Money Moves the Pen

Link Prediction in Congress Bill Co-Sponsorship Networks Using Political Donor Network Information

Yi Zhong, Eddie Chen —— CS224W Fall 2018 II Class Project

Stanford | ENGINEERING
Computer Science

Introduction

Political collaboration is an important part of legislative life, and congress bill cosponsorships provide a rich source of information about the social network between legislators, and serving as a proxy to understand legislators' "connectedness" and collaboration graph. Moreover, according to Mark Twain, "we have the best government that money can buy" - money and politics have already been intertwined. In this project, we applied social network analysis tools on political donation networks and congress bill cosponsorship networks, and framed our research problem as a link prediction task on congress bill cosponsorship networks using political campaign donation records for the US (Congress and Presidential Campaigns) with its network characteristics. We modeled and presented graph characteristics of the two political networks, and showed investigation results of link prediction using various supervised learning techniques for this project. We then compared models' performance to a naive baseline to come up with evaluations.

Method

Our project is made up of two parts: graph modeling, and link prediction. For graph modeling, we aim to construct a tripartite graph of political committees, legislators (we will ignore those failed to get elected to office), and the bills those legislators worked together on. A sample graph can be found in Figure 1. With the graph constructed, we provide a set of statistics and descriptions of the graph structure (including their one-mode projections, for both bills-legislators and committees-legislators subgraphs). After that, we construct a link prediction problem by dividing graph into different years of congress, and select the suitable years for model training and evaluation. Lastly, we report our learnings from the entire exercise.

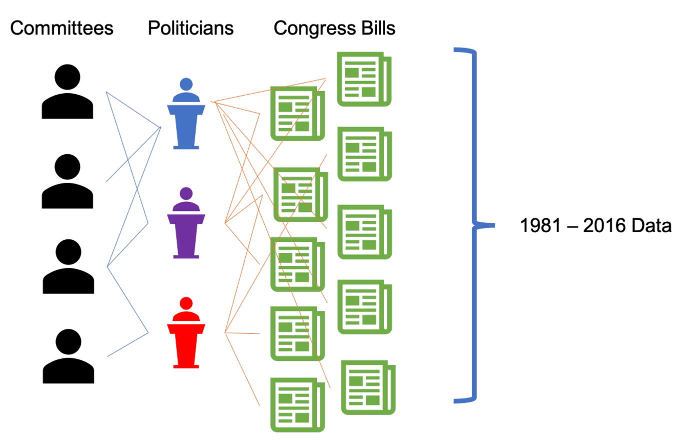


Figure 1: Illustration of the Congress Political Network

Link Prediction

We frame our link prediction problem as follows: predict the link between legislators, where a link exists if two legislators cosponored a bill together, for a specific Congress term. **Features to predict this link all come from the campaign network prior to the Congress going into session**, which is a bipartite network (legislators and committees), where a link between a committee node and a legislator node exists if the committee donates money to the legislator. We want to see if immediate donation has an effect on collaboration, hence we use campaign data two years before to predict the cosponsorship network during a congressional term.

For example, if we are predicting cosponsorship in the 100th congress (1987 - 1988), we would use campaign data from 1985 to 1986, in order to construct the features.

Features and Algorithms

We constructed features from the campaign subgraph solely. Features include:

- Common Neighbors, Union of Neighbors, Jaccard Index
- Degree Difference in a pair of legislator nodes
- Contribution Amount (sum and absolute difference)
- Clustering Co-efficient (sum, absolute difference, mean)
- Degree Centrality difference
- Shortest Distance between two legislator nodes
- Spectral Clusters from Clauset-Newman-Moore greedy modularity maximization

For algorithms, we used supervised learning models like **logistic regression** and **decision trees**. For logistic regression, we used scikit-learn's default implementation with -1,1 notation for labels and $L2$ regularization. The optimization problem formulation is as

$$min_{w,c}\frac{1}{2}w^Tw + C\sum_{i=1}^{n}\log(e^{(-y_i(X_i^Tw+c))+1})$$

Evaluation Method

We used accuracy as our main success measure:

$$Accuracy = \frac{NumberOfCorrectPredictions}{TotalNumberOfPredictionsMade}$$

We define our **naive baseline predicto**r as follows: given a pair of nodesv_1, v_2, we will always predict there will be an edge between these two pairs, i.e. as a complete graph. That is,

$$Accuracy_{NaiveBaseline} = \frac{2||E||}{||V||(||V||-1)}$$

Using the 100th Congress (1987 - 1988) as the training set and the 101th Congress (1989-1990) as the test set. The baseline accuracy is calculated as $96,052/138,075 = 0.695$.

Results and Findings

The basic stats of the tripartitie graph are included below:

- Legislator count: 1,919 (1813 of which are found in campaign financial network)
- Bill count: 221,726
- Committee count: 14,326
- Edges between legislators and bills: 3,086,039
- Edges between committees and candidates: 911,965
- Overall tripartite graph node count: 237,971, and edge count: 3,998,004

Network Description

we have plotted the degree distributions of the overall tripartite graph in 2.

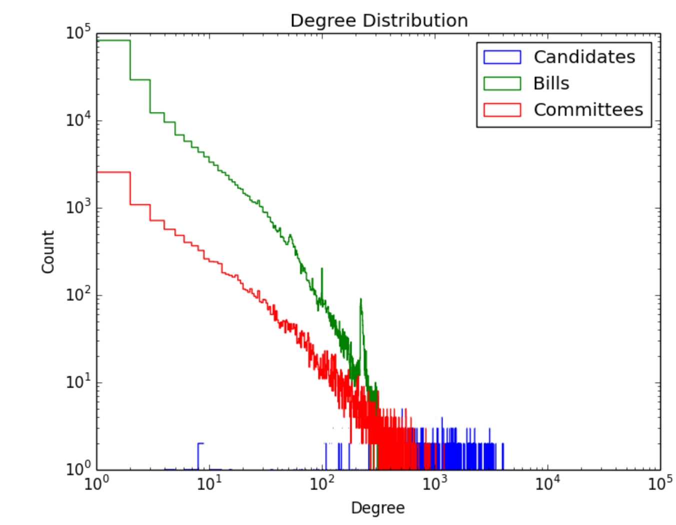
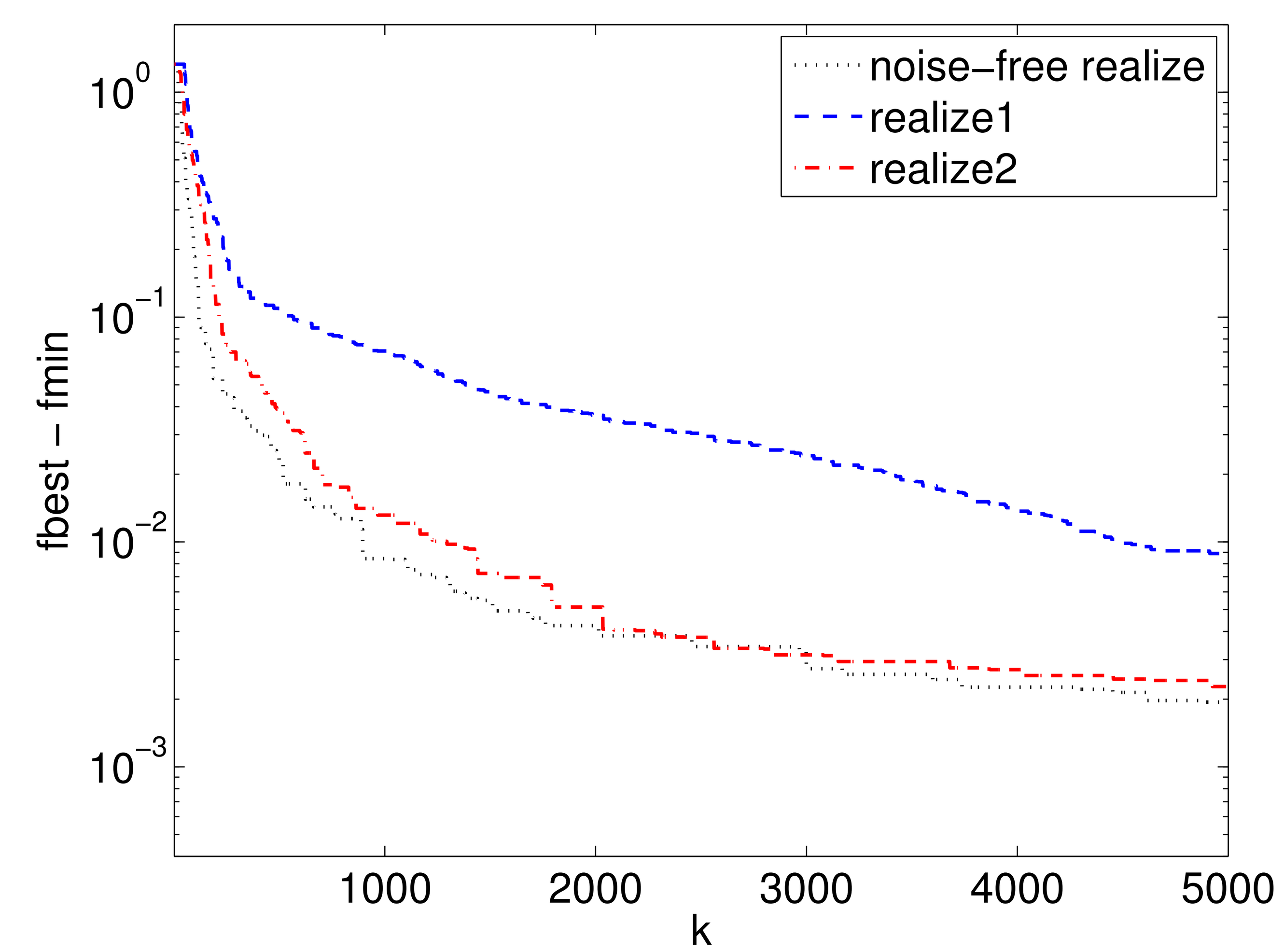


Figure 2: Overall Tripartite Graph Degree Distribution on log-log scale

Bill Co-authorship Graph resembles the academic collaboration graph with a power law pattern (long tail) - the most frequent degrees are the smallest degrees, and it has a very high clustering co-efficient. For the **folded campaign contribution graph, it represents a typical small world pattern with high clustering coefficient**.

Model Performance

PLACEHOLDER



Conclusion

US Congressional Politics is indeed a small world: legislators are connected to other legislators via common donors and co-authorship on bills. We have identified the academic collaboration network-like pattern for bill co-authorship data, and a "small world" pattern among legislators, with consistently high clustering coefficients. Moreover, it does appear that "money moves politics": using features learned from campaign donation networks, we can confidently predict if two legislators will later collaborate on bills together - easily beating a naive baseline. In paricular, decision tree model performed very well to give us 82% accuracy.