Analyzing The Network of Congress
CS224W Project Proposal

Henry Neeb, Taman Narayan, Christopher Kurrus

October 20, 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 co-sponsorships and amendments. Through the lens of growing partisanship, we will 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, and how networks change overtime (say for when an incumbent is replaced with a new legislators). Further, we will investigate to what extent relationships between legislators govern support for the bills the propose and the likelihood of the bill being passed into law.

We plan to use data on the cosponsorship networks of the 93^{rd} through 110^{th} Congress and supplement it with more descriptive information on the bills that were proposed and the legislators that proposed them. We believe that enriching these datasets will provide us with additional opportunities for performing more sophisticated network analysis, such as Latent Multi-Group Membership Graph Models, which will allow us to construct latent clusters on legislators based on their own features, such as sex, party, and ideology.

2 LitReview

2.1 Connecting the Congress: A Study of Cosponsorship Networks (Fowler, 2006)

Fowler [2] 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. With these partitioned graphs, he examines several centrality measures for each graph (i.e. degree, betweenness, closeness, and eigenvector) and sees how these measures track through time for each individual congressman. He hypothesized that the most connected congressmen by centrality members should stay stable over time and evaluated this by tracking the most central members and seeing if they remained the most central across different Congresses.

He continues his study by employing his own weighting metric on a directed graph. He hypothesizes congressman A who cosponsors a bill with congressman B shows different levels of support for each other depending

on the total number of individuals who cosponsor a bill - that is, if A is the only cosponsor for B, this signifies stronger support than if A was one cosponsor out of 25. Fowler constructs a weighted graph such that the weight from A to B is for each bill where A cosponsored a bill with B, the sum of the reciprocal of the total number of cosponsors for that particular bill. He then infers closeness centrality with these weighted edges (a metric he calls connectedness) and clusters between congressman and evaluates that the most connected clusters appear to be amongst the leadership positions (committee chairs and ranking members). Further Fowler constructs a negative binomial model using the centrality and connectedness measures as features to predict the number of bills and amendments that are successfully passed by each representative. He compares the effect sizes for each model as the ratio of the predicted number of amendments passed for a person who had 1 standard deviation greater connectedness or centrality to the number of amendments passed for a person of average connectedness or centrality for each congress and notes that connectedness is the metric that is better at measuring a difference in effect as compared to all centrality measures. The effects sizes range between 39% and 59%.

Fowler concludes his study by trying to predict roll-call votes for a bill given the connectedness of the sponsor and controlling for ideology measures (Poole and Rosenthal 1997). He is able to infer a positive correlation indicating that there are more 'aye' votes for a bill when the individual is connected, but the effect size is small.

2.2 Supervised Random Walks: Predicting and Recommending Links in Social Networks (Backstrom and Leskovec, 2011)

Backstrom and Leskovec [1] propose an algorithm, Supervised Random Walks, designed to naturally combine node and edge information with network structure in order to provide superior link predictions and recommendations.

The link prediction and recommendation problems are considered challenging, as real world graphs are typically very sparse, and it can be difficult to determine to what extent the network structure is responsible for linking, as opposed to node and edge features, or even external factors. To effectively mingle this structural information and feature data, the Supervised Random Walks algorithm performs supervised learning steps using the edge features to bias the random walk over the network so that it will visit certain nodes more often than others. During the learning steps, the edge features are used to learn the 'strengths' of each edge and assign a positive or negative value, such that more positive nodes are visited more frequently. In addition, for link prediction, the positive links are the ones where edges will be created in the future, whereas negative links are left alone. To assign strengths using edge features, they trained an algorithm using a source node and training examples where edges will be created.

When tested on physics coauthorship and Facebook friendship data, Supervised Random Walks outperformed both unsupervised and supervised methods. In addition, Supervised Random Walks doesn't require feature generation or extraction, so it is doubly impressive that it beats algorithms that use supervised machine learning methods with complex feature extraction and generation schemes to account for node and edge features. Overall the algorithm appears to offer significant improvements over well known alternatives like random walks with restarts and supervised machine learning techniques.

2.3 Latent Multi-group Membership Graph Model (Kim and Leskovec, 2012)

Kim and Leskovec [3] propose a model that explicitly links node features with network structure. In particular, they develop a two-stage model to understand this relationship. First, the node features determine which of several latent groups the node is a part of. Nodes can be a part of multiple groups and the groups can have overlapping memberships. Second, the extent and nature of overlapping group membership between

the nodes governs the probability of edge formation between nodes.

The model can be applied to various tasks. The authors document how it achieves success both in predicting node features and in predicting edge formation. Additionally, the intermediate step of group classification allows various types of clustering analyses to be run - for example, which features correspond to membership in different groups and how dense the link structures in the resulting groups are.

3 Critique

3.1 Fowler

Fowler's main analysis was to compare his measure of connectedness to traditional centrality measures to see if it provides a better predictive measure of whether a representative can get an amendment on a bill passed and if the representative can rally more votes for a bill they cosponsored. He did so by constructing a network for each Congress and each house separately with no links going between them and evaluating his models with Congress and house. We have a number of issues with this approach:

- Segregated networks introduce the memoryless property to the overall social network. It could be that it is assumed that two members in a prior Congress will remain connected in a future Congress, but there was no analysis to evaluate this.
- His connectedness metric is based on heuristic parameters with little support beyond intuition. Further, his bucketing strategy for only considering bills within a Congress to determine an edge weight is a bias-variance tradeoff consideration that was not evaluated.
- His baseline model of connectedness is very noisy, which is something he himself acknowledges (pg 459). His cutoff for whether to connect representatives is if they cosponsored a bill for one another. However, just one cosponsorship does not imply support for an individual. We would assume that consistent cosponsor support would imply a relationship.
- There was no analysis done on distinguishing between a cosponsor and a primary sponsor (someone who owns the bill).

Further, Fowler's analysis lacked a descriptive statistics section for his measurement of effect size on amendment passage. He mentions that all coefficients are significantly different from 1 (pg 476), but he never mentions at what significance and how he confirmed the significance. It does not appear that he validated this model on a future Congress.

Overall, Fowler's analysis is too discretized and he constructs models without concern for their statistical underpinnings and methods for evaluation. We would have liked to see a more temporal evolution of congressional relationships. The network grows as a Congress continues and some of the relationships built in previous Congresses clearly carry over to the next. Instead of looking at a Congress as a distinct 2 year interval, we want to see how the network blossoms over time.

3.2 Backstrom and Leskovec - Critique

Link prediction has several possible applications for a network like the one we are considering, but on data that is so feature rich it is important to be conscious of how the algorithm we use is handling the balance

between network structure and node/edge features. The Supervised Random Walks algorithm would be ideal for this application, as we will be able to account for our node and edge features, without requiring costly and time consuming feature engineering. In addition to the savings the algorithm provides over a typical supervised machine learning algorithm, because of the method with which the strength function we train in the supervised step is implemented it is compatible with any of our combination of categorical and continuous variables, allowing us more leeway in the subset of features that we select to guide our link prediction.

One thing to keep in mind is that the Supervised Random Walk model is very sensitive to what an edge means. Given that there are various ways for us to represent edges (e.g. through cosponsorships, committees, etc.), we have to be careful in evaluating the edges that the model predicts. That is, if we do not construct meaningful edges with appropriate edge features, we will get suboptimal results.

3.3 Kim and Leskovec

The Kim and Leskovec paper includes a number of elements that are very relevant to our work. In particular, the focus on node features mirrors with our interest in how much partisanship and ideology affect the workings of Congress. Additionally, the nature of the group memberships is very appealing. It is very natural to assume that Congresspeople belong to multiple overlapping groups instead of being a part of a single cluster. They have their political party, ideological neighbors, committees and subcommittees, geographical neighbors, demographically similar members, and so much more. It is of great interest to understand how important each of these characteristics are and how they have evolved over time.

At the same time, the paper does not perfectly map onto our work. The model is expansive and general with many free parameters that might not be necessary in our context. For example, it is possible to support situations where feature similarity decreases edge likelihood, which is unlikely in the case of cosponsorship networks. Similarly, the model supports core-periphery relationships that might be less applicable in a body with only 435 members; we'd expect less distinction between members. Finally, many formulations of our graph are much denser than the networks studied in the paper. Still, there is a great deal of overlap and it would be interesting to see the results of the latent-group approach.

4 Proposal

4.1 Data

We will be using the data compiled by James Fowler on the 93^{rd} through 110^{th} congress. This data includes the ICPSR id number for each representative, the bills that they have sponsored or cosponsored, and the date of these events. Further, we have data on when a bill was created and when and if it was passed into law.

We have additional supplementary data on candidates and bills. Much of it is compiled by govTrack, which is an entity that scrapes and compiles congressional data from congress.gov. govTrack has more granular detail about each bill beyond its existence and ultimate outcome, including the entirety of the text, what happened to a bill during each stage of the bill's life (such as passage through committee), and who voted for and against a bill. It also has more information about each Congressperson, including their partisanship, state or district, and committee memberships. We also draw upon ideology information (DW- Nominate scores) about each legislator.

In total, this comprises approximately 1,000 representatives (nodes) and about 350,000 bills (potential edges).

This amounts to about 500 Mbs of graph data. In additon, the storage size of the metadata of all bills is approximately 1 Gb. Depending on the method we use to generate a network, there is the possibility of it being quite dense; for example, if we create an edge when two Congresspeople cosponsor at least 10 bills together, almost a fifth of all possible edges are created.

4.2 Problem

The growth of partisanship in recent years has dominated the political conversation in Washington D.C. The Democratic and Republican Parties have grown more ideologically distinct and unwilling to compromise.

We are interested in shedding light on how this evolution has occurred by focusing on the minutiae of governance - the network of sponsorships and cosponsorships of the thousands of bills introduced each two-year cycle, most of which never see the light of day. In particular, we seek to understand the state of working relationships in Congress, how they have evolved over time, and how these evolutions have affected legislative outcomes.

We see a number of direct applications of graph and network analysis that can help us answer these questions. On the qualitative side, we hope to analyze different clusters and coalitions of Congressmembers over time using methods such as the Latent Multi-Group Membership Graph Model discussed above. Which features characterize these coalitions over time; do bipartisan coalitions still exist in any meaningful way? How important are the coalitions to explaining bill outcomes? Who are the power brokers in Congress that have high centrality scores in these graphs; are they senior coalition-builders or party leaders?

There are a number of quantitative models we think would be appropriate as well. To better understand the dynamics of working relationships, we can consider an edge prediction problem: what features best predict whether two Congresspeople are likely to work together? We think the Supervised Random Walks approach is promising. We also think that there might be machine learning approaches to understand why bills pass as a function of their cosponsorship networks - is partisan unity starting to trump early bipartisan support?

Fowler has started down a productive path in thinking about many of these questions, but there is a lot more we can do. On the micro-level, what is the best way to represent the network of Congressional relationships in terms of undirected vs. directed, weighted vs. unweighted, and so on? More broadly, how does partisanship tie into many of the questions he addresses?

4.3 Model Evaluation

Cohorts

Our base model and assumption is that partisanship mostly governs cohorts. When evaluating cohorts that we develop through legislator attributes other than party, we will compute and evaluate party purity as compared to cohorts that are based on party. That is, we will evaluate if a cohort was formed due to party alliances by examining the like features of the cohort - if the predominately like feature is party, we will assume the cohort formed by party.

Bill Passage and Votes

We will assume that the party of the main bill sponsor and cosponsors, as well as the number of cosponsors is what governs the likelihood that a bill will be passed. We will develop a variety of predictive models (logistic regression, supervised random walks) that train on features other than party on early bill networks. We will examine which features govern the likelihood of legislators working together and bills getting passed.

References

[1] L. Backstrom and J. Leskovec. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of Web Search and Data Mining (WSDM)*, 2011.

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

[3] M. Kim and J. Leskovec. Latent multi-group membership graph model. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.