Analyzing The Network of Congress
CS224W Project Milestone

Henry Neeb, Taman Narayan, Christopher Kurrus

December 10, 2016

1 Abstract

Growing partisan polarization has dominated the conversation about Congress in recent years. While there have always been issue differences between the major parties, these differences have grown and calcified along multiple dimensions [8]. Quantitative metrics based on roll-call votes that label legislators as points in multidimensional ideological space have borne out the story of major polarization [7].

At the same time, the United States Congress is a rich social network where each legislator interacts with their peers through committees, house leadership positions, and bill co-sponsorships and amendments. The extent to which these relationships have changed in recent years is not as well understood as the changes in ideology. If ideologically distinct groups are still able to work together on crafting bills, and it turns out that those working relationships are important in legislative outcomes beyond the ideological content of bills, then partisan polarization in roll-call voting might be less worrying than it appears.

We build out a network using bill sponsorship-cosponsorship relationships to track the state of working relationships over time and the nature of legislative deal-making. Legislators may vote together relatively infrequently but still find some areas of common ground and work together on bills in those areas. They may also form communities of multiple legislators and influence which bills get taken up by committee and ultimately succeed. These and many more features of the working relationship network can give more fine-grained insight into Congressional activity, and how it has changed over time, than just the final votes.

In particular, we build a cosponsorship network of the 93^{rd} through 110^{th} Congress and supplement the network with information about each Congressman (node attributes). From this we are able to track broad patterns of how Congress has changed over time such as how many and and what kind of working relationships exist in Congress.

Next, we build stable clusters in the network that represent real-life communities of legislators based on their bill-writing activity. To do this, we implement the Community Detection in Networks with Node Attributes (CESNA) algorithm on the network and track communities and community attributes over time, noting how important partisianship was in defining each community.

Finally, we investigate to what extent the sponsor and early cosponsors of a bill can provide insight on the chances of a bill succeeding. Here, we draw upon not just various features of the working relationship network but also data from multiple sources characterizing the demographics and prominence of different legislators. We are interested not just in how predictable bill outcomes are but on what factors matter over time.

2 Related Work

2.1 Fowler

Fowler [5] presents a graph model for the 93^{rd} through 108^{th} Congress of the United States. He sought to infer the connectedness between legislators primarily through a network of bill cosponsorships. He claims that legislators cosponsoring each others bills is a signal for a relationship between legislators. Specifically, he organized his graph by partitioning by both Congress and house (Senate versus The House of Representatives), which created 32 $[(108-93+1)\times 2]$ distinct partitions. He primarily considered how connected legislators are by representing each legislator as a node with a directional edge between legislator A to legislator B if legislator A cosponsored a bill for legislator B.

Fowler's main analysis is tracking different centrality measurements overtime. He also constructs a network using his own metrics and tries to predict bill passage. We wish to use the data he devised to develop our own network representing the collaboration of Congress and track how collaboration has changed through time as Congress has become more partisan.

We believe that Fowler is correct in using the cosponsorships network as a measure of collaboration between legislators. We, however, think that having designating a working relationship (i.e. an edge) as whether two legislators cosponsored each other's bills at least once is too noisy of an indicator. We develop a less noisy indicator as thresholding on legislators sponsoring at least a number of each other's bills, trying to target an average edge density in our networks (see our data and methods section for more details).

2.2 DW-NOMINATE

Poole and Rosenthal [7] made a major contribution to the quantitative study of political science with their procedure for computing ideological scores of members of Congress. Under the assumption that legislators and bills can be represented as points in two-dimensional ideological space, they solve for an equilibrium which determines these scores in which each legislator probabilistically votes for the bills closer to her.

The resulting DW-NOMINATE scores are useful because ideology is an incredibly important factor in how legislators behave. The polarization between parties is reflected in the widening gap between the DW-NOMINATE scores of Democrats and the scores of Republicans.

While the model is not posed as a network problem, there is a network-related interpretation: legislators who vote together frequently (e.g. have highly weighted edges between each other) end up with very similar ideology scores. Similarly, those who vote together least would be most ideologically distinct. We can therefore think of DW-NOMINATE as encapsulating the information contained in roll call voting similarity. The task of this paper, then, is to see what additional information can be learned from the information contained in bill cosponsorships.

3 Data and Methodology

3.1 Data Overview

Our primary source of data was aggregated by James Fowler [5]. Fowler developed several disjoint of House and Senate social networks for the 93^{rd} through the 110^{th} Congress. His data contains the names and Interuniversity Consortium for Political and Social Research (ICPSR) Ids for all legislators for each congress, as well as the list of all bills, amendments, and resolutions instroduced and whether they passed. Importantly, he also has an indicator matrix of which legislator was a sponsor of a bill and which legislator was a cosponsor of a bill. With this, we know who sponsored and cosponsored each bills, and thus, which legislators each individual Congressman cosponsored a bill with.

We supplement this data with additional information for each legislator. We get party, region, and most significantly ideology scores from DW Nominate [?]. We get other demographic variables like age and sex from GovTrack, another major database of Congressional information [4]. Finally, we get information on committee membership and seniority gathered from the Congressional Record and CQ Press by Garrison Nelson, Charles Stewart, and Jonathan Woon [9, 6]. To our knowledge, we are the first to combine this many sources of information about Congresspeople and their working relationships.

3.2 Constructing the Network

The questions we are primarily concerned about answering in this paper relate to how can we model working and collaborative relationships among legislators. We want to achieve this by creating a network between legislators where a connection represents a collaborative working relationship. We have a number of options to construct the network. We ultimately chose to define a working relationship between Congressman A and B if they meet the following conditions:

- For the House, there exists at least 4 cosponsorships between A and B on which either A or B was the primary sponsor. For the Senate, there needs to be at least 12.
- A needs to cosponsor at least one of B's bills and vice versa.

We discuss our choice of this network over others in the following sections.

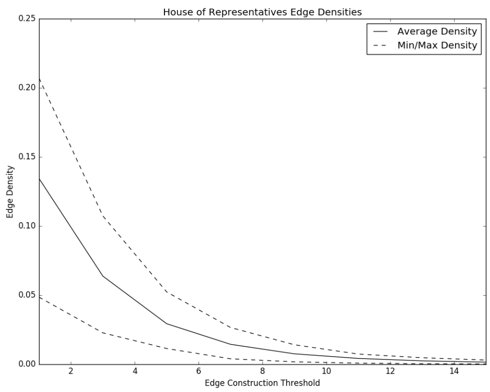
Why Not Mutual Cosponsorship

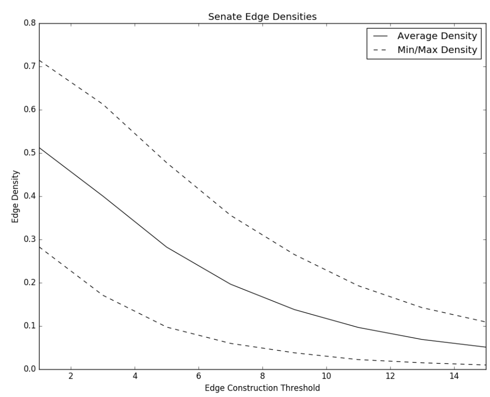
Instead of considering only cosponsorships of one-another's bills, one might consider instead building a network that accounted for mutual cosponsorships of a third legislators' bills. It's certainly true that at times, cosponsoring a bill together might mean working together on shepherding the bill through Congress. However, there are many other times when a simple mutual cosponsorship might be more of a shared ideological expression than anything else and not imply any sort of actual working relationship between the mutual cosponsors. Since we are already capturing ideological affinity with DW-NOMINATE scores, we don't want our collaborative network to cover the same ground. Cosponsoring someone else's bill and having them cosponsor yours, on the other hand, requires direct interaction and a public signal of a willingness to work together.

Determining the Cosponsorship Threshold

Not all bills, resolutions, or amendments are serious affairs that require a great deal of collaboration. Some cover trivial matters like commemorations that are cheap opportunities to signal support for a popular cause via a cosponsorship. As a result, absent a cosponsorship threshold, we will get too noisy of a signal for working relationships. Requiring consistent cosponsorship support between legislators is the best way to smooth the noise out and detect bona fide working relationships. We also add in the requirement that both legislators must cosponsor each other's bills. This is because we want to model mutual working relationships. One-way cosponsorships could include any number of situations that are not working relationships such as party pressure to cosponsor the bills of leadership or the aforementioned cheap cosponsrships of popular resolutions in which you are one of a hundred cosponsors.

To find the specific threshold, we constructed networks with varying levels of mutual cosponship thresholds and examined their edge densities. We had two goals when searching for a threshold. We wanted to hone in on a reasonable average number of working relationships proportional to the size of the network (5-10%, probably with the Senate on the higher end of that range to avoid having too few working relationships in absolute terms), and we wanted the resulting edge density to be similar for each Congress's network.





From the plots of density, we see that the average density of the House dips much more sharply (and gets smaller gaps between Congresses) than that of the Senate as a function of the threshold. Choosing a threshold of 4 for the House and 12 for the Senate works well in balancing our goals: about 5% density for the House (or 20 working relationships on average) and 10% density in the Senate (or 10 working relationships on average), both of which are reasonable, while seeing relatively low variation between Congresses. The higher density on the Senate side is further justified by the high threshold needed just to reach that limit: 12 cases of working together is certainly significant.

3.3 Data Manipulations

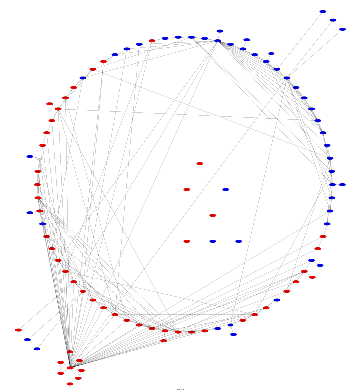
We manually massage the legislator IDs in Fowler, DW- Nominate, Stewart, and GovTrack in order to get them to agree with one another and give legislators unique IDs. Most of these changes occured when a Legislator changes party. For a detailed list of changes, please refer to our github repository.

3.4 Change In Law

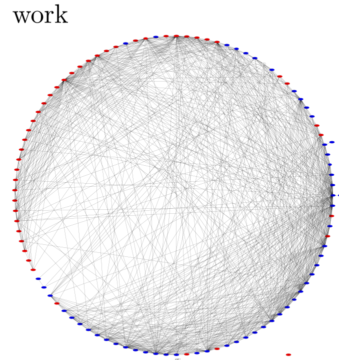
Starting with the 98^{th} Congress, a law passed which allowed a bill to have more than 25 legislators cosponsor it at a time. As a result, legislators started to cosponsor more bills than usual. We can see this behavior manifest itself in the uptick in average network degrees in congresses after the 97^{th}. As such, the behavior and the number of communities we receive from before the 98^{th} congress appear to be fundamentally different.

4 Graph Summary Stats

We start by visualizing some of the working relationship graphs. The two examples we choose are from the Senate (which has fewer nodes) and depict a high-density Congressional session (101st) and a low-density session (104th).



104 Congress - Senate Network

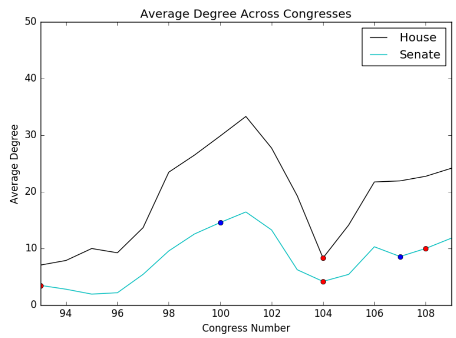


101 Congress - Senate Network

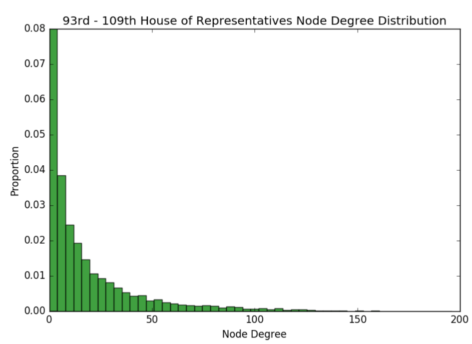
One thing that is immediately noticeable is the party clustering. There is a core of party loyalists in each graph who are densely connected to one another. Nestled among the party loyalists are a couple members from the other party. Then there are members who work closely with members from both parties who have less dense subnetworks but many edges to disparate parts of the graph.

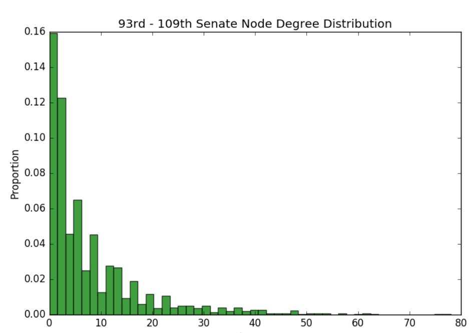
There are also noticeable authority figures in both parties who have strong working relationships with numerous others. On the other end of the spectrum are a handful of Senators with no strong working relationships at all; these members appear to be cashing their paycheck and not doing much legislating.

The two graphs side-by-side also illustrate just how much things can change between Congressional sessions. The definition of an edge is the same in both graphs but the number of edges created is starkly different between the two Congressional sessions depicted. The brutal and combaitve 104th Congress, culminating in a government shutdown, saw substantially less legislative work done and working relationships developed then the 101st just a handful of years earlier.

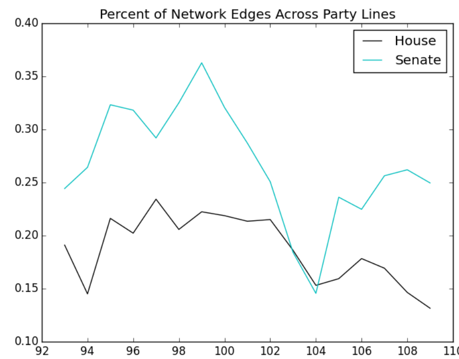


Going along with this last observation, we can visualize the average degree of the working relationship graph across time. Changes in party control are labeled as dots on the chart. After the first few Congresses, which as explained had different rules for cosponsorships, we see a burst of legislative activity and relationships between the 98th and 102nd Congress under Ronald Reagan and George H.W. Bush before a decline and recovery late in the Bill Clinton years and under George W. Bush.





Building on the point about party authorities, meanwhile, we see a degree distribution that like most networks has heavy tails in both the House and the Senate. Most members have only a handful of close working relationships but there is a long tail of legislators who work extensively with a substantial fraction of the entire body. Understanding the work of these high-degree legislators is important to understanding the work of Congress as a whole.



We can also look at the percent of edges over time that cross party lines. Interestingly, this plot is roughly stable over time, though there is a slight decline in general from beginning to end. There is a dip in bipartisan working relationships in the famous 104th congress that gets ‘locked in’ in the House while the Senate is able to bounce back. As a methodological note, we found that the same pattern held even when controlling for the change in possible bipartisan edges due to differently sized majority and minority caucuses.

5 CESNA

5.1 CESNA Clustering Description

We are interested in analyzing clusters of legislators based on their network of working relationships. Crucially, we are particularly important in how these clusters relate to the attributes of the legislators. A promising approach is laid out in Yang et al. [12], who describe an algorithm for finding Communities from Edge Structure and Node Attributes, which they call CESNA. CESNA incorporates node features and edge structure to find communities instead of just relying on one or the other. It explicitly computes the importance of different node features in forming each cluster. Additionally, it allows for overlapping and nested communities, which is one of the key features of our dataset. Legislators are likely to have distinct communities of working relationships with their regional peers, ideological peers, committee peers, and of course party peers.

In particular, we make use of the authors’ C++ implementation and feed in details of our own network. As is necessary for the models, we binarize all of our variables. This means that categorical variables like region are split into distinct dummy variables and continuous variables like ideology are split into buckets. We choose to split each ideological dimension into five groups: left, center-left, moderate, center-right, and right, each of which roughly contain a fifth of the members. The motivation behind selecting an odd number of groups is to allow for a moderate group that straddles the DW-NOMINATE center point of 0. The final list of variables, appropriately binarized, are party, ideology, region, committee, gender, and age.

5.2 Community Sensitivity and Stability

It is important that the clusters we generate be meaningful and stable given the degrees of freedom present in constructing clusters, edges, and features [11]. In particular, in our context we have direct control over the number of clusters we wish to generate using the CESNA algorithm and implicit control over the edges and features due to the decision rules we used: edges are formed only if the number of reciprocal sponsorship-cosponsorship pairs meets a specific threshold in a given term and some features are converted from continuous variables to discrete variables using cutoffs. There are also a large number of features and so some opportunity for 'overfitting' the clusters to noise.

To combat all of these problems, we employ the following strategy to validate our clusters on each Congress:

- Run CESNA to create 20 communities, the maximum of its default search space.
- Repeatedly perturb the edges and features of the network (described below) and recompute 20 CESNA communities for each perturbation.
- Use the Hungarian matching algorithm (described below) to pair up the communities in the original and each perturbed graph and compute the Jaccard similarity score associated with each pair.
- Prune the original clusters that changed too much (i.e. had average similarity scores under **FILL IN CUTOFF**).

The remaining clusters have proven stable to numerous minor alterations in the graph and so are more likely to represent robust communities of working relationships.

Edge Perturbations [HENRY - FIX THIS SECTION] *For each community in a given Congress, we will detect how robust it is to edge deletions. That is, for each edge within a community, we will randomly delete them it with varying probability and rerun CESNA to see how the resulting communities compare to the original. We will use Jaccard similarity between communities to define how similar communities are. A robust community will be one that maintains a high Jaccard similarity with reasonably low probability of random edge deletion. Communities that completely dissolve with small edge probability of edge deletions will be deemed Unstable.*

Feature Perturbations [HENRY - FIX THIS SECTION] *CESNA Only takes in discrete features. However, some of the underlying features (e.g. age and ideology) are really discretized continuous variables. As such, our perturbation strategy for variable perturbation depends on the type of variable.*

For underlying continuous variables, we will add random normal noise to the variables. The average noise will be proportional to that feature's sample average. We will vary the standard deviation for sensitivity purposes. After perturbations, we will measure the community's stability by computing the Jaccard similarity for each community. Again, communities with low Jaccard similarities will be considered unstable, and we will then reduce the number of communities that we detect.

For genuinely discrete variables (such as sex), we will employ a strategy similar to edge perturbations. Except, instead of randomly deleting edges, we will randomly switch features for all individuals within the cluster. We will use Jaccard similarity to gage how stable a community is, imputing that low Jaccard similarity means an unstable community, thus necessitating lowering the number of communities that we use.

Cluster Similarity and Pruning

Given two sets of clusters, the original and the perturbed, we wish to find a 1-to-1 matching such that the most similar clusters are paired up. The way we use this is with the Hungarian Matching algorithm, which can be used to find a maximum weight perfect matching on a complete bipartite graph. Here, the nodes in the graph are the two sets of clusters and the weight on the edge between original cluster i and perturbed cluster j is the Jaccard similarity (intersection over union) between i and j.

Upon finding the perfect matching, we assign each original cluster the weight of its edge in the perfect matching; that is, the Jaccard overlap with its similar cluster in the perturbed results. We can then average the similarity score of each original cluster across 100 perturbations to get a stability score, which we use to make our pruning decisions.

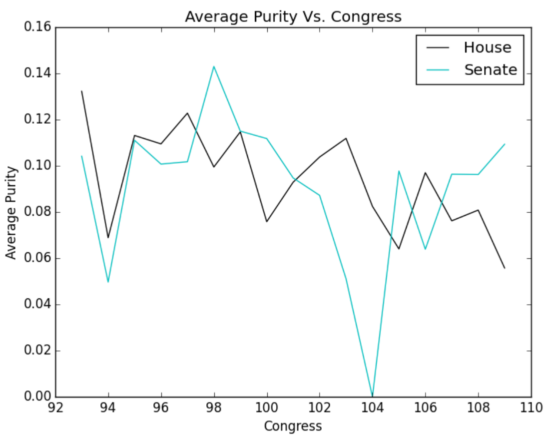
6 CESNA Analysis

6.1 Party Purity Analysis

This analysis tries to estimate how the community's purity in relation to party has changed over time. To compute purity for a binary feature, we first determine the proportion of members in a community that exhibit that feature. We will call this proportion p. The purity of a community is then defined as $p \times (1-p)$. Note that p ranges from 0 to 1. A value of $p = 1$ or $p = 0$ corresponds to a community comprised of only one type of feature, and results in a purity of 0. We note that the largest the purity criterion can be is when $p = 0.5$, then purity equals 0.25. Then, our community is very 'impure'.

We proceed with this analysis by first computing the purity of of each community in each congress. We then compute the average purity between each community within each congress. We then examine the trend of average community party purity as it differs from congress to congress.

We hypothesize that working relationships in Congress have become more strenuous over time as Congress has become more partisan. If this were true, then we would expect to see a decrease over time in community party purity as Republicans and Democrats mutually refuse to work with the other party.



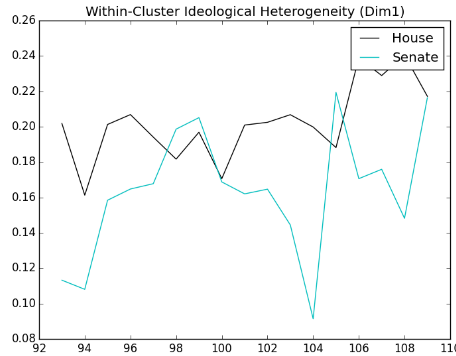
We notice first that the senate does not appear to have an overall trend up or down. It does, however, have a very sharp decline for the 104 Congress. Recall that this was the Congress that saw the very sharp decline in average degree.

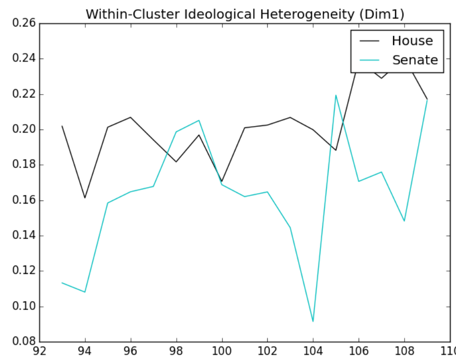
The House plot, on the other hand, appears to have a very mild decline in average purity. Given that we have not finalized our communities or performed perturbation analysis, and the fact that the decline is so mild, we would not argue that this is evidence of a more partisan House. We will re-plot this after we obtain definitively stable communities.

6.2 CESNA Ideology

An interesting question is how ideologically diverse the clusters of working relationships are and how this has changed over time. Is it true that there used to be more diverse clusters in the past but that in an era of high polarization we now find only like-minded clusters? The metric we use to answer this question is the average of the within-cluster ideological variances in each Congress.

The evidence does not convincingly support the view of increasing homogeneity. There is some slight evidence that along the second ideological dimension (social issues), working-group clusters have become more similar. On the primary dimension, however, there does not appear to be much of a trend. The 104th Congress, which





has shown up a few times throughout our analysis, exhibits abnormally low heterogeneity in the Senate but levels return to normal thereafter.

7 Bill Passage

The web of working relationships in Congress is also interesting because of the effect it may have on legislative outcomes. A basic model of legislation, along the lines of the Median Voter Theorem [2], would say that legislation is most likely to pass if its ideological content matches most closely the median ideology of legislators. A more sophisticated take would include the effects of party such as the level of support by the majority party and party leaders, who can exert control on which amendments or bills will come up for a vote.

But there are several reasons to believe that there are factors beyond just party and ideology (as defined by DW-NOMINATE) that would determine the success of a piece of legislation. For one, relatively few bills are passed relative the the amount introduced, so there is the question of which bills will be prioritized. Additionally, bills are often technical in nature without a major ideological component or are compromises between various factions. Finally, bills themselves might be subjects of larger compromises, in which legislators vote for each others' bills in what is known as 'logrolling' [10].

Each of these factors can be influenced by considering the network of working relationships. Congresspeople with more numerous, important, and strategic relationships may be more likely to get the bills they write or cosponsor passed, even holding fixed party and ideology variables.

In many ways, this is similar to the question of community growth in the network literature. The success of a bill is analagous to the spreading of the early community formed by the early cosponsors to cover a majority of Congress. Backstrom et al. [1] identify several features of early networks that they see as important for growth. In particular, they find a lot of significance in the number of nodes that have edges to someone in the early network, which they call the 'fringe'. Additionally, they look at features related to clustering in the early network, such as the ratio of closed to open triads which is negatively correlated with growth. Finally, they look at the 'activity' of the initial group, which in our context might include the legislative productivity of the early cosponsor networks.

7.1 Features

We split our feature set into *demographic* features, *visibility* features, and *network* features over the early cosponsorship network. In the case where a bill does not have any cosponsors, the features are computed for the sponsor alone. To focus on the main backers and supporters of a bill and to avoid including cosponsors who join onto a bill when it looks like it will pass, we limit the pool of considered cosponsors to only those who sign on in the first 30 days after the bill was introduced.

The demographic features most prominently include party and ideology, and include age, gender, and region as well. These features are meant to capture the palatability of the bill to various factions. If a bill's chance of success really does depend primarily on its content, we would expect these demographic variables to be the important ones. In particular, we consider

- whether sponsor is from majority party
- party impurity of early cosponsors
- ideology of sponsor
- mean ideology of early cosponsors
- variance of ideology of early cosponsors
- regional diversity of early cosponsors
- gender impurity of early cosponsors
- variance in age of early cosponsors

The visibility features are meant to capture how prominent or well-known a legislator is in Congress. In some ways, these can be thought of as a poor man's version of the network features, which explicitly capture how embedded a legislator is in the Congressional network. The idea is that more visible legislators will have greater success in attracting support for their initiatives. All variables that are computed in 'real time' are measured as of the first of the month after the bill was sponsored to reflect the state of the world after the initial cosponsorship network was formed. Our specific considered features are

- number of terms served by sponsor
- whether the sponsor is a chairman or ranking member of a standing committee
- number of bills written by sponsor so far
- number of bills cosponsored by sponsor so far
- number of consponsorships received by sponsor so far

Finally, we introduce our network features gleaned directly from the constructed web of working relationships. We construct the network in 'real time' for each bill, seeding it with the network from the prior Congress and keeping it up to date as of the bill's date so as to capture recent relationships and the situation for new Congressmembers. We view these relationships as the operationalization of the theory mentioned above that more visible legislators will succeed more often with their bills. Legislators with more useful working relationships with their peers may be more likely to get their bills prioritized and passed. It is useful to have more working relationships but it is also useful if those working relationships are with other influential members of Congress. Furthermore, it may be better if the early cosponsorship network has a diverse set of working relationships and don't cluster too tightly among themselves. These intuitions lead us to the following set of features

- degree of sponsor
- eigenvector centrality of sponsor
- cut size of set of early cosponsors
- clustering of early cosponsors
- presence of early cosponsors in stable CESNA clusters

References

[1] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *IN KDD 06: PROCEEDINGS OF THE 12TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING*, pages 44–54. ACM Press, 2006.

[2] Duncan Black. On the rationale of group decision-making. *Journal of Political Economy*, 56:23–34, 1948.

[3] Royce Carroll, Jeff Lewis, James Lo, Nolan McCarty, Keith Poole, and Howard Rosenthal. Dw-nominate scores with bootstrapped standard errors, 2015.

[4] LLC Civic Impulse. Govtrack.us, 2016.

[5] James H. Fowler. Connecting the congress: A study of cosponsorship networks. *Political Analysis*, 2006.

[6] Charles Stewart III and Jonathan Woon. Congressional committee assignments, 103rd to 114th congress, 2017.

[7] Howard Rosenthal Keith T. Poole. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384, 1985.

[8] Geoffrey C. Layman, Thomas M. Carsey, and Juliana Menasce Horowitz. Party polarization in american politics: Characteristics, causes, and consequences. *Annual Review of Political Science*, 9:83–110, 2006.

[9] Garrison Nelson. Committees in the u.s. congress, 1947-1992, 1994.

[10] Thomas Schwartz. Collection of issues and vote trading. *The American Political Science Review*, 71(3):999–1010, 1977.

[11] Vassilios Tzerpos and Richard C. Holt. On the stability of software clustering algorithms. In *Proceedings of the 8th International Workshop on Program Comprehension*, IWPC '00, pages 211–. IEEE Computer Society, 2000.

[12] Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *2013 IEEE 13th International Conference on Data Mining*, pages 1151–1156. IEEE, 2013.