Analyzing The Network of Congress
CS224W Project Milestone

Henry Neeb, Taman Narayan, Christopher Kurrus

November 17, 2016

1 Abstract

The United States Congress is a rich social network where each legislator interacts with their peers through committees, house leadership positions, and bill cosponsorships and amendments. Through the lens of growing partisanship, we investigate qualitative measures of the network, including the difference in congressional networks using different relationship features, how legislators are clustered beyond party lines and what features are common in those clusters. We investiage how these clusters and communities evolve over time through different congresses and evaluate how partisianship has affected these communities and working relationships in Congress.

We achieved this by first building a cosponsorship network of the 93^{rd} through 110^{th} Congress and supplemented the network with information about each Congressman (node attributes). We then implemented Community Detection in Networks with Node Attributes (CESNA) on the network and tracked communities and community attributes over time, noting how important partisianship was in defining each community.

Further, we investigate to what extent early features within the same networks can predict the success of a bill being passed into law.

2 Related Work

2.1 Fowler

Fowler [**?**] 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 [**?**] 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 [**?**]. 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 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, including ideology scores provided by DW Nominate [**?**], party, age, sex, and region of the United States. Data sources include DW-NOMINATE and GovTrack [**?**].

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 with 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.

Cosponsorship Versus Role Call Votes

We can infer that a cosponsorship is almost equivalent to voting for a bill. If you are willing to cosponsor a bill then you will most likely vote for the bill. We decided against using roll call votes as our measure of collaboration for the following reasons:

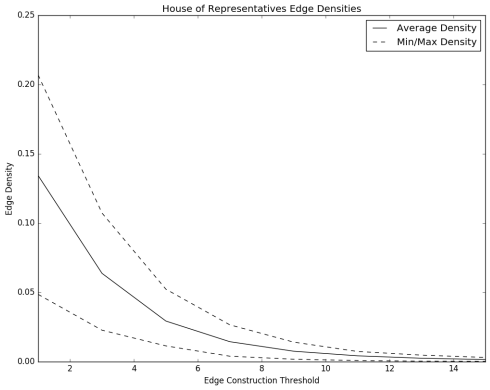
- Cosponsorship contains support signaling. It is a method legislators use to signal to the rest of Congress that they support a bill.
- Voting for a bill does not necessarily imply a collaborative working relationship. There are reasons to vote for a bill that goes beyond a working relationship, including voting for a partisan issue or voting for an issue that affects your constituents.

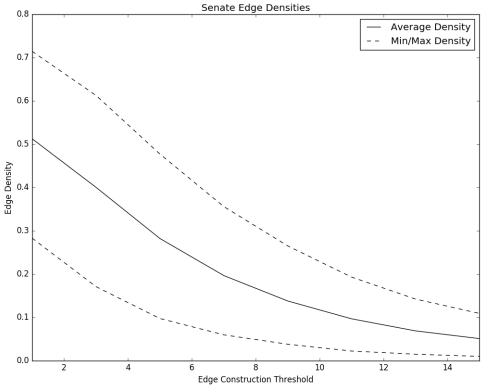
In short, we believe that embedded in roll call votes there is a signal of a working relationship, but we believe that the signal is obscured by noise relating to other voting reasons.

Why 5 Mutual Cosponsorships

We believe that there is a stronger signal in mutual cosponsorships. We believe, however, that without some thresholding on the number of cosponsorships, we may get too noisy of a signal for working relationships. We think that consistent cosponsorships support between legislators is the best way to smooth out and detect working relationships. We also added in the requirement that both legislators must cosponsor each other's bills. This is because we want to model mutual working relationships. Allowing one-way cosponsorships allows for relationships where party and committee leaders force the rank and file legislators to cosponsor their bills.

We constructed networks with varying levels of mutual cosponship thresholds and examined their edge densities. We wanted 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, and we wanted the density to be on the same magnitude across all congresses. For the House, we thought that on average, having about 20 working relationships (relative to the usual 435 members of the house) would be ideal. This would corresponds to an average density of approximately 0.05. For the senate, this would be about 0.1 average density.





From the plots of density, we see that we achieve the desired average density with minimal spread between congresses for the house at 4 and for the senate at 12.

Final Networks

We implement this methodology across all congresses. Since the Senate and the House cannot cosponsor each others bills, we construct disjoint graphs for the Senate and House. Also, since we are looking at the evolution of communities over time, we construct these networks separately for each congress.

3.3 Data Manipulations

We manually changed some of the Fowler unique identifiers to match those in DW Nominate and GovTrack. 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.

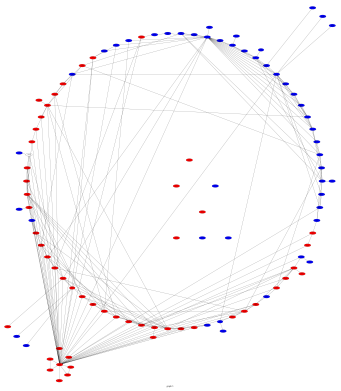
4 Graph Summary Stats

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).

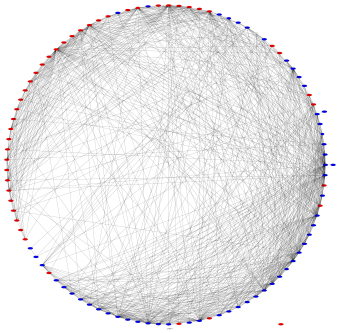
One thing that is immediately noticeable is the party clustering. There is a core of party loyalists in each graph who are densely connected to one another. Nestled among the party loyalists are a couple members from the other party. Then there are members who work closely with members from both parties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

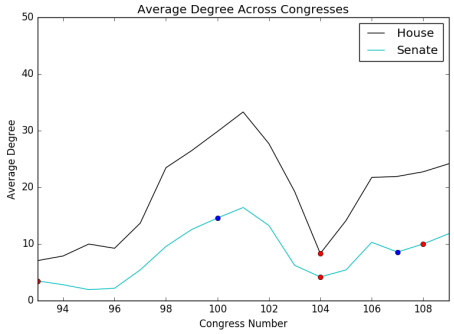
The two graphs side-by-side also illustrate just how much things can change between Congressional sessions. The definition of an edge is the same in both graphs but the number of edges created is starkly different between the two Congressional sessions depicted. The brutal and combaitve 104th Congress,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



104 Congress - Senate Network

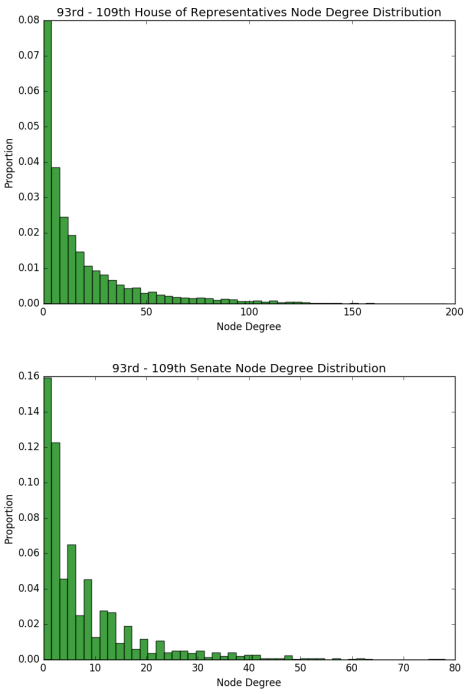


101 Congress - Senate Network



Going along with this last observation, we can visualize the average degree of the working relationship graph across time. 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 mirrors the well-known power law 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.

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. [**?**], 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

Currently our analysis is done without community sensitivity. We allow CESNA to pick the number of communities and we performed our preliminary analysis on these communities. In our final report we plan to do more extensive testing on the stability of the communities that we develop.

The CESNA algorithm and community assignment is dependent on several things: the number of total communities that we wish to create, the edges distribution in the graph, and the attribute values within the graph. When we analyze community stability, we will perturb each individual and jointly.

Edge Perturbations

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

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.

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.

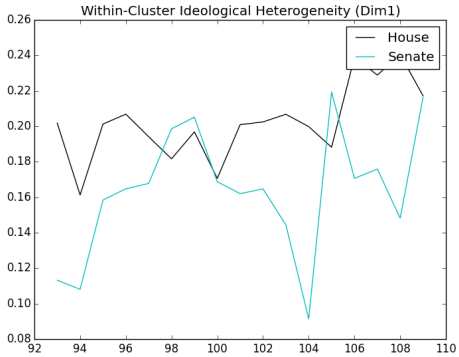
We notice first that the senate does not appear to have an overall trend up or down. It does, however, have a very sharp decline for the 104 Congress. Recall that this was the Congress that saw the very sharp decline in average degree.

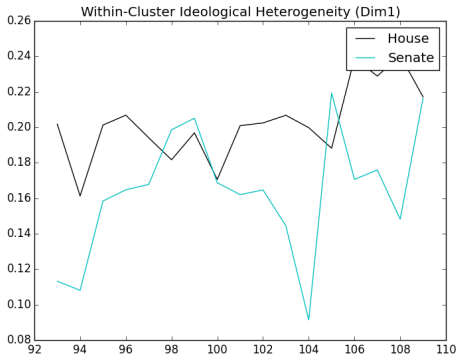
The House plot, on the other hand, appears to have a very mild decline in average purity. Given that we have not finalized our communities or performed perturbation analysis, and the fact that the decline is so mild, we would not argue that this is evidence of a more partisan House. We will re-plot this 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 (Black, Duncan (1948). "On the Rationale of Group Decision-making". Journal of Political Economy. 56: 2334. doi:10.1086/256633), 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' (Schwartz,

Thomas (1977). "Collection of Issues and Vote Trading". The American Political Science Review. 71 (3): 9991010. doi:10.2307/1960103).

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 traditional problems of community growth in the network literature. We are interested in whether the early cosponsor network of a bill will grow to such a point that the bill is able to pass or whether it will remain small and die off at the end of the Congressional cycle. One way to tackle this problem would be trying to predict who will join in as cosponsors, which is an edge prediction problem. Another, the way we pursue here, is to see which bills end up passing, a traditional binary machine learning problem. Both are interesting problems from an intellectual standpoint and we focus on the latter using the reasoning that bill outcomes are ultimately more interesting from a political standpoint than 'intermediate' successes like marginal cosponsors.

Backstrom et al. (BACKSTROM CITE) 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.

Our final goal, then, is a predictive model for bill passage based on the sponsor and early cosponsors of a bill. Predictive accuracy is of course important but it's really the variable importance that is of broader interest - for example, do network features matter and how has the importance of party and ideology changed over time relative to network features. Possible features include

- party of sponsor
- number of cosponsors
- partisan makeup of cosponsors (pct Democrat)
- ideological makeup of network (mean, variance)
- number of bills written by sponsor
- degree of sponsor in network
- number of edges leaving the cut of cosponsors
- size of connected component of sponsor (graph is NOT connected)