Co-Authorship Networks In Congress

CS 224W Final Project Report | December 11th, 2016

Patricia Perozo, Gaspar Garcia, Cheenar Banerjee

Introduction

The legislative branch of the United States government plays a central role in the health and operations of the country. Of all the powers and duties vested in Congress, the most important is the power to create and pass national legislation. However, though the legislation proposed in Congress affects every single United States citizen, it's often slow to pass, if at all. Oftentimes, it's difficult to tell whether or not crucial legislation will pass one or both chambers, which can cause major obstacles and uncertainty for policy and legislators. If there were a way to successfully predict whether or not a bill will pass, it would be groundbreaking for many people involved in government.

In the United States Congress, social networks play a major role in the production and passing of legislation. Congressmen and women can co-sponsor legislation, or publicly declare their support for a piece of legislation proposed by a colleague. Congressmen and women spend a significant amount of time and energy building these co-sponsorship relations. Furthermore, co-sponsorships on legislation are frequently brought up during sessions of Congress as well as to constituents and other parties. In summary, co-sponsorships play a major role in the legislative process. These co-sponsorships create a social network that may be telling of legislation success or failure -- each sponsor on a bill can be defined in terms of his/her place in the network. How connected is he/she? Who is he/she most closely associated with? What network characteristics best summarize him/her? What network characteristics determine the outcome of the legislation that this person chooses to sponsor?

In our project, we investigate the importance of these social networks on legislation outcome by (1) identifying network characteristics of successful and unsuccessful bills, (2) using the features we identify to try and predict legislation success based on co-sponsorship social networks, and (3) predicting future co-sponsorships.

Literature Review

Legislative co-sponsorship Networks in the U.S. House and Senate by James Fowler

To the best of our knowledge, no work has been done to use the features of co-sponsorship networks in Congress to predict whether or not legislation is successfully passed. We will be building off of previous work related to the congressional network itself by James Fowler and separate work surrounding using subgraph features in prediction. Fowler does an extremely thorough analysis summary statistics and comparisons between the different years of both the house and senate. He then uses that information to come up with a metric of "Connectedness" of the legislators. Where Fowler focuses on the connectedness and then the success of the most connected legislators by becoming party leaders in following years, we will be focusing on predicting the chances of a bill passing the house or even becoming law. This type of prediction speaks to very different measures of success. Whereas Fowler defines success to be entering party leadership, we will be focusing on passing legislation. Furthermore, Fowler constructs his graph based on "distance" between legislators, rather than using the directed edge co-sponsorships based directly on bills.

The Link Prediction Problem for Social Networks by David Liben-Nowell and Jon Kleinberg

This paper describes methods for the structural analysis of graphs based solely on topology. It sets aside the use of node or edge features[3]. Specifically it focuses on different algorithms for link prediction and their greater use in measuring larger scale graph characteristics. The overarching question is, to "what extent can the evolution of a social network be modeled using features intrinsic to the network itself?" The paper offers a robust explanation and description of the algorithms used, including preferential attachment. However, the paper fails to take a strong stance on which link prediction models are best suited for different graph structures. Instead it just offers general statistics. Nonetheless, it still provided us a strong perspective and argument for why there is much to learn about a graph from link prediction methods alone. This encouraged us to continue exploring link prediction for our use case.

Overview

We model the co-sponsorship network of the House as a directed graph, where each node represents a single legislator, and a directed edge from legislator A to legislator B indicates that legislator A co-sponsored some piece of legislation proposed by legislator B. Our approach to this project consisted of three main components: (1) Feature extraction from successful bill "subgraphs," (2) Bill success prediction based on said feature extraction, and (3) link prediction within the co-sponsorship network graph.

In this project, we chose to focus on the House networks exclusively. We do not consider the bill's status in the Senate or at the Presidential level. Before trying to capture the extra complexity of the interplay between the House and Senate networks, we wanted to focus on a more basic network to see what salient features stood out to us. We chose the House network because it is larger, and based on Fowler's findings, the network is less dense, so it might allow us to see a wider variety of network interactions [1]. We define *success* of a bill as a binary variable, where a bill is successful if it passed in the House, and a bill is unsuccessful if it did not pass in the House.

Prior to the 96th Congress (1979), House rules prohibited more than 25 total sponsors on any bill. During the 96th Congress, this co-sponsorship cap was lifted and to this day, there does not exist a co-sponsorship cap on bills. We hypothesize that the presence or absence of this co-sponsorship cap may significantly influence the co-sponsorship network structure, because if a cap is present, legislators may put more weight into their decision to co-sponsor a bill. Thus, in our analysis, we focus on two separate sets of Congressional sessions: the 93rd-95th Congresses (co-sponsorship cap), and the 97th-109th Congresses (no co-sponsorship cap). We leave out the 96th Congress because the co-sponsorship cap rules changed during this session of Congress, and we cannot place it neatly into either set. Each session of Congress contains around 10-20,000 bills in total.

We use a combination of a dataset created by James Fowler of the co-sponsorship networks from the 93rd Congress (1973) through the 109th Congress (2007), and a GitHub dataset containing metadata for all United States legislators from 1789 to the present. We used the Fowler data to create the directed graph. We used the GitHub data to store information about each legislator in the 93rd-109th Congresses. The datasets had many unexpected inconsistencies that we had to work around. First, we expected that every bill in the dataset would have a sponsor, but there were bills without marked sponsors. We chose to ignore these bills since they didn't contribute information about co-sponsorship. Second, we found that approximately 15 representatives per session had no political party affiliation recorded in the GitHub data. We chose to treat party missing as its own political party class, because we feel that the case where a legislator chose not to disclose his or her political party is equally as informative as the case where a legislator did choose to disclose his or her political party. Finally, there are 435 total seats in the United States House of Representatives. However, every graph has more than 435 nodes because during each session of Congress, some representatives left and their spots were filled by new representatives. Each representative played a unique role in the co-sponsorship network, so we chose to include every one of them.

Feature Extraction

Methodology

We attempt feature extraction by comparing the *"subgraphs"* of successful bills with the *"subgraphs"* of unsuccessful bills. These aren't subgraphs in the strict sense of the term, because we don't isolate nodes into a separate subgraph -- rather, for each bill, we look at all nodes that sponsored or co-sponsored that bill, and compute statistics for those nodes within the context of the entire graph. We chose to do this because it gives us a better sense of how nodes that participate in a certain bill interact with the entire House network as a whole. We hypothesized that there may be certain network features within bill subgraphs that indicate bill success or failure.

We chose to look at the 93rd and 103rd Houses in our initial feature exploration. We randomly chose each of these sessions from the sessions with and without a co-sponsorship cap, and we expect that the findings will generalize to other sessions of Congress. After constructing the directed co-sponsorship graphs for each session, we computed a variety of network structure statistics using SNAP for subgraphs of each unsuccessful and successful bill in each of the 93rd and 103rd sessions of the House. For each statistic, we computed the mean value for all nodes in the given subgraph and then compared the mean of means using the Mann-Whitney significance test. We also computed statistics based on intrinsic characteristics of each node's party, gender, and states with the highest representation in the House. We chose these particular features based on our domain knowledge of how relationships in Congress develop.

Results and Findings

The results of the features we considered are summarized in Tables 1 of the Appendix. We used the Mann-Whitney test to compare the means for each statistic between successful and unsuccessful network subgraphs. We defined potential predictive features as features that had a statistically significant difference in means across all successful and unsuccessful bills in a particular session of Congress, where statistical significance was determined by a significance level of 0.05. The 93rd Congress network has 446 total nodes and 31,660 edges and the 103rd Congress has 447 nodes and 65,072 edges. There is total of 28,6486 bill subgraphs between these two sessions. From this alone, we can see that the absence of the co-sponsorship cap significantly increased the density of the co-sponsorship network. We were surprised to find that in the 93rd Congress, almost every metric we examined appeared to be statistically significant in terms of differentiating successful bill subgraphs from unsuccessful bill subgraphs, while only a handful of features were significant in the 103rd Congress. This is a strong indication that the co-sponsorship cap had a significant effect on the significant of co-sponsorship networks in Congress, as we hypothesized earlier.

Interestingly, most the graph structure metrics we used didn't show a significant difference in both Congresses, with the exception of mean number of sponsors and mean betweenness centrality. Mean number of sponsors indicates that having a larger number of sponsors on a bill means that the bill is more likely to pass. This corroborates the amount of time and effort Congressmen and women spend on their co-sponsorship relationships. Qualitatively, betweenness centrality of a node indicates the number of shortest paths between all vertices in the graph that pass through that node. The fact that betweenness centrality is significant here indicates that in order for a bill to pass, it might be helpful if on average, sponsors of the bill are on shortest paths within the network. In the context of Congress, this indicates that it may be important for a bill to get support from legislators that form bridges between clusters, where clusters may be by political party, by committee, by state, or any other number of factors. We also found that several node-based statistics seem to be significantly different, indicating that characteristics of legislators themselves may be indicative of co-sponsorship relationships that will form, and ultimately bill success.

Bill Success Prediction

In this phase of the project, we wanted to investigate the predictive power of these features to see whether or not we could actually use them to predict the success or failure of a bill in a particular session of the House.

Methodology

Define *significant features* as those features from the previous part which were significantly different between successful and unsuccessful bill subgraphs within each set. In order to focus on the predictive power of each potentially significant feature, for each set of sessions, we first computed the mutual information of all significant features. We then used a Naive Bayes classifier to predict bill success or failure within each set of sessions. We chose to use this relatively simple classifier in order to focus on the predictive power of features rather than the overall performance of the classifier. For each session, we computed a feature vector and corresponding label for each bill in the session based on the significantly different features from the previous part. We do a binary classification where a label/prediction of 0 indicates bill failure and a label/prediction of 1 indicates bill success. We used leave-one-out-cross-validation on each set of sessions and ran the predictor three times for each set, using a different feature vector for each run (once using all significant features, and once each using top three and top five features as determined by mutual information.

Results and Findings

The most predictive features in the capped sessions were the graph-based features, while the most predictive features in the uncapped sessions were mostly node-based features. This indicates that as the network becomes more dense, the overall structure of the graph is less important than the features of the individual nodes in overall bill success. Using the features with highest mutual information increased the overall accuracy, but actually decreased the proportion of correct successful bill predictions. On the whole, about 95% of bills are unsuccessful and about 5% of bills are successful in any given session, meaning the expected true positive rate is 0.05 if the classifier were guessing at random. Our predictor shows a consistent true positive rate of around 0.08 when we predict using all significant features, which is higher than our chance estimate. Because we've run several cross-validations on the true positive rate and it's still higher than 0.05, we believe that this increased true positive rate is not due to chance. While our classifier is far from perfect, these results lead us to believe that it is possible to use network features of the Congressional co-sponsorship graph to get some idea of bill success.

Link Prediction

In this section we explored two different approaches to the classic link prediction problem. The first is the Supervised Random Walks algorithm proposed by Backstrom and Leskovec, in which random walks of PageRank are combined with learned node and edge features[4]. While Backstrom and Leskovec developed the algorithm to address the issues of link prediction in sparse networks, we still felt that there could be something gained in our dense network by using their combined supervised random walks approach. The second is a modified Network Evolution algorithm proposed by Leskovec, Backstrom, Kumar, and Tomkins[5]. Again, while this algorithm was developed for sparse networks we felt that it could be applied to our dense network since the backbone of the algorithm is based on completing triangles which work even in dense graphs.

General Link Prediction Methodology

The formal link prediction task is: given a training interval [t_0, t_j] and a test interval [t_k, t_l] where $t_j < t_k$ we want to create a ranked list of edges that are predicted to appear in the interval [t_k, t_l] but don't appear in [t_0, t_j]. We use a variety of algorithms to accomplish the link prediction task.

Methodology - Random Walks

Let *LP* be the link prediction algorithm which outputs the ranked list referenced above. In our use case we want to predict co-sponsorship networks for successful bills. More formally, we want to predict edges that appear in a co-sponsorship network of successful bills and but do not appear in either . More details follow below.

Our formal methodology specific to our co-sponsorship data is as follows:

1. Model the data as a directed graph with bill sponsors as nodes. Primary bill sponsors have inbound directed edges from co-sponsors.
2. Create three co-sponsorship networks. (1) The first co-sponsorship network is the entire network of all bills proposed, successful or not. (2) The second network is made using only successful bills in the training interval [t_0, t_j]. (3) The final network is made using only successful bills in the test interval [t_k, t_l]. We calculate intervals [t_i, t_j] using timestamps when the bill was co-sponsored.
3. Use LP on the second network from step 2.
4. Evaluate prediction performance using networks 1 and 3 from step 2. We test whether the predicted edges appear in graph 3 and not in graph 1.

Algorithms

Personalized Pagerank and **Supervised Random Walks**
We run a personalized pagerank on nodes labeled sponsors to calculate the node “closest” to them. Here we aim to extend the notion of consistent triads, “a friend of my friend is my friend”, to the main bill sponsors.

The Supervised Random Walks algorithm employs an approach similar to Personalized Pagerank with the added component of edge strengths. The formal optimization problem is as follows:

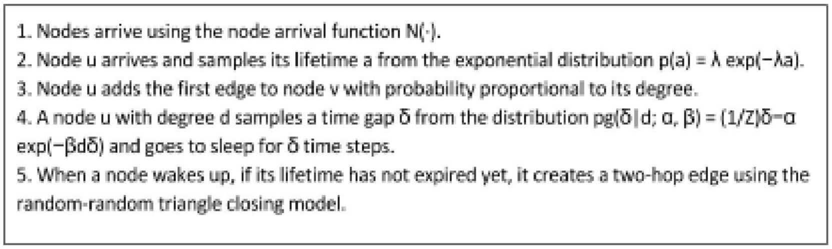
$$\min_{w} F(w) = ||w||^2 + \lambda \sum_{d \in D, l \in L} h(p_l - p_d)$$

The optimization problem is modeled as a random walk where we aim to visit nodes in D but not nodes in L. p_d and p_l above are the pagerank scores for nodes in D and L respectively. We are minimizing the cost function h, which is taken over the difference over p_l and p_d. To solve this optimization problem we use simple gradient descent.

Feature Selection: In selecting features we are looking for edge features that are relevant to both edge creation and the success of a bill being passed. We were limited in the features available largely due to incomplete datasets. However, we explored details of nodes including party affiliation, gender, and state. Additionally, since each edge in the graph is created for a co-sponsorship on a bill, we tried to add features relating to bills such as the bill topics. To learn the weights for these features we used two variants of h: the Wilcoxon-Mann-Whitney loss function and the Squared Loss function.

Triads

In the algorithm developed by Leskovec et al., we see:



1. Nodes arrive using the node arrival function N(·).
2. Node u arrives and samples its lifetime a from the exponential distribution p(a) = λ exp(−λa).
3. Node u adds the first edge to node v with probability proportional to its degree.
4. A node u with degree d samples a time gap δ from the distribution pg(δ|d; α, β) = (1/Z)δ−α exp(−βdδ) and goes to sleep for δ time steps.
5. When a node wakes up, if its lifetime has not expired yet, it creates a two-hop edge using the random-random triangle closing model.

Due to the nature of congress, there is not a significant number of nodes arriving once a session has started and nodes don’t wake up for periods of time since legislators are continuously active during sessions of Congress. Finally since a representative’s “lifetime” is simply the amount of time they are in office, the we simply run one round of two-hop edge creation or triadic closure for each node. For a given node, we would choose a neighbor, n, using one of four methods and then choose a neighbor of n using the same method. We ran four different types of triangle closure, random-random, degree-degree, random-random within party, and finally degree-degree within party. In our baseline, random-random, we randomly pick a neighbor, n, and then randomly pick a neighbor of n, m and finally create edge (node, m). For degree-degree, we picked the first neighbor proportionally to its degree and the same for its second neighbor. Random-random within party simply runs random random only choosing among members of the same party as the original node, while degree degree within party runs degree-degree choosing proportionally by degree among the node’s neighbors within the same party.

Results and Findings

Personalized Pagerank and **Supervised Random Walks**

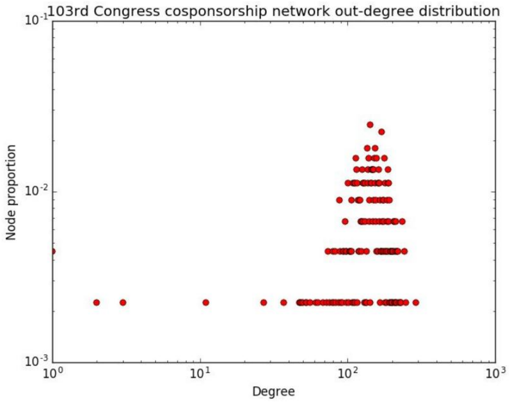


Fig 1 Out-degree distribution of one House session

Unfortunately, our Supervised Random Walks weight vector **w**, saw very little deviation from the **0** vector. Furthermore, we found the predicted pagerank scores on our test data, as described in our methodology above, to be uniform across all nodes. This is largely due to the inadequate data features available that led to a faulty training set. We could not adequately split the data into a training and test set because the granularity of times between co-sponsorships was not fine enough. This is because many legislators co-sponsor legislation right when it is introduced. Additionally, the Supervised Random Walks algorithm did not perform well with our co-sponsorship network's degree distribution shown. The degree distribution is fairly isomorphic and high in relation to the number of nodes (see Figure 1).

The results for the Personalized PageRank did not show promising results when evaluated on the test set. Unlike in Supervised Random Walks, we discovered non uniform pagerank scores. However, the notable variance in these scores was limited to the existing edges. The derived PageRank scores using Personalized PageRank did not provide meaningful signals to predict future co-sponsorships, with largely random PageRank scores.

We were unable to predict any successful bill co-sponsorships with a significant degree of accuracy using either Personalized PageRank or Supervised Random Walks.

Triads

We ran triads on the congressional sessions with-out the co-sponsor cap or sessions 97-109. We found the average accuracy of our added edges by calculating the percentage of our predicted edges that existed in the original graph. As shown in our results summarized in Table 4, we received a 4% bump in accuracy when only choosing edges within the original node's party but did not receive better results from choosing proportionally by degree. This makes sense since there node out degree is a highly right skewed distribution and while most representatives have out degree > 100, there are a representatives that have much lower out degree. Essentially choosing proportionally based on out degree is mostly the same as our random-random triadic baseline. To calculate the statistical significance of the triadic accuracies, we compared our results for each congressional session to the accuracy of looping through every node in the graph and creating a new edge to a random destination. Via the Mann-Whitney test, we found that every single triadic closure approach we tested was statistically significant.

Summary

In this project, we explored the notion of predicting bill success in one chamber of Congress using two main methods: binary classification using network features, and link prediction within the graph. We were able to find some potentially predictive features of bill subgraphs that indicate that certain network and node features of the Congressional co-sponsorship network may be indicative of bill success or failure. While we found moderate accuracy with Triadic closure, the Personalized PageRank and Supervised Random Walks were largely plagued by isomorphic degree distribution and dense graph structure. Given the nature of Congress, there are many things that play into the success or failure of a bill, including public perception, special interest groups, and the political climate of Washington. The co-sponsorship network is just one of several factors, and it's not enough to tell the full story of a bill. However, our results indicate that the cosponsorship network can be an informative factor in the success or failure of a bill.

Appendix

	93rd House		**103rd House**	
	U statistic	P value	U statistic	P value
Mean clustering coefficient	9443086.5	0.0902	2344384.0	0.1941
Mean number of sponsors	9110409.5	**4.2416e-05**	2134076.0	**3.6296e-07**
Mean in degree	8478781.0	**2.6156e-11**	2330189.5	0.13106
Mean out degree	6413777.0	**8.0468e-71**	2367365.0	0.3284
Mean PageRank	8758259.5	**2.2122e-07**	2320199.5	0.09622
Mean HITS Hub score	6658491.5	**6.0872e-61**	2367003.5	0.3260
Mean HITS Authorities score	8230861.5	**1.3132e-15**	2330820.5	0.1335
Mean betweenness centrality	8374944.5	**5.1396e-13**	2269831.5	**0.01318**
Mean closeness centrality	7720143.5	**6.6685e-27**	2312814.0	0.07522
Proportion of co-sponsors in same political party as sponsor	8799296.0	**9.8030e-11**	2189680.0	**3.1030e-05**
Proportion Democrat	7291282.0	**1.4609e-44**	2391436.5	0.4976
Proportion Republican	7545266.5	**8.5773e-37**	2377681.0	0.3948
Proportion Independent	N/A	N/A	2315399.0	**0.0001576**
Proportion of co-sponsors from same state as sponsor	9046365.5	**8.6183e-10**	2268059.5	**0.002803**
Proportion from CA	9501991.5	**0.02937**	2385608.0	0.4161

Proportion from TX	9261386.0	**3.00293e-11**	2375648.0	0.2285
Proportion from FL	9418208.0	**0.000111**	2355647.5	**0.02365**
Proportion from NY	8968385.0	**4.5711e-13**	2378919.0	0.2840
Proportion of co-sponsors of same gender as sponsor	9122669.5	**4.07811e-05**	2197107.0	**6.5345e-05**
Proportion women	9245695.0	**1.6661e-05**	2253033.0	**0.001553**
Proportion men	9097631.5	**4.02455e-07**	2301004.5	**0.04008**

Table 1 Mann Whitney test results for mean statistics. Statistically significant results are in boldface.

Feature (* indicates significance in 103rd Congress)	**Capped Congress sessions (93rd-95th)**	**Uncapped Congress sessions (97th-109th)**
Mean clustering coefficient	1.61783382e-02	
Mean number of sponsors*	8.60671189e-04	0.00064667
Mean in degree	5.07789747e-03	
Mean out degree	6.50157544e-03	
Mean PageRank	1.30916488e-02	
Mean HITS Hub score	1.63671940e-02	
Mean HITS Authorities score	1.44093480e-02	
Mean betweenness centrality*	1.39972555e-02	0.02086913
Mean closeness centrality	1.11629182e-02	
Proportion of co-sponsors in same political party as sponsor*	5.49782517e-03	0.01349688
Proportion Democrat	2.65442639e-03	
Proportion Republican	2.75891189e-03	
Proportion Independent*	0.00000000e+00	0.00562028
Proportion of co-sponsors from same state as sponsor*	6.42036678e-03	0.
Proportion from CA	1.20125953e-03	
Proportion from TX	1.89553495e-04	
Proportion from FL*	1.42685044e-04	0.00589299
Proportion from NY	9.14086620e-05	
Proportion of co-sponsors of same gender as sponsor*	1.10831165e-02	0.00553177

Proportion women*	4.28834847e-04	0.
Proportion men*	1.06496868e-02	0.00570794

Table 2 Mutual information scores of significant features for capped (93-95) and uncapped (97-109) Congresses

	Capped Congressional sessions (93, 94, 95)		**Uncapped Congressional sessions (99, 100, 106)**	
Features	**Average true positive rate**	**Average accuracy**	**Average true positive rate**	**Average accuracy**
All significant features	0.085444	0.8899	0.0750	0.8156
Top 3 features by mutual information	0.0020768	0.9392	0.02720	0.8362
Top 5 features by mutual information	0.020000	0.92629	0.01309	0.8302

Table 3 Average rates for leave-one-out-cross-validation experiments

	Random-random	Degree-degree	Random-random within party	Degree-degree within party
Average accuracy across sessions 97-109	0.589	0.581	0.631	0.641
P-value from Mann-Whitney	**8.601e-06**	**8.125e-06**	**7.548e-06**	**6.048e-06**

Table 4 Average Accuracy for Triadic closure.

Citations

1. Fowler, James H. “Legislative Co-sponsorship Networks in the U.S. House and Senate.” *Social Networks.* 28 (4): 454-465 (October 2006)

2. Burkett, Tracy. “Co-sponsorship in the United States Senate: A Network Analysis of Senate Communication and Leadership, 1973-1990” (Unpublished Doctoral Dissertation)

3. Liben‐Nowell, David, and Jon Kleinberg. "The link‐prediction problem for social networks." *Journal of the American society for information science and technology* 58.7 (2007): 1019-1031.

4. Backstrom, Lars, and Jure Leskovec. "Supervised random walks: predicting and recommending links in social networks." *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011.

5. J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins. Microscopic Evolution of Social Networks. In Proc. KDD 2008

Datasets used:
Fowler dataset: jhfowler.ucsd.edu/cosponsorship.htm
GitHub dataset: https://github.com/unitedstates/congress

Division of Labor

Cheenar: Feature extraction from subgraphs, ML predictions using subgraph features

Patricia: Data import and network creation, Link prediction using triadic closure
Gaspar: Link prediction using personalized PageRank and supervised random walk.