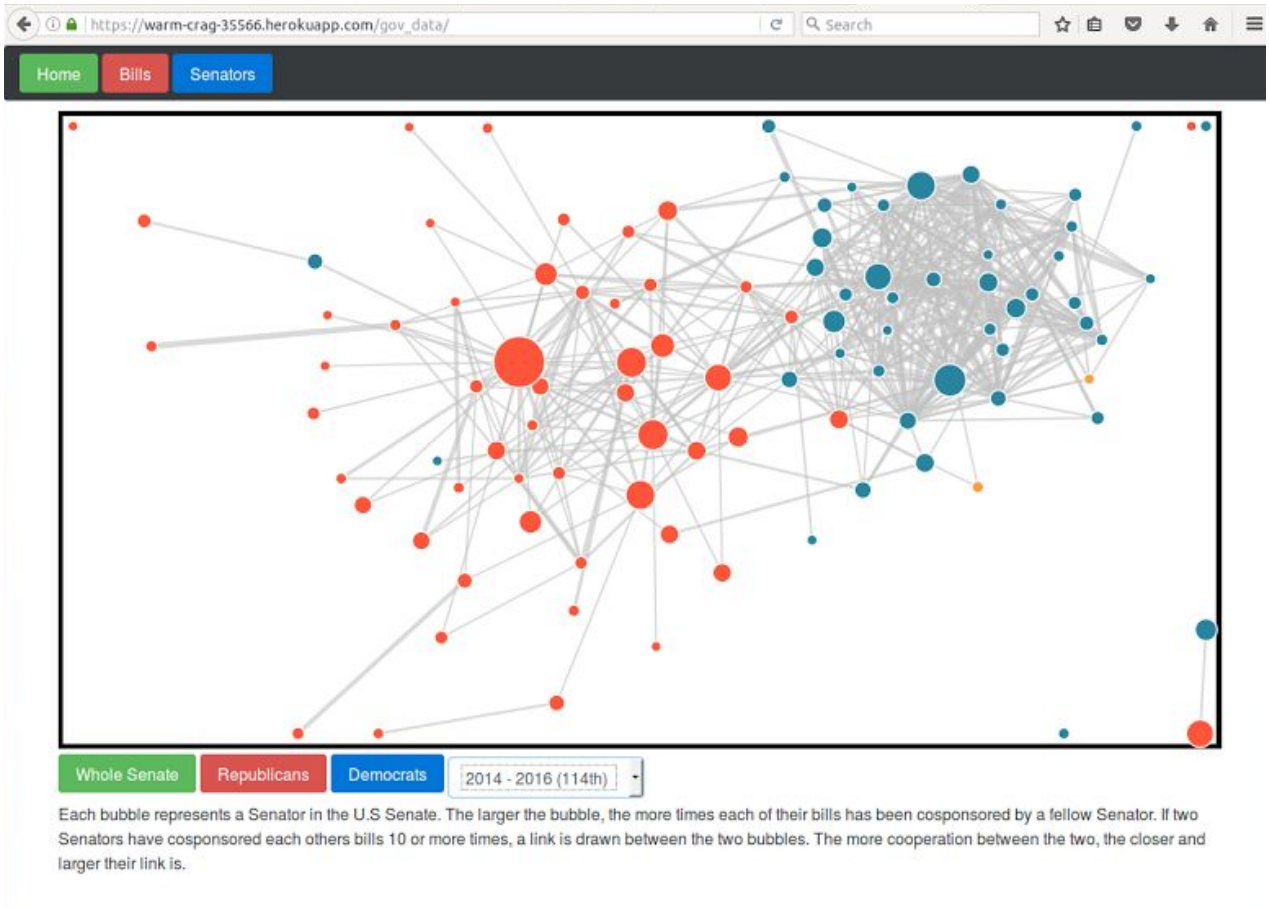
CAPP 30255 Final Project Proposal: A Combined Network and Text Analysis of Senatorial Influence

Gregory Adams



Introduction:

Congressional influence is a common topic of research for Public Policy. Viewing Congress as a network of individual actors connected by sponsorships and co-sponsorships of bills allows us to measure, however imperfectly, everything from legislative effectiveness to corruption. In my CS122 project,[1] I was able--with my

[1] The project code is located here: https://github.com/mvasiliou/Congressional-Cosponsor-Relationships.

partners--to construct a visualization[2] of this network using data from YouGov's congressional API. In this project, I hope to build off of that idea. I plan to use Natural Language Processing methods to predict the likelihood of different bills (or different types of bills) passing; Network Analysis, including the PageRank algorithm, to measure influence of different members of Congress; and more traditional statistical and machine-learning techniques to combine those approaches in an integrated classifier. I hope to develop basic insights into important questions like "how important is the topic or substance of a bill relative to the standing of its supporters?" and "which members of congress are best at getting unpopular bills passed?"

Methods:

For this project, I plan to limit my analysis in two important ways. First, I will only focus on the Senate. Limiting it to only one house drastically reduces the complexity of the model because it does not have to account for two independent networks. It also limits the data I need to gather, reduces the size of the network to a more manageable size, and eliminates the need to constantly check and recheck bills that get passed back and forth between the houses to prevent them skewing the analysis. Second, I limit my initial analysis to the 109th congress. This was the last time there was a not-obviously-polarized network (as shown by my visualizations) and neither party had a supermajority. Time permitting, I will expand this to other congresses (e.g. the 112th) which are more polarized, to note changes.

4 Main Goals:

1. Scrape data from congress.gov. Congress.gov has the full text of every bill introduced in the Senate, along with summary data (sponsorships, cosponsorships, etc.). Should this prove a problem, my backup is to use summaries from https://www.gpo.gov/fdsys/bulkdata/, which is designed for bulk data gathering.
2. Use the text of the bills to build a TF-IDF model (or models) of bill texts, with the goal of predicting whether or not it will pass. I may also integrate some of the summary data, including when the bill was introduced, which party introduced it, etc. into a more complex model. I could also extend this analysis to show what types of bills are most characteristic of each senator, more likely to pass, etc.

[2] The deployed visualization is located here: https://warm-crag-35566.herokuapp.com/gov_data/ (Note: one of my partners did the deployment independently after the project was submitted; this is just to show what our visualization looks like)

3. Measure influence--as defined by centrality--on the network. I have the network set up already as a networkx object for the visualization, so I would perform analyses on that existing data.
    a. I would use multiple measures of centrality, from the simplistic (e.g. raw counts of connections) to the sophisticated (e.g. PageRank)[3]
    b. I would also like to demonstrate correlation between influence of cosponsors and likelihood of passage. Potential methods include:
        i. Regression
            1. Logistic (passed/not passed)
            2. SLR (influence vs. stage bill reached, 1-9)
        ii. Combining centrality measures and building my own classifier
4. My ultimate goal is to sample from the above methods to build one, final, integrated classifier to predict whether or not a bill will pass the Senate. The final evaluation for this will be how effective the classifier is by F1 score. Optimistically, I plan to evaluate my final model by testing it on other congresses.

3 Optimistic Goals:

1. Run these analyses for other, more polarized congresses (e.g. 112th) to see what changes and how.
2. Develop a measure of polarization to include in the model.
3. Depending on time it takes to get the text data/build the models, compare F1s (or other eval. metrics) across varying levels of polarization -- see if it changes, and to what extent.

Tools/Libraries I will Use:

1. NetworkX (for network analysis incl. PageRank)
2. Scikit-learn (for text analysis and building the classifiers)
3. Sqlite
4. Various others, including OS, Sys, urllib, and django.

[3] A relevant discussion of different centrality measures can be found here: http://fowler.ucsd.edu/best_connected_congressperson.pdf

Schedule:

4th Week:	Build Web scraper and begin downloading the information about each of the bills. It will likely take a long time to pull down that many bills, so this needs to be done early so I can begin pulling down data quickly.
5th and 6th Week:	Being building text analysis models (Goal 2). Experiment with different types of models and optimize classifier.
7th Week:	Prepare for mid-quarter presentation. Write slides, make sure code is commented, polish project, etc.
8th and 9th Week:	Work on Goal 3: run the network centrality analyses and experiment with different types of models for predicting passage.
10th Week:	Integrate text and network models. Optimize and validate final model.
Finals Week:	Buffer Period. Prepare final presentation slides and polish project.