in the legislative process with specificity, even if the content of legislation would hit core investor interests. This is mitigated by the fact that, in a regime where bicameral consent is required for nearly any substantive legislation, investors can mobilize Senate clients peeled off from partisan blocs through focused investment strategies, thereby blocking House action that is insufficiently attentive to their stakes.[8]

7.3 From the perspective of the legislator

House members are younger, more poorly resourced, and less experienced, with nearly a third of the 117th House having served less than two terms. The legislative work-load for a House member is about twice that of a Senator in terms of the number of roll call votes taken.[9] As well, the short election cycle means that Representatives must spend a disproportionate amount of time fundraising—reportedly, up to 30 hours per week, a phenomenon known as "call time" [15, 45]. Both congressional ethics rules and the federal criminal code prohibit members from fundraising in their congressional offices [29], making legislators dependent on centralized call-time facilities made available at the party HQ, several blocks from Capitol Hill. With investors only pursuing differentiated sponsorship with Senators or House leadership, House members lacking a continuous flow of funds are incentivized to defer to party leadership where possible, thereby maintaining access to this and other party fundraising instruments. This naked dependence on party resources for even legislator-driven fundraising efforts provides party leaders with tremendous leverage.

Turn now to the Senate, roughly half of whose members have served in the House. In an environment where investors are eager to procure individuated political sway—particularly as a bicameral stopgap to House action—a Senator may be given the choice between affiliation with party machinery on the one hand, and individuated patronage on the other.

[8] We can see a version of this dynamic in the 2021 debates surrounding the Build Back Better agenda, which passed smoothly through partisan voting in the House of Representatives, but was thereafter blocked by Senators Joe Manchin and Kyrsten Sinema. There is good evidence that the pair's deviation from the governing party consensus coincided with a targeted intervention from patrons in affected industries [39].

[9] 620 per Senator per session, vs. 1,170 for Representatives.

If choosing the latter, the legislator may be guaranteed a stream of campaign resources conditional on participating in a non-partisan voting bloc with respect to issues relevant to the patron—an appealing trade-off.

8 Conclusion

The behavior predicted from this fact pattern essentially corresponds to what we observe in our data. With few exceptions, House members' votes align with the party consensus, which itself aligns with aggregated financial interests that maintain and strengthen party blocs with a predictable platform content roughly matching the investors' interests. In the Senate, both legislators and investors may develop long-run strategies of interdependence that are relatively low-risk (on the investor side) due to the length of Senate terms, and that substantially relieve work burdens (on the legislator side) and allow a greater measure of independence from the party mechanism. The House-Senate asymmetry emerges as a result of this array of incentives.

It is worth mentioning that while the Investment Theory emphasizes the impact of organized industrial interests in party competition, non-partisan or cross-partisan voting blocs can be financed by other structural actors, including grassroots movements. Digital technologies reduce the costs of mass communication with voters as well as the costs associated with making recurring transactions, allowing legislators to operate from a base of investors whose interests and preferences more closely match those of the "median voter". Differentiating the systemic political effects of this brand of giving, in relation to that of organized industry, should be a focus of future quantitative research.

References

[1] Clio Andris, David Lee, Marcus J. Hamilton, Mauro Martino, Christian E. Gunning, and John Armistead Selden. The rise of partisanship and super-cooperators in the U.S. House of Representatives. *PLOS ONE*, 2015.

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, 2020. `https://arxiv.org/abs/2004.05150`.

[3] Adam Bonica. Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning. *American Journal of Political Science*, pages 830–848, 2018.

[4] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning structured embeddings of knowledge bases. In *AAAI*, 2011.

[5] J. Alexander Branham, Stuart N. Soroka, and Christopher Wlezien. When do the rich win? *Political Science Quarterly*, 132, 2017.

[6] Center for Responsive Politics. OpenSecrets. `https://www.opensecrets.org/`, 1983.

[7] Tyler Cowen. Why politics is stuck in the middle. *The New York Times*, February 6 2010.

[8] Gary W. Cox and Keith T. Poole. On measuring partisanship in roll-call voting: The U.S. House of Representatives, 1877-1999. *American Journal of Political Science*, 46:477–489, Jul. 2002.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*, pages 4171–4186, June 2019.

[10] Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern component, and representational similarity analysis. *PLOS Computational Biology*, 13, 2017.

[11] Carol Doherty, Jocelyn Kiley, and Bridget Johnson. The Public, the Political System, and American Democracy. *Pew Research Center*, April 2018.

[12] Anthony Downs. *An Economic Theory of Democracy*. Harper & Row, 1957.

[13] Benjamin Eidelson. The Majoritarian Filibuster. *The Yale Law Journal*, January 2013.

[14] Karl Evers-Hillstrom. Most expensive ever: 2020 election cost $14.4 billion. *OpenSecrets*, 2021.

[15] Brent Ferguson. Congressional disclosure of time spent fundraising. *Cornell Journal of Law and Public Policy*, 23:1–43, 2013.

[16] Thomas Ferguson. *Golden Rule: The Investment Theory of Party Competition and the Logic of Money-Driven Political Systems*. The University of Chicago Press, 1995.

[17] Thomas Ferguson, Paul Jorgensen, and Jie Chen. How money drives us congressional elections: Linear models of money and outcomes. *Structural Change and Economic Dynamics*, 2019.

[18] Thomas Ferguson, Paul Jorgensen, and Jie Chen. How much can the U.S. Congress resist political money? A quantitative assessment. *INET Working Paper 109*, Jan. 2020.

[19] Thomas Ferugson, Paul Jorgensen, and Jie Chen. The Knife Edge Election of 2020: American politics between Washington, Kabul, and Weimar. *INET Working Paper 169*, Nov. 2021.

[20] Catherine Fisk and Erwin Chemerinsky. The Filibuster. *Stanford Law Review*, 49:181–254, 1997.

[21] Patrick Flavin. Campaign finance laws, policy outcomes, and political equality in the American states. *Political Research Quarterly*, 68:77–88, March 2015.

[22] Brian Frederick. Gender patterns of roll call voting in the U.S. Senate. *Congress & the Presidency*, 37:103–104, 2010.

[23] Brian Frederick. Gender and roll call voting behavior in congress: A cross-chamber analysis. *American Review of Politics*, 34:1–20, 2013.

[24] Sean M. Gerrish and David M. Blei. Predicting legislative roll calls from text. In *ICML 11*, pages 489–496, June 2011.

[25] Margin Gilens and Benjamin I. Page. Testing theories of American politics: Elites, interest groups, and average citizens. *Perspectives on Politics*, 12:564–581, 2014.

[26] Matthew H. Goldberg, Jennifer R. Marlon, Xinran Wang, Sander van der Linden, and Anthony Leiserowitz. Oil and gas companies invest in legislators that vote against the environment. *PNAS*, 117:5111–5112, March 2020.

[27] Ryan Grim. Dark-money group to donors: Reconciliation bill can still be killed. *The Intercept*, September 2021.

[28] Eric Heberlig, Marc Hetherington, and Bruce Larson. The price of leadership: Campaign money and the polarization of congressional parties. *Journal of Politics*, 68:992–1005, Nov. 2006.

[29] House Committee on Standards of Official Conduct. House Ethics Manual, 2008. 111th Congress of the United States. `https://ethics.house.gov/sites/ethics.house.gov/files/documents/2008_House_Ethics_Manual.pdf`.

[30] Gary C. Jacobson. Campaign spending effects in U.S. Senate elections: Evidence from the National Annenberg Election Survey. *Electoral Studies*, 25:195–226, 2006.

[31] Gary C. Jacobson. Measuring campaign spending effects in U.S. House elections. In Henry Brady and Richard Johnson, editors, *Capturing Campaign Finance Effects*, pages 199–220. University of Michigan Press, 2006.

[32] Shannon Jenkins. How gender influences roll call voting. *Social Science Quarterly*, 93:415–433, June 2012.

[33] Hamid Karimi, Tyler Derr, Aaron Brookhouse, and Jiliang Tang. Multi-factor congressional vote prediction. In *IEEE*, 2019.

[34] James B. Kau and P.H. Rubin. *Congressman, Constituents, and Contributors: Determinants of Roll Call Voting in the House of Representatives*. Springer Science & Business Media, Nov. 2013.

[35] Jonathan Wayne Korn and Mark A. Newman. A deep learning model to predict congressional roll call votes from legislative texts. *Machine Learning and Applications*, 7:15–27, December 2020.

[36] Anastassia Kornilova, Daniel Argyle, and Vladimir Eidelman. Party Matters: Enhancing legislative embeddings with author attributes for vote prediction. In *ACL 56*, pages 510–515, 2018.

[37] Peter E. Kraft, Hirsh Jain, and Alexander M. Rush. An embedding model for predicting roll-call votes. In *Proceedings of EMNLP*, pages 2066–2070, 2016.

[38] Nikolaus Kriegeskorte, Elia Formisano, Bettina Sorger, and Rainer Goebel. Individual faces elicit distinct response patterns in human anterior temporal cortex. *PNAS*, 51:20600–20605, 2007.

[39] Matthias Lalisse. Sinema and Manchin Flush With Lobbyist Contributions as They Hold Up Biden Agenda. *Data for Progress*, October 2021.

[40] Jeffrey B. Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. VoteView: Congressional Roll-Call Votes Database. `https://voteview.com/`, 2021.

[41] Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117:30046 – 30054, 2020.

[42] MapLight. Federal Money and Politics Data Set. `https://maplightarchive.org/`, 2005. Retrieved in Nov. 2021 from Bulk Campaign Finance database.

[43] John J. Nay. Predicting and understanding law-making with word vectors and an ensemble model. *PLOS ONE*, May 2017.

[44] Fox News Network. Presidential Election Exit Polls. https://www.foxnews.com/elections/2020/general-results/voter-analysis, Nov. 2020. Retrieved Dec. 6, 2021.

[45] Norah O'Donnell. Are members of Congress becoming telemarketers? *CBS News*, April 2016.

[46] Spencer Overton. The Donor Class: Campaign finance, democracy, and participation. *University of Pennsylvania Law Review*, 153:73–118, Nov. 2004.

[47] Pallavi Patil, Kriti Myer, Ronak Zala, Arpit Singh, Sheshera Mysore, Andre McCallum, Adrian Benton, and Amanda Stent. Roll call vote prediction with knowledge augmented models. In *Proceedings of ACL 23*, pages 574–581, Nov. 2012.

[48] Kathryn Pearson. *Party Discipline in the U.S. House of Representatives*. University of Michigan Press, 2015.

[49] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543. Association for Computational Linguistics, October 2014.

[50] Thomas Piketty. *Capital and Ideology*. Harvard UP, 2020.

[51] ProPublica. ProPublica Congress API. https://projects.propublica.org/api-docs/congress-api/, 2016.

[52] ProPublica, MapLight, GovTrack, and FiveThirtyEight. congress-legislators database. https://github.com/unitedstates/congress-legislators, 2013.

[53] John P. Sarbanes. H.R.1 - For the People Act of 2021, January 2021. 117th Congress of the United States. https://www.congress.gov/bill/117th-congress/house-bill/1/text.

[54] Kurt Schrader. H.J.Res.21 - Proposing an amendment to the Constitution of the United States giving Congress power to regulate campaign contributions for Federal elections, January 2021. 117th Congress of the

United States. https://www.congress.gov/bill/117th-congress/house-joint-resolution/21?s=1&r=4.

[55] Dennis M. Simon and Barbara Palmer. The roll call behavior of men and women in the U.S. House of Representatives, 1937-2008. *Politics & Gender*, pages 225–246, 2010.

[56] Samuel Smith, Jae Yeon Baek, Zhaoyi Kang, Dawn Song, Laurent El Ghaoui, and Mario Frank. Predicting Congressional Votes Based on Campaign Finance Data. In *ICML 11*, 2012.

[57] James M. Snyder, Jr. and Tim Groseclose. Estimating party influence in Congressional roll-call voting. *American Journal of Political Science*, 44:193–211, April 2000.

[58] Nicholas O. Stephanopoulos. Aligning campaign finance law. *Virginia Law Review*, 101, Sept. 2015.

[59] Joseph Stiglitz. *The Price of Inequality*. W.W. Norton & Co., 2012.

[60] Arturo Vega and Juanita M. Firestone. The effects of gender on Congressional behavior and the substantive representation of women. *Legislative Studies Quarterly*, 20:213–222, May 1995.

[61] Abby K. Wood and Christian R. Grose. Campaign finance transparency affects legislators' election outcomes and behavior. *American Journal of Political Science*, 2021.

[62] Yuqiao Yang, Xiaoqiang Lin, Gent Lin, Zengfeng Huang, Changjian Jiang, and Zhongyu Wei. Joint representation learning of legislator and legislation for roll call prediction. In *Proceedings of IJCAI-29*, 2020.

9 Appendix

9.1 Materials and Methods

Data processing. Metadata for legislators was retrieved from the `congress-legislators` database (`https://github.com/unitedstates/congress-legislators`), using custom routines to match FEC and BioGuide identifiers for congressmembers. Models (PCA and RF classifiers) were drawn from the `scikit-learn` v0.23.2 toolbox. All data processing was done in `python` v3.7.9 using `pandas` v1.1.1, using `pingouin` v0.5.0 to compute true and permuted Spearman/partial Spearman values (Section 6).

RSA metrics. To measure the distance between legislators' voting profiles in a given congressional session, we used the Manhattan metric ($\mathcal{L}_1$ norm). Let $\boldsymbol{v}^a, \boldsymbol{v}^b$ denote the vote vectors for legislators a and b, with components v_i^a and v_i^b. The Manhattan distance between the two legislators' voting profiles is:

$$\text{ManhattanDist}(\boldsymbol{v}^a, \boldsymbol{v}^b) = \sum_i |v_i^a - v_i^b| \tag{1}$$

where $|x|$ is the absolute value of scalar x. The distances between `LEGFIN` vectors were computed using the cosine measure. For `LEGFIN` vectors $\boldsymbol{u}^a, \boldsymbol{u}^b$, the distance is:

$$\text{CosineDist}(\boldsymbol{u}^a, \boldsymbol{u}^b) = 1 - \frac{\boldsymbol{u}^a \cdot \boldsymbol{u}^b}{\|\boldsymbol{u}^a\|_2 \left\|\boldsymbol{u}^b\right\|_2} \tag{2}$$

where $\boldsymbol{x} \cdot \boldsymbol{y}$ is the dot product between vectors $\boldsymbol{x}, \boldsymbol{y}$ and $\|\boldsymbol{x}\|_2$ the Euclidean norm. When $\boldsymbol{x}$ and $\boldsymbol{y}$ are collinear with a positive constant of proportionality, the cosine distance takes a value of 0. The maximum value is 2, and occurs when the vectors are collinear with a negative proportionality constant, indicating perfect anticorrelation.

9.2 Extended vote prediction results

Table 3 includes results for ten-fold *cross-validation by legislator* in which votes are partitioned with non-overlapping sets of *legislators*. The training set therefore includes the remaining legislators' votes on bills seen in the test set. This is provided for comparison with [3, 37, 47]. We also provide results for models where the 50d `LEGSTATE` indicator features (a one-hot vector for each American state), which made a small difference in vote prediction accuracy relative to the main variables of interest: political party and fundraising sources.

9.3 Review of prior results on roll call prediction

To allow comparison of our results with prior work, we provide a summary of every roll call prediction paper we could find and map their evaluation (data subsets, cross-validation schemes, etc.) into the closest-matching task setting among our three cross-validation setups. We believe this list of prior results is fully comprehensive.

9.3.1 Methods using campaign finance features

Bonica (2018) [3] explores an array of classifiers to do vote prediction on the congressional roll calls dating back 40 years. The method is welded onto classical "ideal point" models in political science, which are measures of ideological affiliation derived from multidimensional scaling applied to voting records. Taking "ideal point" DW-NOMINATE projections of legislators' voting records to 2 dimensions, Bonica uses supervised regression classifiers to learn mappings from campaign finance features to legislators' "ideal points". Vote prediction is then done by estimating a roll-call-wise "optimal cutting line" to separate the 2d projection of the `Yeas` from the 2d projection of the `Nays`. He reports a maximum of 89.5% (House) and 88.2% (Senate) accuracy for the best supervised classifier (Random Forest Regression + decision boundary), in ten-fold cross-validation with an underspecified partitioning scheme and a dataset that contains all votes, including procedurals. Because of the methodological differences as well as Bonica's concern with mapping

		Accuracy								
		Cross-val by bill			Cross-val by Congress			Cross-val by legislator		
	Model specification	House	Senate	Total	House	Senate	Total	House	Senate	Total
Baselines	MAJORITY_CLASS	.599	.772	.605	.599	.772	.605	.599	.772	.605
	BILLVEC	.599	.767	.605	.599	.772	.605	.599	.770	.605
Excludes financials	LEGPRTY+BILLVEC	.816	.700	.812	.714	.703	.713	.961	.850	.957
	LEGPRTY+SPONPRTY	.926	.646	.916	.925	.648	.916	.926	.646	.916
	LEGPRTY+LEGSTATE+SPONPRTY	.926	.670	.917	.925	.666	.916	.918	.665	.909
	LEGPRTY+LEGSTATE+BILLVEC+SPONPRTY	.924	.717	.917	.923	.688	.915	.958	.853	.955
	LEGPRTY+BILLVEC+SPONPRTY	.924	.719	.917	.921	.689	.913	.961	.850	.957
Includes financials	LEGFIN+BILLVEC	.845	.832	.845	.734	.769	.735	.942	.904	.940
	LEGFIN+SPONPRTY	.927	.777	.922	.896	.772	.892	.893	.768	.889
	LEGFIN+LEGSTATE+SPONPRTY	.927	.777	.922	.896	.772	.892	.891	.770	.887
	LEGFIN+LEGSTATE+BILLVEC+SPONPRTY	**.933**	**.866**	**.931**	.926	**.821**	.922	.949	.907	.947
	LEGFIN+BILLVEC+SPONPRTY	**.932**	**.863**	**.930**	**.927**	**.816**	.923	.949	.906	.948
	FULL_MODEL	**.932**	.859	**.930**	**.927**	.812	**.924**	**.963**	.909	**.961**
	LEGSTATE+FULL_MODEL	**.933**	.865	**.931**	**.927**	.818	**.927**	.961	**.910**	.960

Table 5: **Vote prediction results**. Results on roll call prediction in the two cross-validation regimes from the main text, plus cross-validation by legislator to make these comparable to some results drawn from the literature review. Top results from the main text are **bolded**, and top results overall when LEGSTATE is included are additionally **underlined**, with the results from Table 3 highlighted in a lighter shade if they were affected. The FULL_MODEL refers to the specification with the main variable sets included (LEGFIN+LEGPRTY+BILLVEC+SPONPRTY). The best results without including financial features are highlighted in red.

classifier results to vote prediction through a theoretically loaded measure of legislator ideology (2d DW-NOMINATE projections), it is difficult to align this set of experiments with our procedure. However, since estimating the bill-wise decision boundaries requires test bills to appear in the training sample, the closest fit is *cross-val by legislator*.

Smith et. al. (2012) [56] use machine learning to predict roll call from 16d to 397d campaign finance summary vectors with individual donors collapsed into 16 industrial categories (or 397 subsector categories) as annotated by Maplight [42] and post-processed with PCA. Cross-validation is *by legislator* with a 70%/30% train-test split, using an array of classifiers (kNN and an SVM) each having hard convexity priors. Since they do not report the actual accuracy numbers for each method, they are omitted from Table 6.

9.3.2 Methods using bill text embeddings

Gerrish et. al. (2011) [24] combine textual features with a Bayesian ideal point model of each legislator fit to a subset of their votes. For each legislator, the ideal point model generates the probability that the legislator will vote for the bill on the basis of the reduced-dimension legislator projection and a decision boundary specific to each bill. Bills are projected onto the same space as legislators to yield a decision boundary, with a bag-of-words (BOW) text model allowing decision boundaries for new bills to be generated. Topic modeling of n-gram distributions in legislative text yields document embeddings, which are combined with the ideal-point projections of legislators. Gerrish et. al.'s top model predicts votes with 89.7% accuracy in 6-fold bill-held-out cross-validation within each congressional session (*cross-val by bill*).

In Kornilova et. al. (2018) [36], an ideal point model of each legislator is augmented with metadata about the bill, such as its cosponsors' party split (the proportion of cosponsors from each party). Text embeddings are additionally provided via a convolutional neural network (CNN) architecture, or alternatively with a BOW model. Peak performance is 86.21% for in-session cross-validation and 77.3% in leave-one-congress-out cross-validation averaged across testing folds (*cross-val by bill* and *cross-val by Congress*).

Korn and Newman (2020) [35] develop a custom text embedding architecture for bill out-

come prediction, meaning prediction of whether a bill will pass or fail taking into account not just the probability of passing a roll call, but the probability that the bill dies in committee or otherwise does not come to a floor vote. The architecture is based on a combination of long short-term memory (LSTM) and convolutional neural network (CNN) layers, and outputs the predicted number of `Yea` vs. `Nay` votes, from which a bill is predicted to pass or fail. To cope with the fact that only 3.6% of congressional bills pass roll call [43], the authors sample a balanced set. They report 67.3% accuracy across folds at a chance level of 50% after the rebalancing. Since they do not predict individual votes, but only bill outcomes, their results are not comparable to roll call prediction.

Karimi et. al. (2019) [33] deploy a multi-factor deep learning model that combines ideological features of legislators (both voters and sponsors) gleaned from Wikipedia text and samples of legislators' Tweets, as well as more conventional features such as past voting records plus party and bill embeddings. Train, valid, and test partitions are all within one Congress (the 113th), with the entire training set temporally preceding the test set in a forecasting approach. They report a maximum roll call prediction accuracy of 77% (*cross-validation by bill*).

Yang et. al. (2020) [62] jointly learn latent features for legislators and legislation using an array of neural network architectures. Directly learned legislator embeddings with party/state metadata are combined with an embedding of the legislator "network" derived from cosponsorships (legislators are linked if they cosponsor the same legislation). A number of graph embedding architectures are explored, with legislation encoded using an LSTM. In model evaluation, the authors take all legislation from 1993 to 2018 with separate evaluation tasks obtained with a 5-year sliding window. The model is trained on the first four years of each slice and tested on the fifth year, in an incremental forecasting approach. They report accuracy averaged across the 22 slices: 81.86%. The best match among our procedures is *cross-validation by Congress*.

Kraft et. al. (2016) [37] use a BOW model initialized with `GloVe` vectors [49] combined with directly learned legislator embeddings. The cross-validation is five-fold within each of the 106th to 111th Congresses (1999-2011) without partitioning out any group (e.g. bills or legislators). The closest match is *cross-val by legislator*. Their top-performing model

has 90.6% accuracy (though a reimplementation by Patil et. al. [47] finds only 88.5% with a tweaked evaluation regime). Since the embeddings are atomic rather than compositional, they can only evaluate on the subset of legislators who occurred in one of the congresses in the training set.

9.3.3 Other methods

In Patil et. al. (2019) [47], the authors introduce two external resources: news text about legislators, and a structured knowledge base in which legislators appear as entities. The news text model is a BOW of fine-tuned `GloVe` vectors for frequent unigrams from articles mentioning legislators. This is further augmented with learned politician embeddings trained on the Freebase knowledge graph with the graph completion objective [4]. Evaluation is in four Congresses (106th to 109th), with each model evaluated within-Congress with 60%/20%/20% train/valid/test partitioning. Two cross-validation schemes are explored: one in which 5% of legislators are held out (*cross-val by legislator*), and one in which train and test are freely mixed. Best-model accuracy is 85.98% for CV-by-legislator, and 89.47% in the free-mixing cross-validation (we report both in comparison to CV-by-legislator as best-match in Table 6).

9.3.4 The present results in context

We advise caution in comparing results in this sparse literature due to the non-uniformity of methods and their sometimes incomplete reporting—especially with respect to cross-validation partitions. However, we have made our best effort to match each paper with the closest analogue among our cross-validation regimes. Paper-by-paper comparison with our results is provded in Table 6.

Table 6: **Top results from the literature in comparison with ours.** Results from our survey of the roll-call prediction literature with each reference's best-reported result as well as ours. Each paper is matched to its closest analogue among the three cross-validation regimes used in this article. See Section 9.3 in the above text for details about the computational methods used in each reference.

Reference	**Best-matching CV scheme**	**Best result (reference)**	**Best result (ours)**
Bonica 2018 [3]	Cross-val by legislator	89.5% (House)	96.3% (House)
		88.2% (Senate)	91% (Senate)
Gerrish et. al. 2011 [24]	Cross-val by bill	89.7%	93.1%
Kornilova et. al. 2018 [36]	Cross-val by bill	86.21%	93.1%
	Cross-val by Congress	77.3%	92.7%
Karimi et. al. 2019 [33]	Cross-val by bill	77%	93.1%
Yang et. al. 2020 [62]	Cross-val by Congress	81.9%	92.7%
Kraft et. al. 2016 [37]	Cross-val by legislator	90.6%	96.1%
Patil et. al. 2019 [47]	Cross-val by legislator	85.6%/89.5%	96.1%

	House				Senate			
Congress	ρ	$\rho_{\text{part}}(\texttt{LEGSTATE})$	$\rho_{\text{part}}(\texttt{LEGPRTY})$	ρ_{part}	ρ	$\rho_{\text{part}}(\texttt{LEGSTATE})$	$\rho_{\text{part}}(\texttt{LEGPRTY})$	ρ_{part}
113th	.476	.476	.047	.047	.548	.548	.291	.291
114th	.529	.528	.057	.057	.679	.677	.348	.347
115th	.534	.534	.075	.076	.681	.680	.180	.179
116th	.523	.520	.059	.056	.670	.669	.275	.274
117th	.419	.417	-.007	-.009	.635	.632	.269	.269

Table 7: Spearman and partial Spearman correlations separating out the effect of the LEGPRTY and LEGSTATE variables from the RSA (cf. Table 4 in the main text). $\rho_{\text{part}}(\texttt{LEGSTATE})$: RDM correlations with the same-state indicator as a covariate only; $\rho_{\text{part}}(\texttt{LEGPRTY})$: RDM correlations with the same-party indicator as a covariate only. ρ_{part}: RDM correlations with both the same-state and same-party indicators as covariates.

9.4 Replication of the analyses using the Center for Responsive Politics dataset

To check that the results were robust to small changes in the data annotations, we replicated the analyses initially conducted using the MapLight data tables. Bulk data tables were obtained from the Center for Responsive Politics (CRP) for the 2010-2020 election cycles, and we repeated the data processing protocol from the original analysis: removing individual contributions and transfers from party organizations, filtering out legislators who clocked in less than $100,000 in donations, and keeping only donors who gave at least $1,000 to a legislator who served. A PCA model was fit to the resulting inventory of donors and legislators, and we repeated the roll call prediction and RSA experiments using the `LEGFIN` vectors obtained from the new data.

9.4.1 Results

Roll call prediction. The replication corroborates every main result. The best accuracy across the board is obtained when including financial features, with top accuracy of .931/.923/.963 in cross-validation by bill/congress/legislator (see Table 8). The effect of including `LEGFIN` is greater in the Senate than in the House, with a sizable improvement (13.5%/11.8%/6%) in the Senate within each cross-validation regime when comparing the best models with and without financial features.

RSA. Table 9 reports the Spearman and partial Spearman correlations between legislators' vote and financial profiles|a replication of Table 7. Vote and financial similarities were highly correlated in both chambers across all sessions. After controlling for party and state, we found small positive correlations in all but one session (the 117th) in the House, and strong positive correlations in the Senate, both of about the same magnitude as in the original analysis.

This provides a complete replication of all of our main findings: the significance of campaign finance contributions as a factor in predicting votes and in structuring congressional voting blocs, as well as the House-Senate asymmetry.

		Accuracy								
		Cross-val by bill			Cross-val by Congress			Cross-val by legislator		
	Model specification	House	Senate	Total	House	Senate	Total	House	Senate	Total
Baselines	MAJORITY_CLASS	.599	.772	.605	.599	.772	.605	.599	.772	.605
	BILLVEC	.599	.767	.605	.599	.771	.605	.600	.771	.606
Excludes financials	LEGPRTY+BILLVEC	.822	.725	.819	.719	.701	.718	.961	.848	.958
	LEGPRTY+SPONPRTY	.926	.646	.917	.925	.649	.916	.926	.646	.917
	LEGPRTY+LEGSTATE+SPONPRTY	.926	.671	.917	.926	.667	.917	.917	.658	.908
	LEGPRTY+LEGSTATE+BILLVEC+SPONPRTY	.925	.728	.919	.925	.693	.917	.958	.852	.955
	LEGPRTY+BILLVEC+SPONPRTY	.925	.714	.918	.924	.689	.916	.961	.848	.958
Includes financials	LEGFIN+BILLVEC	.853	.840	.852	.733	.772	.734	.925	.903	.924
	LEGFIN+SPONPRTY	.927	.779	.922	.895	.775	.891	.885	.774	.882
	LEGFIN+LEGSTATE+SPONPRTY	.927	.779	.922	.895	.774	.891	.877	.776	.874
	LEGFIN+LEGSTATE+BILLVEC+SPONPRTY	**.934**	.862	**.931**	.919	**.819**	.916	.930	.904	.929
	LEGFIN+BILLVEC+SPONPRTY	**.934**	**.863**	**.931**	.922	.818	.919	.933	.904	.932
	FULL_MODEL	**.934**	**.863**	**.931**	**.927**	.813	**.923**	**.965**	**.912**	**.963**
	LEGSTATE+FULL_MODEL	**.934**	.862	**.931**	.926	.816	.922	**.965**	.910	**.963**

Table 8: **Vote prediction: Replication**. Replication of the vote prediction results using the Center for Responsive Politics (CRP) dataset. The FULL_MODEL refers to the specification with the main variable sets included (LEGFIN+LEGPRTY+BILLVEC+SPONPRTY). The best results without including financial features are highlighted in red. Results from the baseline models and those without financial features differ marginally from those in Table 5 due to stochasticity in the model initializations.

	House				Senate			
Congress	ρ	$\rho_{\text{part}}(\texttt{LEGSTATE})$	$\rho_{\text{part}}(\texttt{LEGPRTY})$	ρ_{part}	ρ	$\rho_{\text{part}}(\texttt{LEGSTATE})$	$\rho_{\text{part}}(\texttt{LEGPRTY})$	ρ_{part}
113th	.370	.370	.049	.049	.482	.481	.196	.195
114th	.404	.405	.044	.044	.622	.620	.241	.237
115th	.403	.406	.060	.062	.686	.684	.218	.216
116th	.375	.370	.067	.061	.638	.636	.224	.223
117th	.287	.283	.001	-.005	.581	.579	.206	.205

Table 9: Correlations from a replication of the RSA using data from the Center for Responsive Politics. Compare with Table 7. $\rho_{\text{part}}(\texttt{LEGSTATE})$: RDM correlations with the same-state indicator as a covariate only; $\rho_{\text{part}}(\texttt{LEGPRTY})$: RDM correlations with the same-party indicator as a covariate only. ρ_{part}: RDM correlations with both the same-state and same-party indicators as covariates.

9.5 RDMs for the five Congresses analyzed

Figures 5-14 display the Representational Dissimilarity Matrices (RDMs) for the five Congresses analyzed (113th-117th) in the Representational Similarity Analysis (RSA). The heatmap rows and columns are sorted according to the cosine similarity of row/column elements either by vote profile similarity (Figs. 5, 7, 9, 11, 13) or financial profile similarity (Figs. 6, 8, 10, 12, 14) using a nearest-neighbors hierarchical projection of the row/column identifiers onto one dimension (the axis).

113th Congress: sorted by voting profile

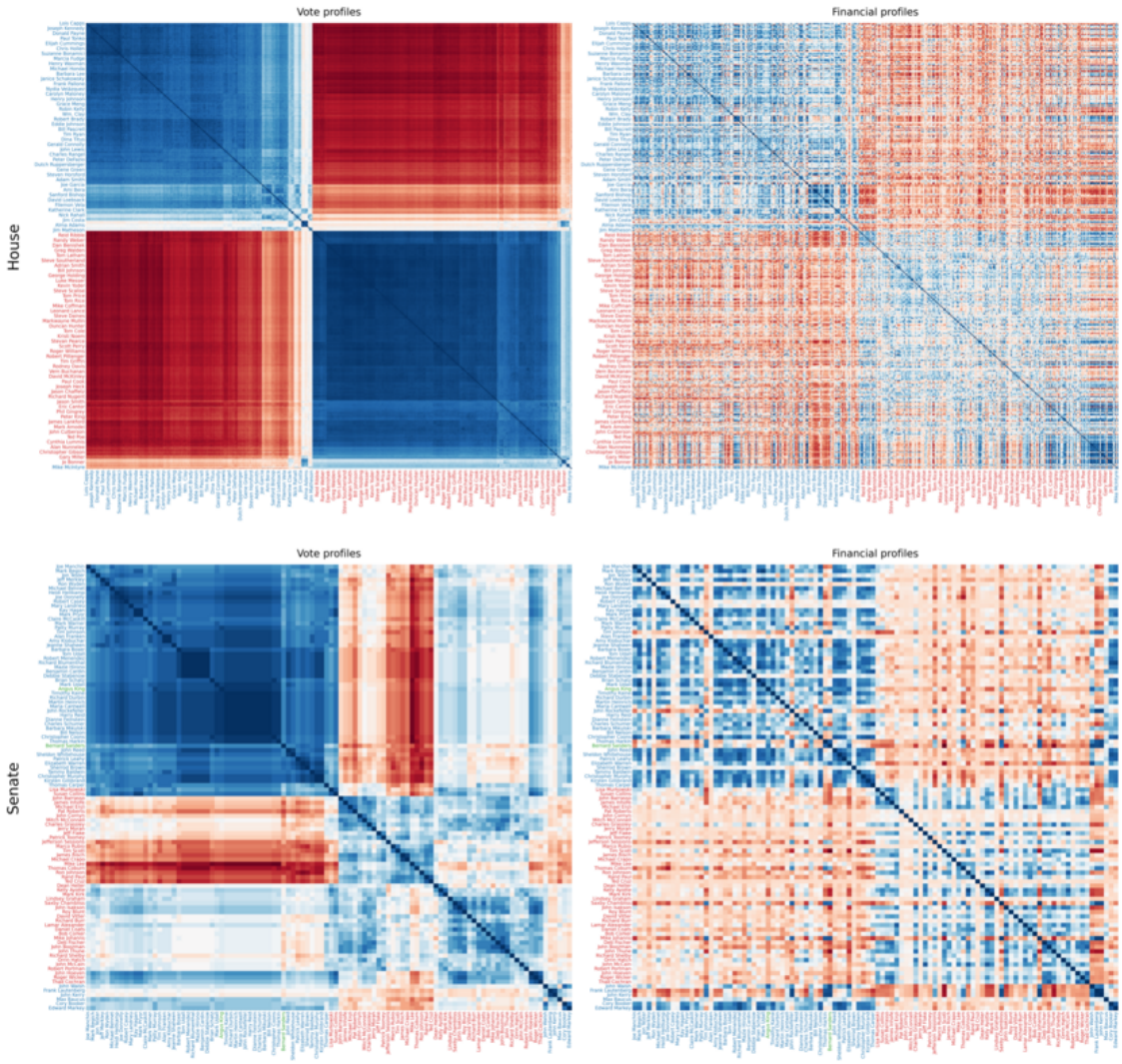

Figure 5: 113th Congress: Vote profile and legislative profile RDMs, row- and column-sorted according to vote profile similarity (voting blocs).

113th Congress: sorted by financial profile

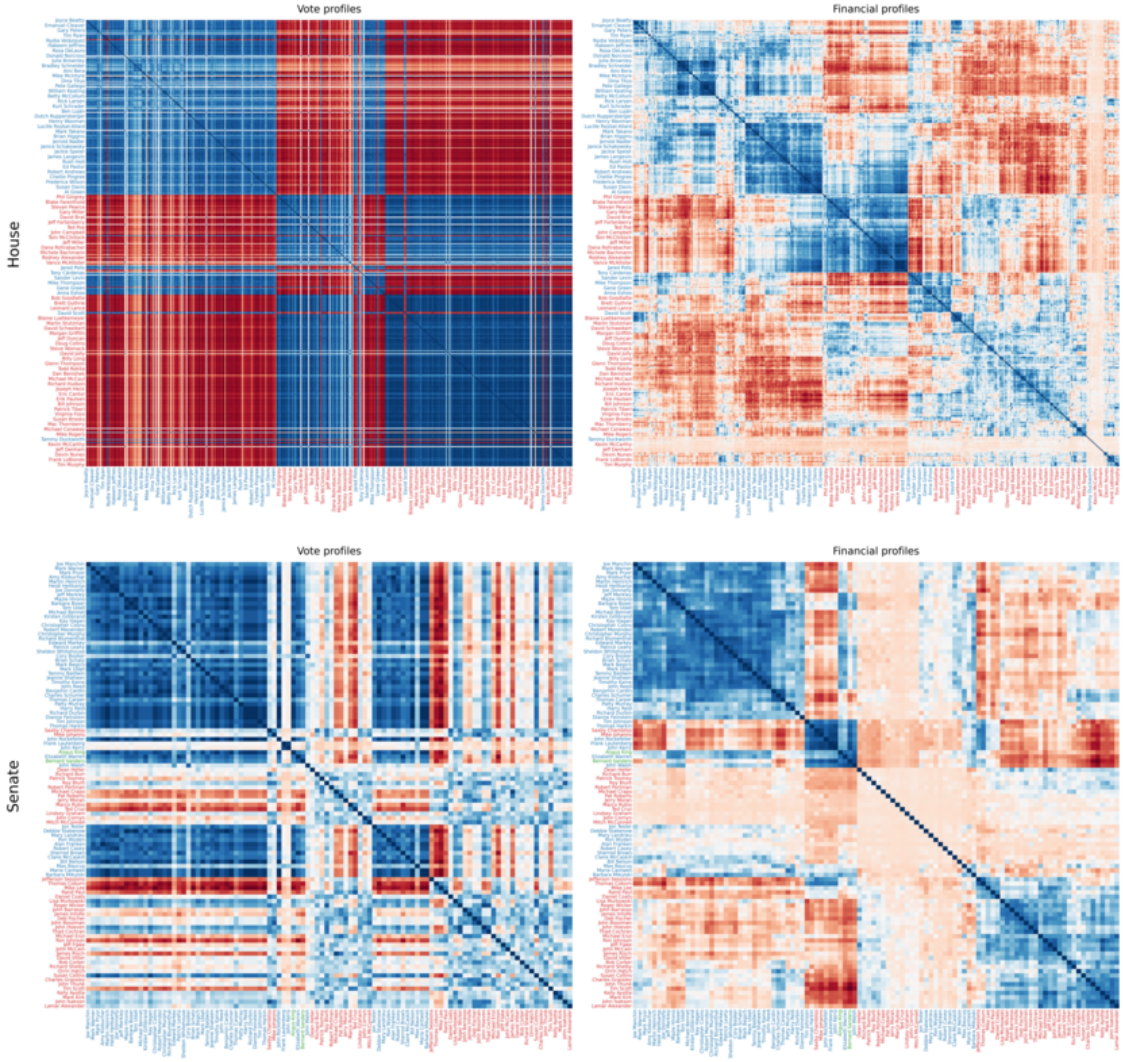

Figure 6: 113th Congress: Vote profile and legislative profile RDMs, row- and column-sorted according to financial profile similarity (fundraising blocs).

114th Congress: sorted by voting profile

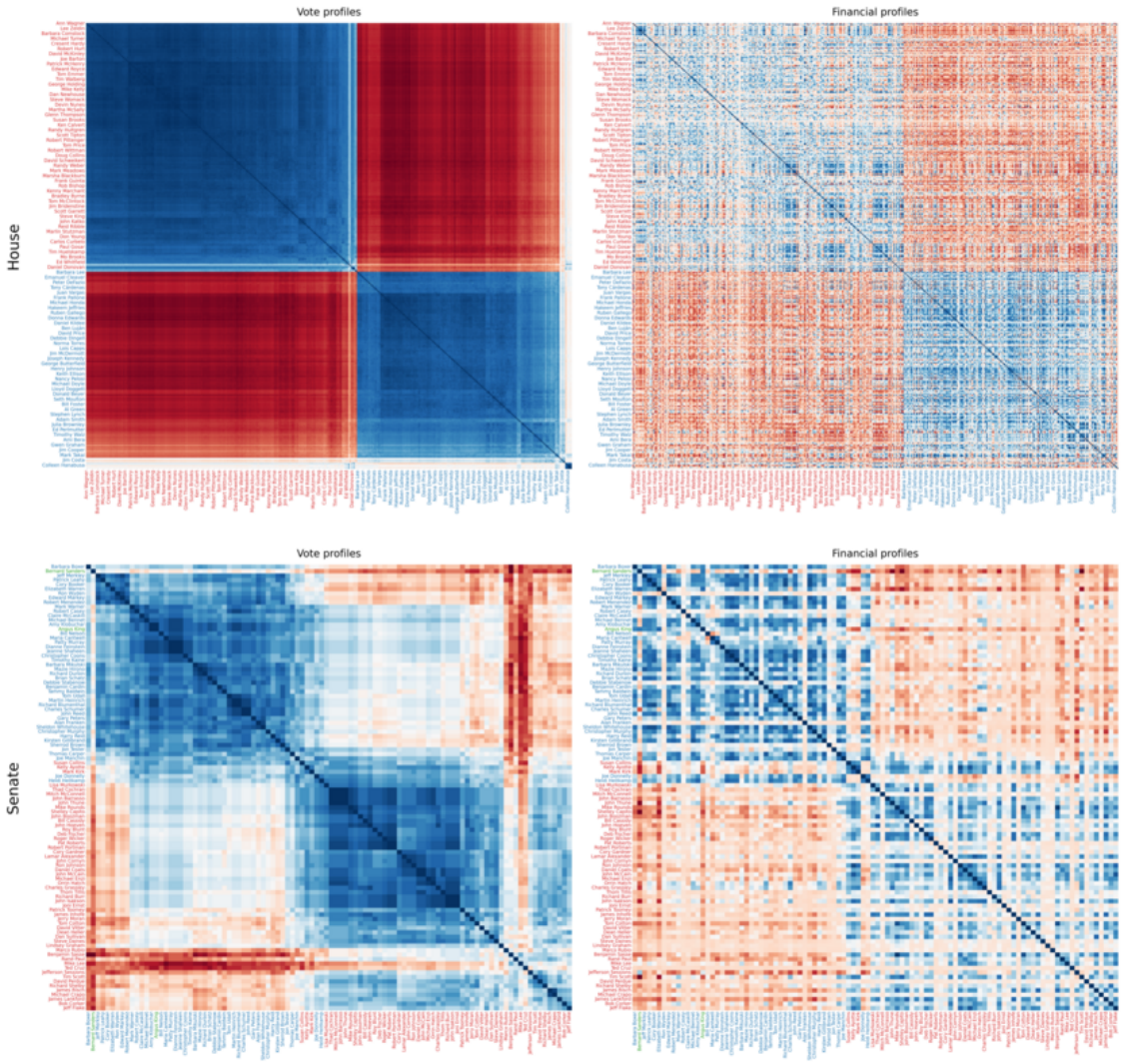

Figure 7: 114th Congress: Vote profile and legislative profile RDMs, row- and column-sorted according to vote profile similarity (voting blocs).

114th Congress: sorted by financial profile

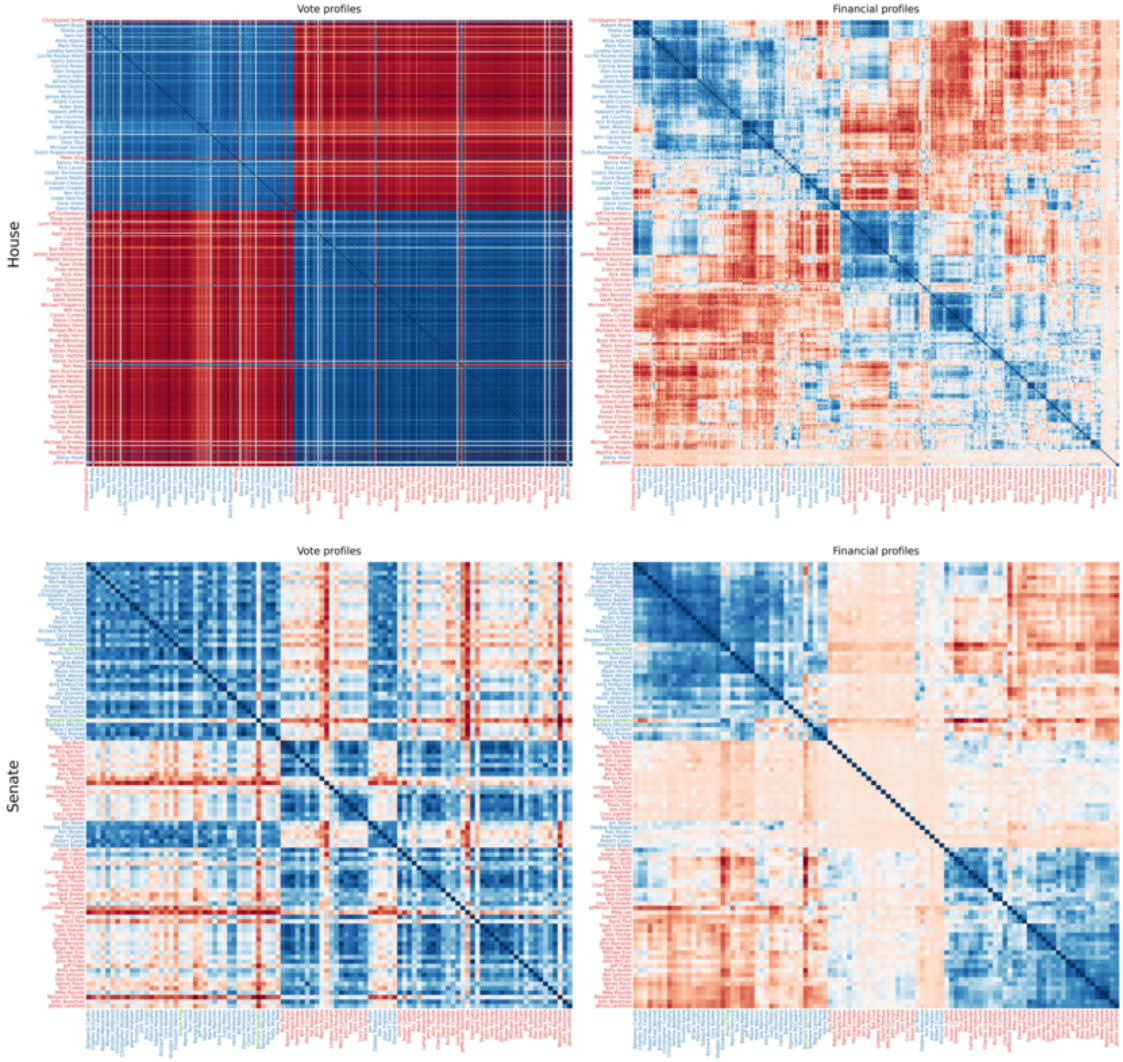

Figure 8: 114th Congress: Vote profile and legislative profile RDMs, row- and column-sorted according to financial profile similarity (fundraising blocs).

115th Congress: sorted by voting profile

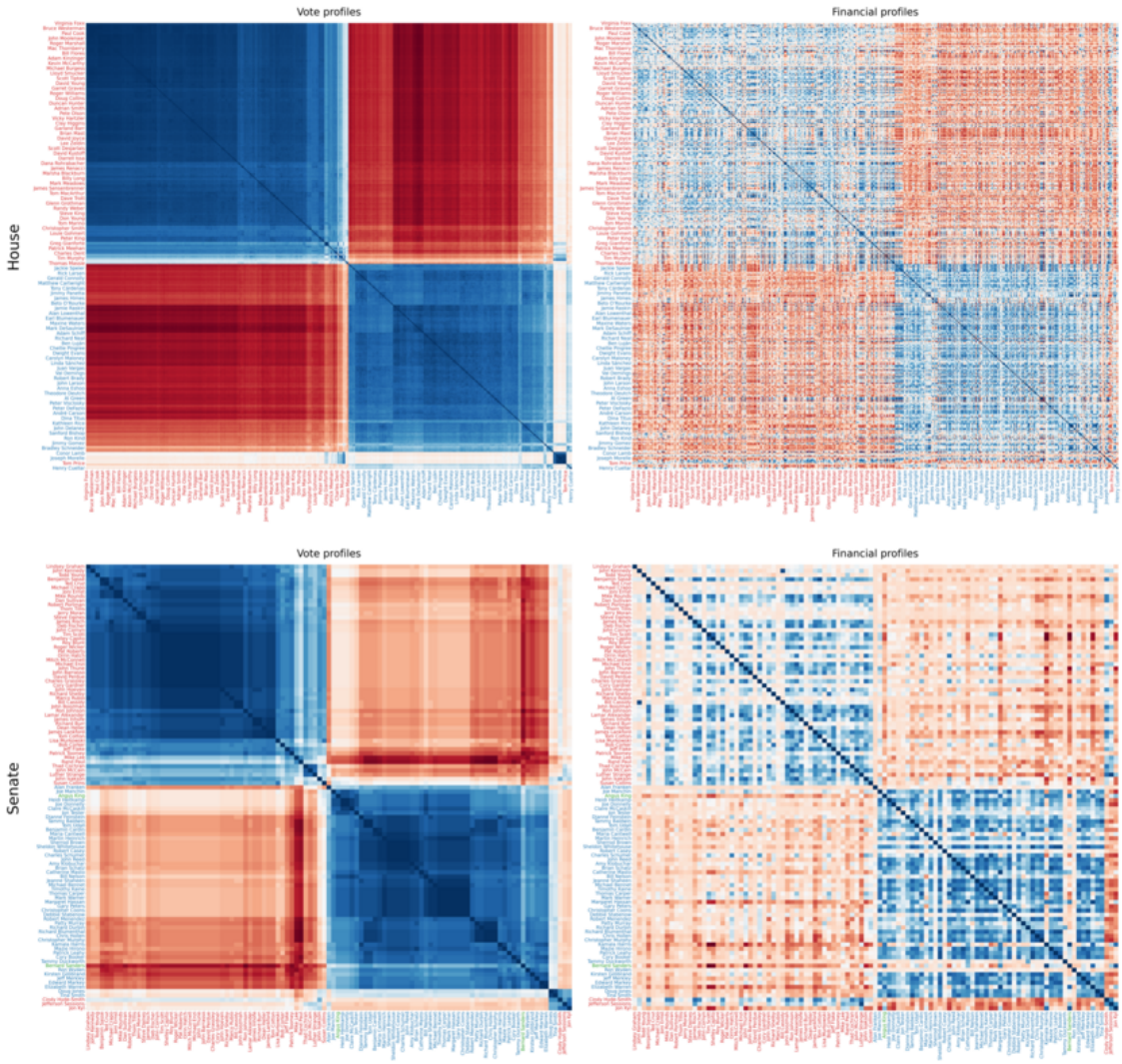

Figure 9: 115th Congress: Vote profile and legislative profile RDMs, row- and column-sorted according to vote profile similarity (voting blocs).

115th Congress: sorted by financial profile

Figure 10: 115th Congress: Vote profile and legislative profile RDMs, row- and column-sorted according to financial profile similarity (fundraising blocs).

116th Congress: sorted by voting profile

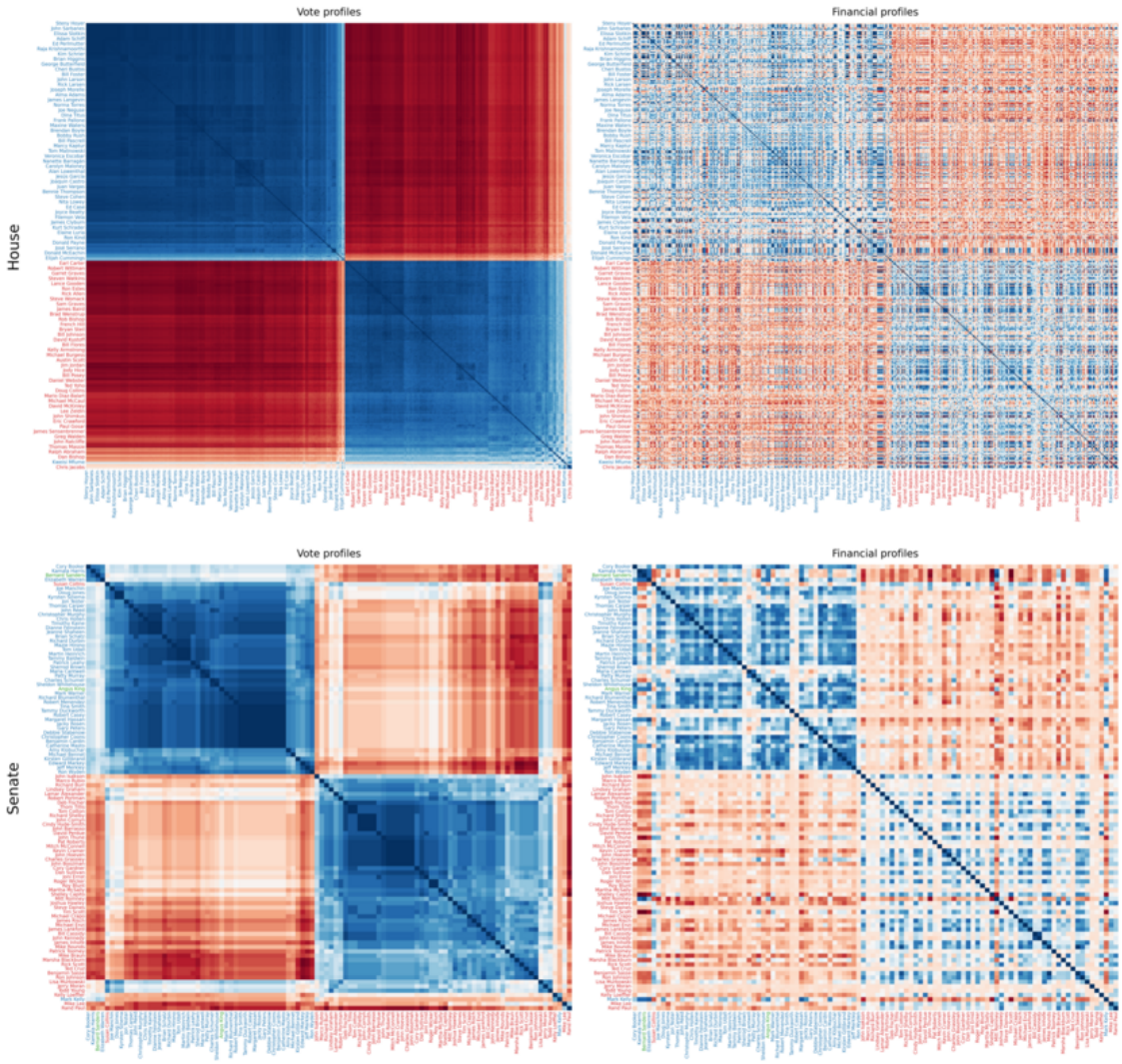

Figure 11: 116th Congress: Vote profile and legislative profile RDMs, row- and column-sorted according to vote profile similarity (voting blocs).

116th Congress: sorted by financial profile

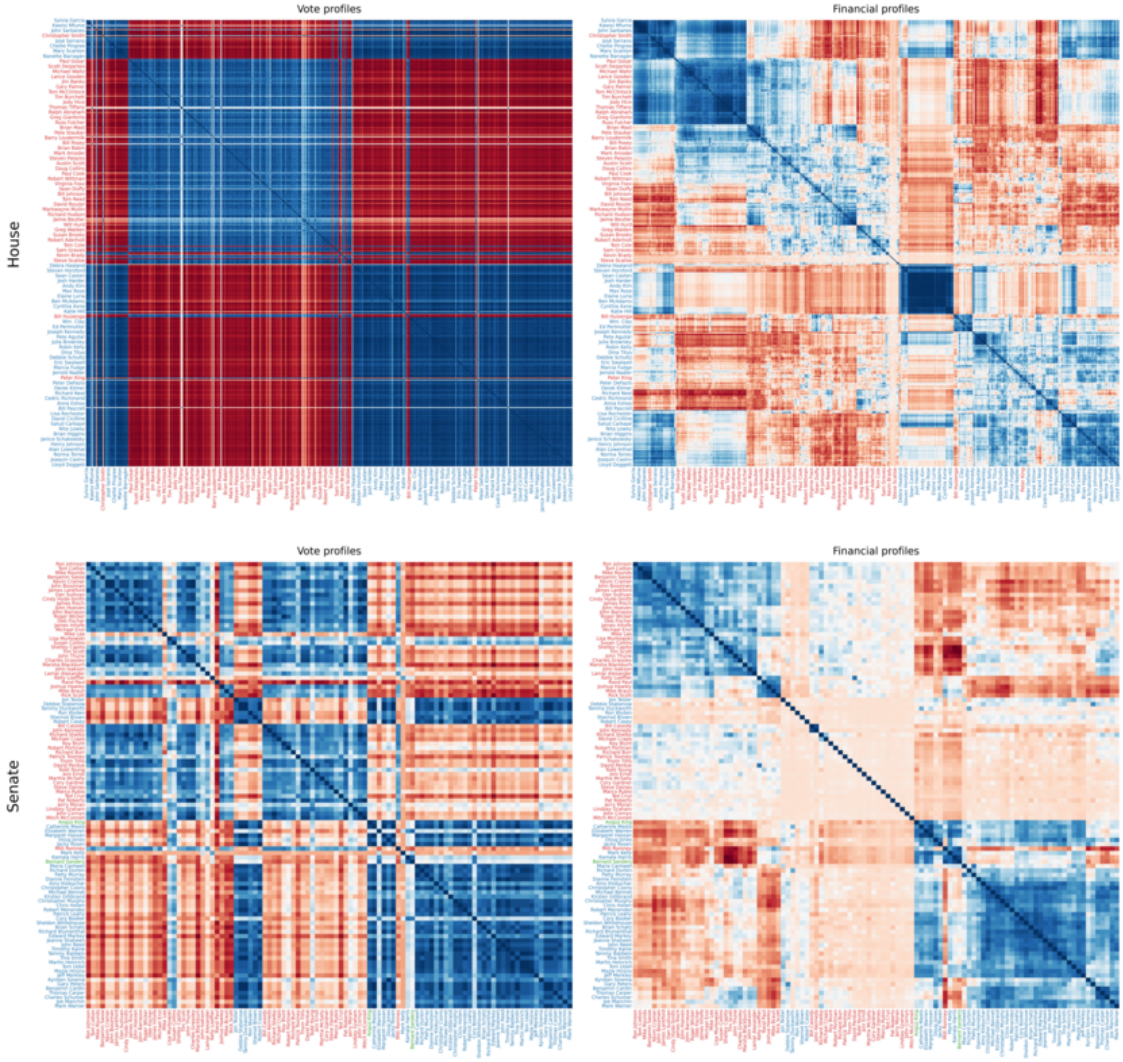

Figure 12: 116th Congress: Vote profile and legislative profile RDMs, row- and column-sorted according to financial profile similarity (fundraising blocs).

117th Congress: sorted by voting profile

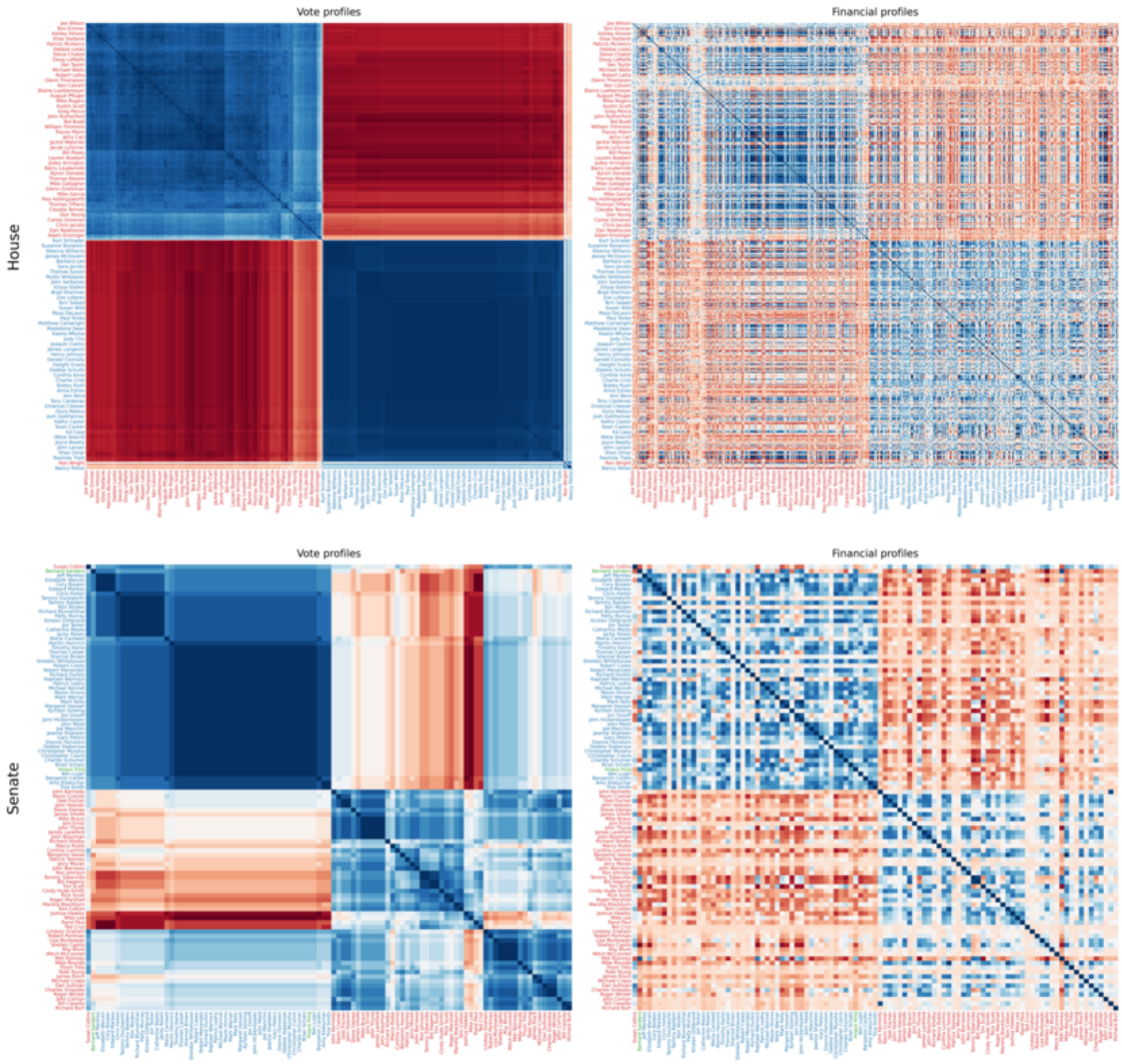

Figure 13: 117th Congress: Vote profile and legislative profile RDMs, row- and column-sorted according to vote profile similarity (voting blocs).

117th Congress: sorted by financial profile

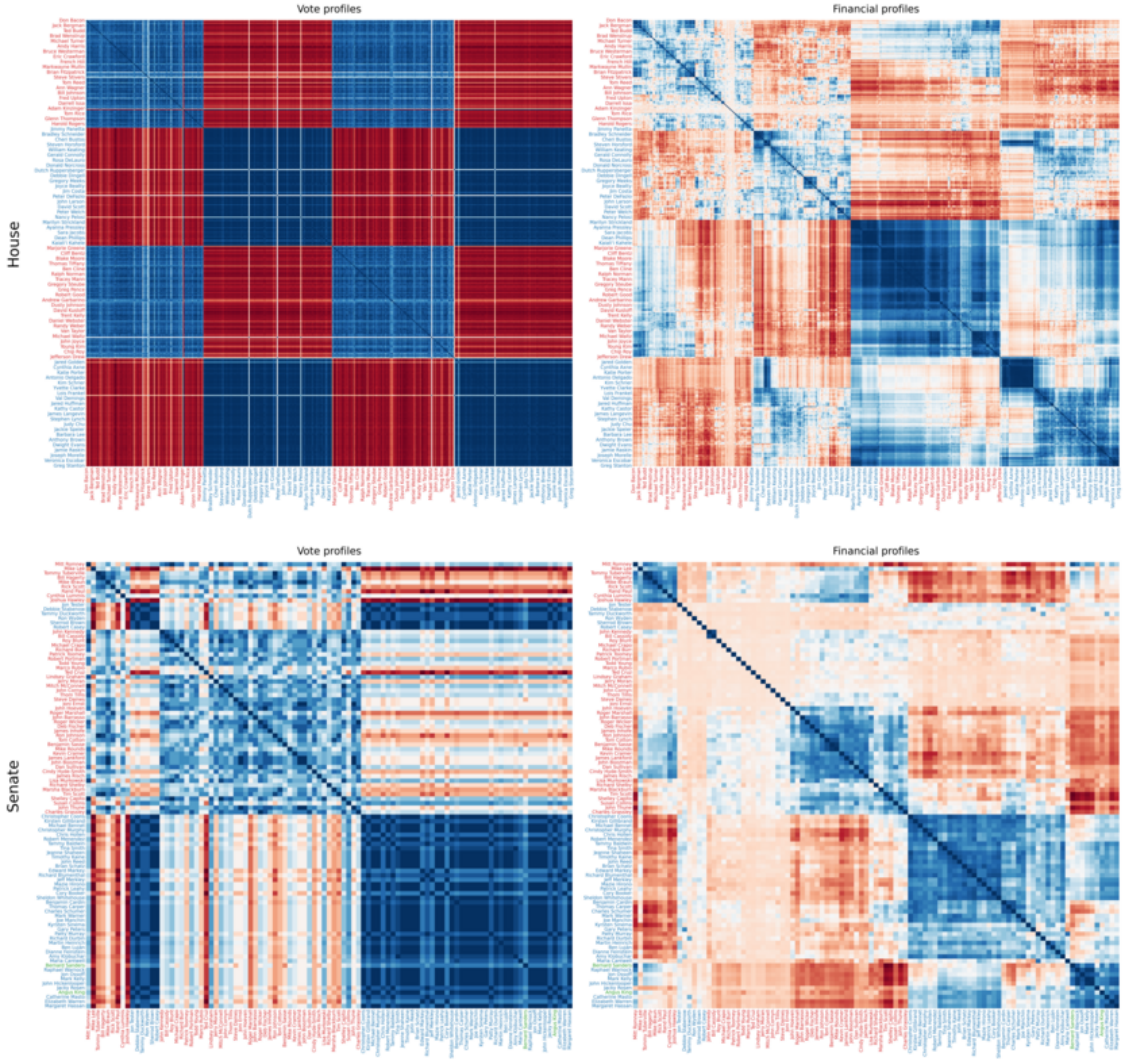

Figure 14: 117th Congress: Vote profile and legislative profile RDMs, row- and column-sorted according to financial profile similarity (fundraising blocs).

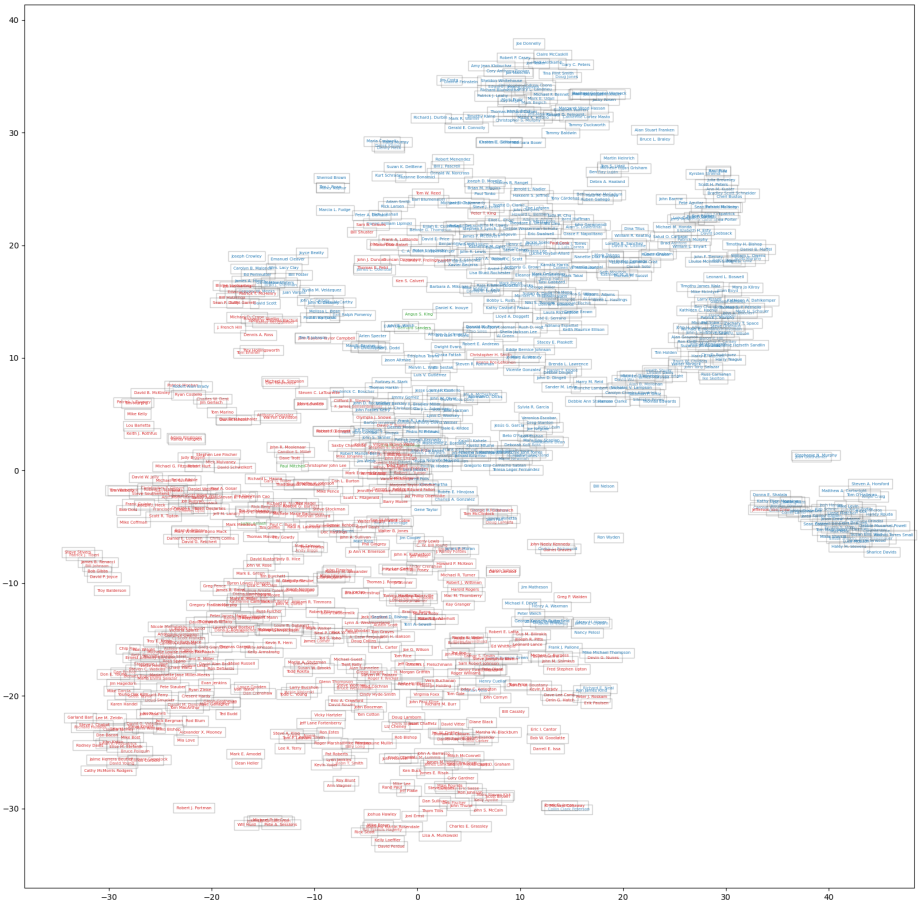

Figure 15: TSNE visualization of legislators' PCA `LEGFIN` vectors, annotated with name and party.