Identifying Factors in Congressional Bill Success

CS224w Project Milestone Report
Travis Gingerich, Montana Scher, Neeral Dodhia

The Problem

During an era of government where Congress has been criticized repeatedly as being the most polarized and unproductive (in terms of bills passed) in American history [9, 10], understanding group decision making and the power dynamics behind it has peaked our interest. Thus our project investigates the factors that are involved in the success of Congressional bills, and how we can use this data to develop a model for predicting the success of a bill.

Several metrics and techniques can be used to investigate the factors that make up a bill's success. We have started with some of the basic properties of bills, i.e. date bill is introduced, how partisan it is, the number of co-sponsors, etc… We will also investigate network centrality, collaboration networks between Congress members based on memberships of parties, committees, co-sponsors, etc can be used with an algorithm such as PageRank, or measures like size of influence set and network constraint to better inform our models. Another source of interest is the connection between bill success and campaign finance sources for which a tripartite graph linking campaign finance sources to candidates to bills they sponsored can be used.

We are also using the congressional voting data as another component in predicting bill success. We are interested in modeling individual voting behavior of congress members and investigating how Congress members relate to a particular bill across several dimensions. A starting point for this investigation is the matrix factorization methods presented by Paterek. Vectors modeling Congress members' preferences can be derived from historic voting data. When attempting to predict behavior on a new bill, a vector describing representing the bills' characteristics can be derived from the vectors modeling the bills' sponsors. These vectors could be used to predict the voting behavior of each Congress member individually, which can be aggregated to predict a bill's outcome.

In order to establish a basis for prediction will use a number of the aforementioned properties, based on how strong of an indicator they are, as features in a machine learning model. We will validate this model by splitting historical data on bills as well as voting records available from GovTrack.us into a training set and a test set. Extensions, time permitting, could include training the model multiple times with different features using cross validation and for datasets from different congress sessions.

Review of the relevant prior work

Governance in Social Media: A Case Study of the Wikipedia Promotion Process
by Leskovec, Huttenlocher, Kleinberg
This paper studies group decision making, using data from the Wikipedia promotion process. Using measures of assessment to understand how individual users will vote on a candidate and how these votes will aggregate to form a final election decision, Leskovec et al. show that there are

non-monotonic effects of relative merit and a clear difference in voting patterns when a voter has less merit versus more merit than the candidate. Since all votes are public and sequential, they notice a “herding” model that is representative of averages over populations but not any particular individuals. There is little insight into group deliberation habits.

Propagation of Trust and Distrust
by Guha, Kumar, Raghavan, Tomkins
In their paper ‘Propagation of Trust and Distrust’, Guha et al investigate the flow of trust (and distrust) through the web of trust. On a dataset from the Epinions website, they modeled the spread of trust as a matrix operation. Their factors included how trust is propagated: by transitivity (direct propagation), by symmetry (transpose), by implication of shared beliefs (co-citation), etc; whether distrust is propagated and, if it so, how. They ignored various factors which are not as important in the online world, such as distance between any two people, and ignoring the effect of homophily in age, gender, ethnic, etc..

Improving regularized singular value decomposition for collaborative filtering.
by Paterek, Arkadiusz
This paper describes several techniques used in the Netflix Prize competition, which uses historic ratings data to provide better recommendations. Paterek et al present several models including: regularized singular value decomposition (also referred to as matrix factorization in other works) of a user-rating matrix, K-means using vectors of user ratings, and using K-nearest neighbors with the results of matrix factorization. The most successful techniques built upon matrix factorization. The main weakness in their methods was limited information about the user or movies. For example, including data on demographics, users’ connections, genre, age and running time could improve the predictions.

Bill Prognosis Analysis.
GovTrack.us. 2013
GovTrack.us have performed similar work in analysing factors that contribute to a bill’s success. They split asked two questions: whether a bill will make it out of committee, and secondly. whether a bill will be enacted given it is out of committee. Two of their more interesting factors are *ideology* which measures the similarity between two senators’ beliefs and a *leadership* score which measures whether more people follow you than vice-versa (this uses PageRank on the co-sponsorship network). With around fifty features, many of which require domain-specific knowledge to understand, they trained a logistic regression classifier to predict whether a bill would succeed or not.

Data Collection

The data used is all publicly available at GovTrack.us and at from the Federal Electoral Commission website. GovTrack.us provides: the voting record of each Congress member, committee membership data, detailed information about each legislator, e.g. their party affiliation their position in congress, terms served, etc, and detailed information on bills, including their current status, past actions, sponsors and cosponsors, summaries and votes. Bills are organized by congress, the 2 year congressional term, and by bill type - house, senate, joint resolution, concurrent resolutions, and simple resolutions. For the purposes of this project, we will only be using the first three types as these are the forms used for all legislation concerning the public with the goal of becoming law. The Federal Election Commission provides data on campaign contributions to public officials from action committees and individual contributors. We are able to combine the information across different data files by using legislator and bill ids, as the library of congress has unique identifiers for both.

Initial Findings

Bill Success Level

As our goal is to predict the path a bill will take through congress, we used the bill data to formulate a standard for calculating the "success" of each bill. The ultimate enactment of a bill requires it to pass through a lengthy sequence of steps. Our notion of bill success stems from how far a bill makes it through this process. The basic steps are as follows...

1. Introduction of bill
2. Referral to committee
3. Reported out of committee to floor
4. Passed by House/Senate
5. Passed by opposite house
6. Signed by the president

After ingesting all of the data we were able to get a few summary statistics about the bills and where in the process they tend to die.
Number of bills that die after ...

Introduction: 1
Referral to Committee: 9206
Reported to Floor: 498
Passed by House: 70
Passed by Senate: 292
Passed by both: 2

Number Enacted: 277

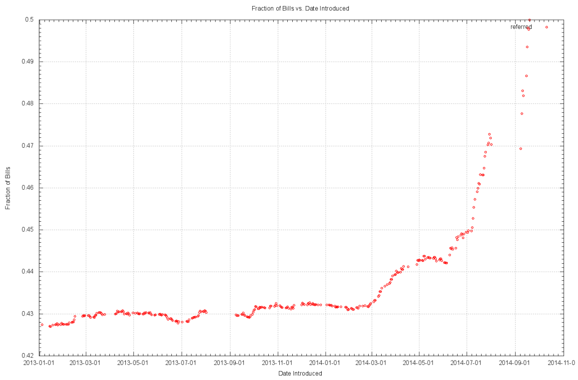
Bill Properties

We define our notion of a bill's "success rate" as the numerical step after which the bill dies or gets vetoed. Using this definition, we can now compare the relative success of different bills as well as try to find correlations between success and different dimensions of the bill. Some of the basic dimensions we have considered include the date the bill was introduced and the "bi-partisan-ness" of the bill's cosponsors, we plan to also look at the number of co-sponsors a bill has, the committee the bill was referred to, and the influence of the sponsor that introduced the bill.

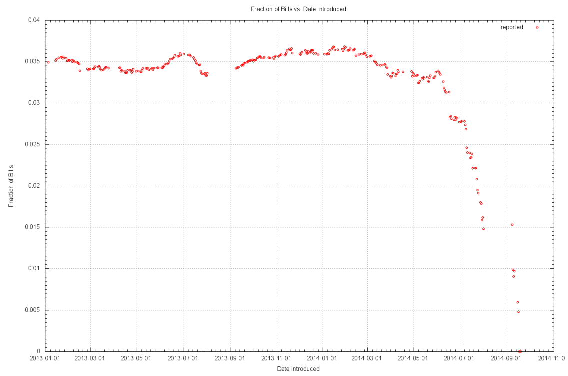
Bill Introduction Date

We thought there could be a relationship between the introduction date of the bill and its success given that there may be certain periods of the year, i.e. over the holidays or during campaigning and elections, during which congress would be less productive and thus could have an effect on how a bill advances. In order to understand whether the probability that a bill passes a certain step goes down over time, we graphed the fraction of bills that fail after a given step at time $X >= x$. You can see below the 5 different fail steps and how their probability changes based on the date they were introduced. While much of the data is noisy and does not give us any clear correlations, the most significant graph is that of % of bills enacted, graph E.

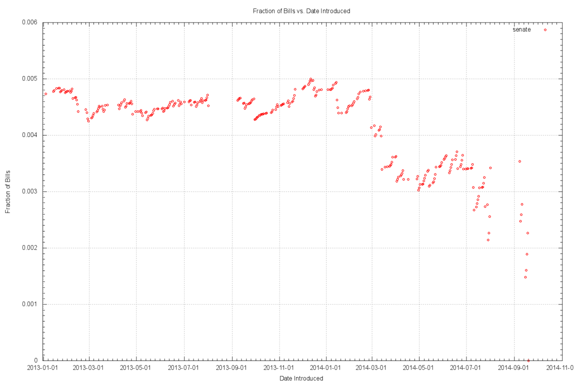
A. Died after Referral to Committee



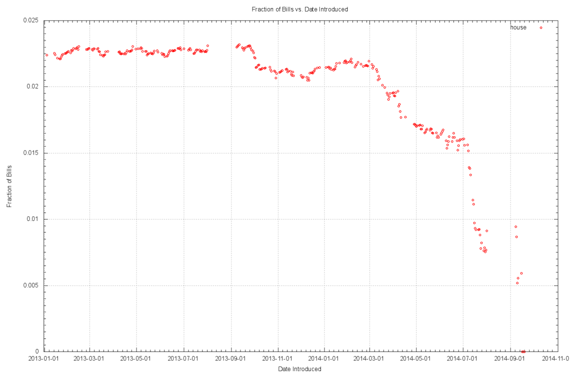
B. Died after Reported to floor



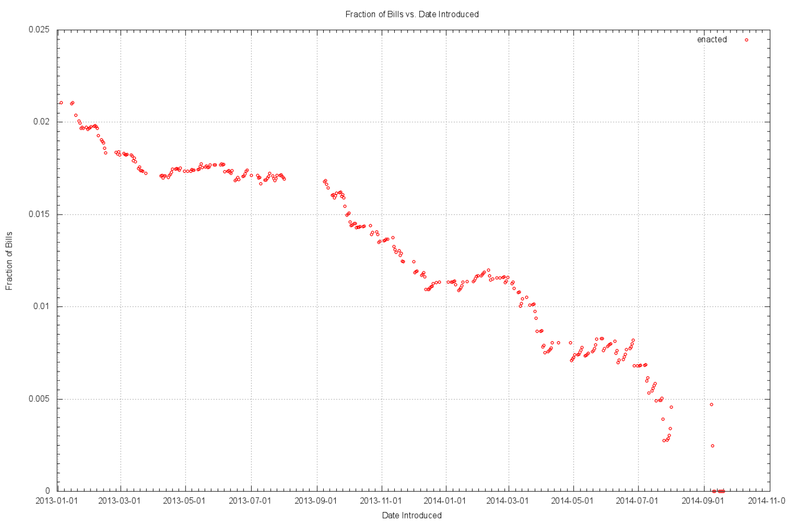
C. Died After Passed By Senate



D. Died After Passed By House



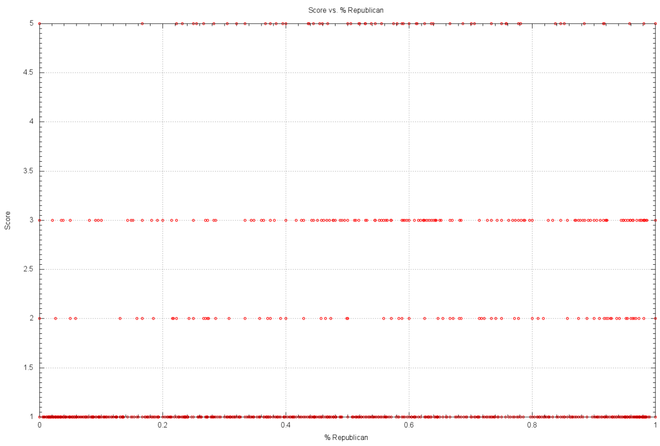
E. Bills That Were Enacted



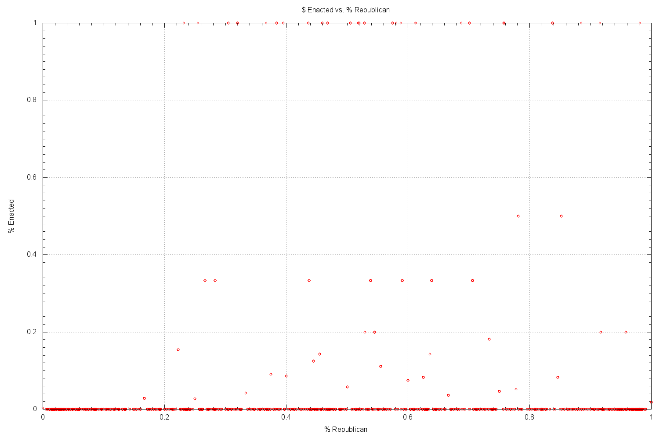
You can see a rather significant downwards trend in graph E, implying that the later a bill is introduced the lower the probability is has of being enacted. Only about 2.7% of all bills are enacted total, so a 1% decrease by the middle of the congressional term is rather significant.

“Bi-Partisanness” of Bill

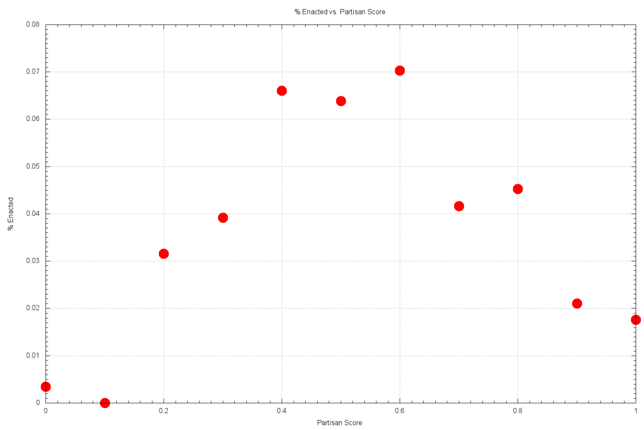
We also looked wanted to measure how bi-partisan a bill was. We did this simply by calculating the percentage of republicans that sponsored it. We graphed this percentage vs. the bill’s success level.



In the top graph to the left, you can see that there is not too much of a correlation between the co-sponsor make-up and the bill’s success. However if you look towards the top left corner of the graph, you see a gap, i.e. there is only one bill with less than 20% republican co-sponsors that was enacted. While this isn’t too surprising, as the republicans controlled the house, it could still be an interesting indicator.



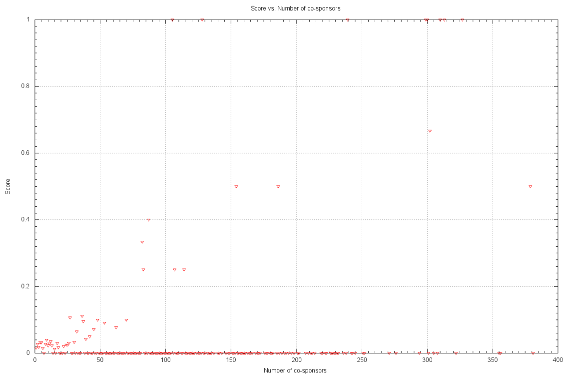
The following graph maps the probability of enactment given the percentage of republican co-sponsors. Again we see that if a bill has 20% or fewer republicans as co-sponsors, it has almost 0% probability of being enacted.

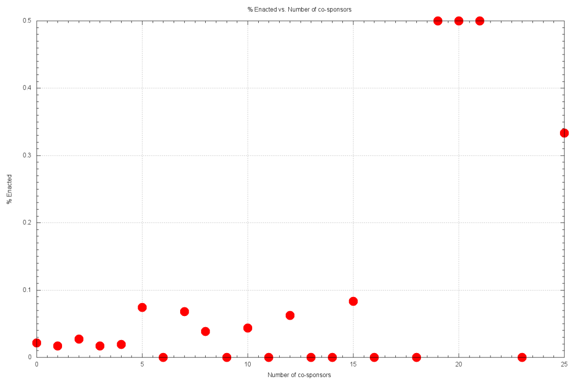


In order to more clearly see a correlation, the third graph bins the partisan score by rounding to the 10th of a percent. Since there is likely to be very few bills for a given percentage of republicans. By binning, you can actually see that there is a possible correlation between partisanship and bill success. Since on average a bill has a 2% chance of succeeding, and bills with almost equal partisanship have about a 7.5% probability of being enacted, thats a significant improvement. Indeed, it seems that the more even the co-sponsorship in terms of partisanship, the more likely it is to be enacted.

Number of Co-sponsors

Below are similar graphs as above however with the x dimension being the number of co-sponsors a bill has. We have the % enacted vs the number of co-sponsors, where the right graph has binned the number of co-sponsors by 15.





After initially looking through the bill data, it seemed that bills with a large number of co-sponsors could fall into two categories, those that everyone wants to support because they aren't very controversial and deal with things like honoring war heros, or they are very partisan and controversial, .e.g. a bill to repeal the affordable care act has 189 co-sponsors. It seems there is a possible correlation between more co-sponsors and stronger likelihood of success, however it doesn't seem to be a very strong indicator.

Referral Committee

Since the large majority of bills die in committee and never make it to the house floor, in fact, only 15%, we feel this could be a very interesting area to explore. There is an entire hearing and voting process that goes on inside a committee for a given bill, thus there is potential to dig into this information more deeply to see if either any actions taken by the committee or the committee itself could be an indicator for a bill that makes it to the house floor.

Sponsor

The bill's sponsor also has the potential to affect a bill's outcome. Our goal here is to generate an influence score for each legislator and relate it to their bill's success rates. This would be in a similar vein as the govtrack.us' leadership score, however in their case they use this as one of many low-weight features in their prediction model. We are hoping to find a stronger indicator, and can use their leadership score as a benchmarl. There are many ways we plan to generate an influence score including calculating the size of the sponsor's influence set using breadth first search in two possible networks, a directed graph of co-sponsors ← sponsors, and a directed graph of committee-members ← committee-chairs. We will also calculate the network constraint for a legislator in both of these graphs. Other indicators of influence we will look at are how much money each legislator received during campaigning and the length of time they have spent in congress.

Bill Graphs

Beyond looking at the basic aspects of a bill, we worked on generating graphs that could give us information about the structure of the congressional network and the bills they introduce. Below is a table with overview information for three of the graphs that we've generated so far.

	Bill <--[cosponsors]--> Bill	Legislator <--[cosponsors]--> Legislator	Committee co-membership
Nodes	10439	549	537
Edges	2145385	63173	19417
Zero Deg Nodes	938	0	6
Non-zero In-Out Deg Nodes	9501	549	531
Clustering coefficient	0.668244	0.750051	0.700155
Effective Diameter	2	1	2.789925
Full Diameter	5	4	
Max degree	29		140

Bill to Bill

The first graph looks at how bills are embedded in a bill to bill network, where the edges are common co-sponsors. You can see from the clustering coefficient and diameter that this network is pretty tightly connected. We will use this graph to calculate the centrality of a bill using a pagerank algorithm. Once we have that, we can see how this may correlate to a bill's success.

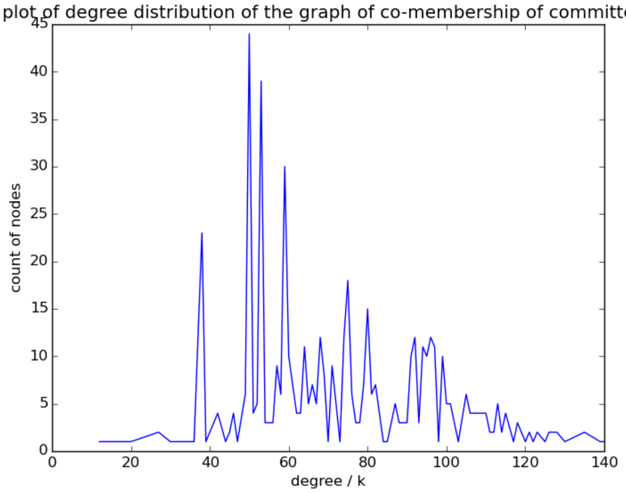
Legislator Co-sponsorship

This second graph represents how much collaboration there is amongst members of congress on bills. Again, this network is very well connected, which we found slightly surprising. The current congress has gotten much attention for being highly partisan and unable to compromise or collaborate. We want to investigate the properties of this network further by looking primarily at the degree distribution, which will help us understand whether this connectivity is due to only a few powerful legislators.

Committee Co-membership

Another aspect we looked at was the network on committee co-membership, where there is a node for each legislator and an edge between two nodes if the corresponding two legislators sit in the same committee. Apart from six legislators, the remainder all are connected. The mode of the degree counts is 45 degree. The plot shows that most legislators are well-networked by committee and so know a lot of their colleagues. There are very few legislators who know less than 30 others.

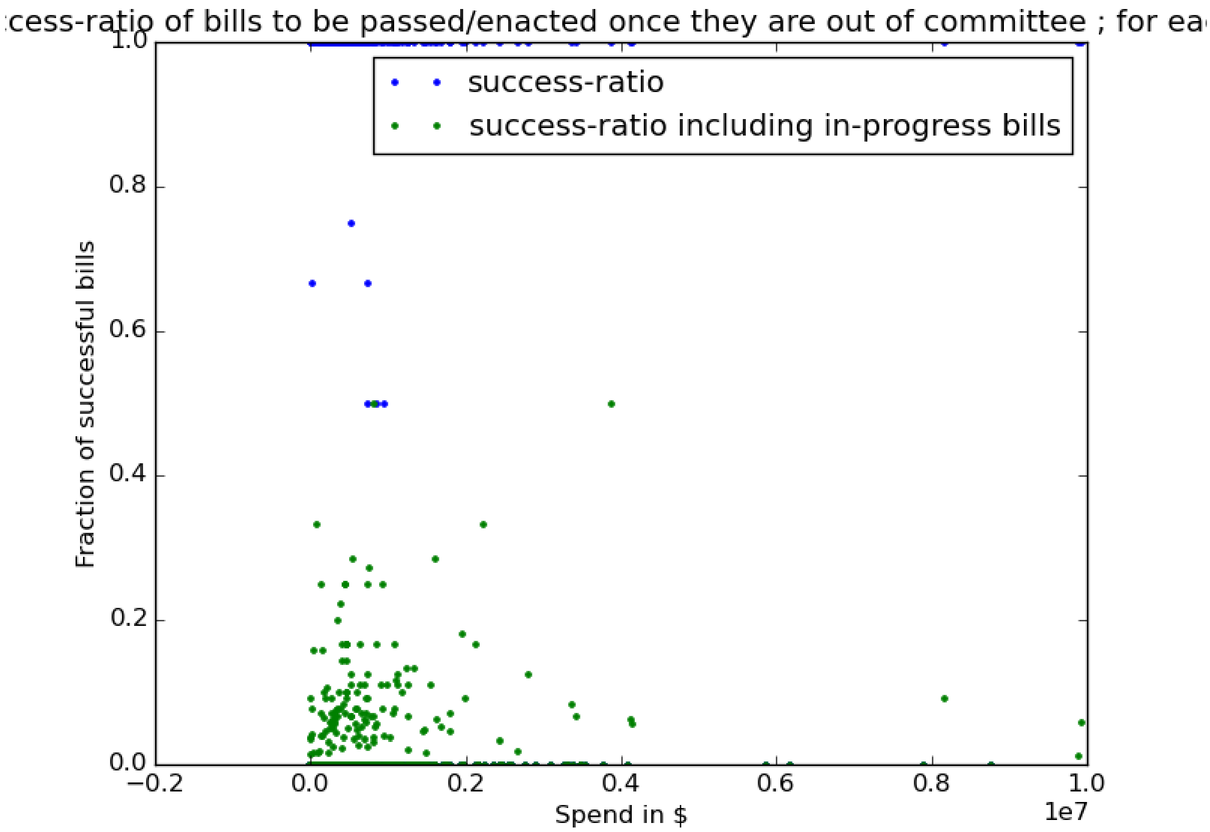
This makes intuitive sense as politicians are good at networking and will seek out opportunities to do so.



Financial Contributions

The link between campaign financing and bill success is another area we investigated. Intuition would lead to the belief that the more money spent on a bill, in various activities like lobbying, publicizing, on interest groups, would improve a bill's chance of success. The Federal Electoral Commission provides data on all contributions made by a PAC, party committee, candidate committee, or other federal committee to particular candidates.

This data was aggregated to a candidate level. Hence, the bills success can be viewed through two angles: from the candidate and from the bill. Although there were 8750 bills in the previous session, only 200 of them reached a completed state (whether that be successful or unsuccessful). Treating all the remaining bills as in progress skewed the results of the plots by having the majority of points in a line along the bottom x-axis. Excluding did not reveal any immediate trend. As mentioned above, the majority of bills get stuck in the committee stage. Separating the dataset into two subsets, of those bills that have not been reported out of committee and those that have, would allow more meaningful questions to be asked for each. On the dataset of bills that have already made it out of committee, there is a linear relationship between spending and bill success between \$0 and \$40,000. There are very few examples that have spending of more than \$40000 so making any inference of trends in this range is not possible.

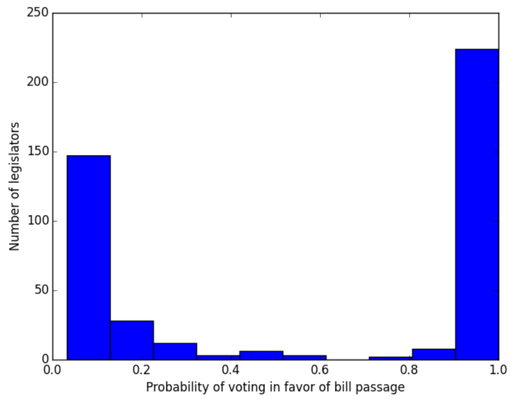


While we were not able to find any strong indicators of campaign funding to bill success, we will continue to use this data to inform our influence scores for legislators.

Bill Passage Votes

The votes of individual congress members are of interest in predicting a bill's success. One technique to evaluate is the use of matrix factorization to predict whether a legislator will vote for or against a bill. Matrix factorization is a technique used extensively by recommender systems to predict user's ratings of items based on a sparse set of known ratings (1). The basic idea is that a $|U| \times |I|$ user-item rating matrix R can be factored into matrices P and Q, where P is a $|U| \times k$ matrix representing a set of k latent features representing the user's preferences, and Q is a $|I| \times k$ matrix representing each item by the same set of k latent features. The product $P * Q^T = \hat{R}$ is used to predict each user's rating across all items. P and Q can be found by using gradient descent to minimize the mean squared error of the resulting rating predictions.

Initial findings are based on analysis of votes for bill and resolution passage from the current session of congress, which consists of a set of 279 votes. The most striking feature of this data is the tendency of individual legislators to consistently vote for or against each bill presented for vote. Why this is the case merits further investigation. The figure below shows a histogram of each legislator's ratio of votes in favor and votes against passage of presented bills.

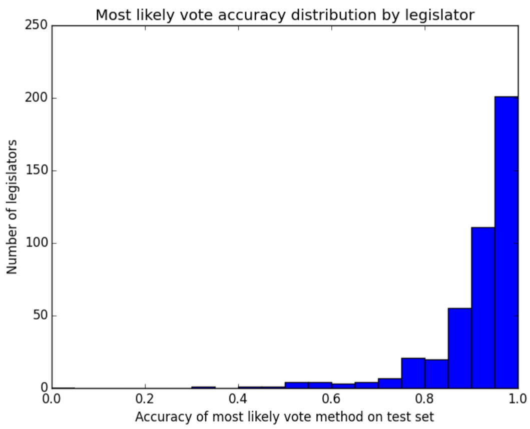


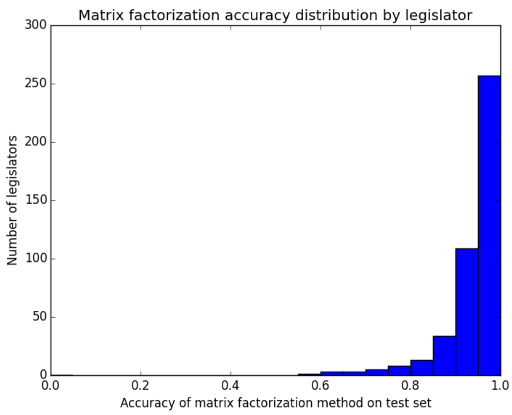
For initial data exploration, 15% of known votes (i.e. votes in which the legislator participated in the vote) are withheld as a test set, and the rest of the data is used for training. These are randomly selected on a per-legislator basis. Training is done on the remaining 85% of known votes. Accuracy is computed over the test set. For initial exploration this technique seems sufficient. A more rigorous approach would involve cross-validation over several folds.

A simple matrix factorization method with random initialization (with values between 0 and 1) of a set of k=40 latent features and subsequent gradient descent was used. The initial few iterations result in a dramatic drop in error on the training data, and subsequent iterations gradually decrease error as the slope levels out. 2500 iterations were performed.

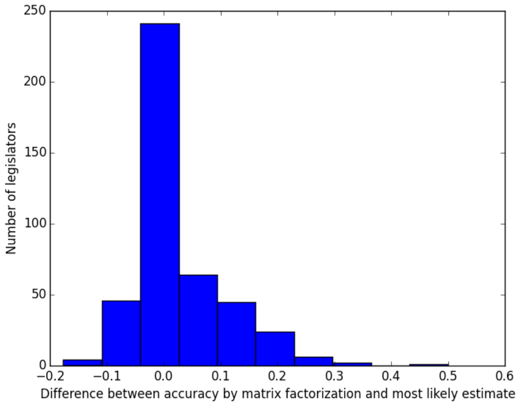
The consistency of each legislator's voting record implies that a very simple technique of always choosing the vote the legislator is most likely to make would perform quite well. This is used as a baseline against which to evaluate matrix factorization.

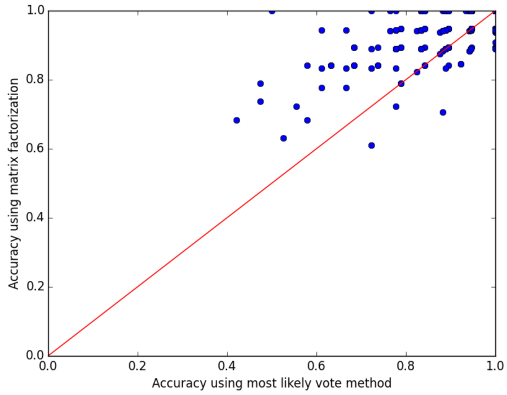
Using the most likely vote gave surprisingly good results, giving 93.0% accuracy on the training set. Matrix factorization improved upon this significantly, giving 96.0% accuracy. The distribution of accuracies by each legislator is shown below for both methods. Matrix factorization improves overall accuracy and reduces the spread of the distribution.





For most legislators, matrix factorization provides better estimation of their votes. The histogram below shows the distribution of the differences between prediction accuracy using matrix factorization and using the baseline technique. The second figure below plots the accuracies of both techniques for each legislator; points below the red line indicate that matrix factorization performed worse than choosing the legislator's most frequent vote.





Summary and Remaining Work

We have made headway in understanding different aspects of the data, e.g. the stages a bill goes through to become law, the likelihood of becoming law, the connectedness of legislators, the funding that legislators receive, and their individual voting behavior. We still have many properties of both legislators and bills that we plan to explore to see if we can find stronger indicators of bill success, most notably the "influence score" of a legislator. As explained previously, we will analyze how legislators are embedded in the congressional network among other legislator properties in order to develop a meaningful indicator of influence. Once we have this defined, we can use that measure to potentially weight individual voting behavior to more accurately predict a bill's outcome.

Another area we will explore more in depth is the structure of the congressional and bill networks. We will both do more analysis on the graphs we have already generated as well as look at a few more graphs of interest. Below is the list of relationships we plan to look at, both to better understand the collaboration that goes on in congress as well as the influence/leadership different legislators have.

Types of Networks and Relationships
Congressman ---[co-sponsorship]--- Congressman (done)
Congressman ---[co-membership of committee]--- Congressman (done)
Bill ----[shared sponsor]--- Bill (done)
Interest group ----[contribution]----> Congressman
Congressman ---[vote (+ | -)]---> Bill
Congressman ---[vote (+ | -)]---> Bill Sponsor
Interest group ---[contribution]---> Congressman ---[vote (+ | -)]---> Bill

Once we have identified the bill, legislative, and congressional network properties that most strongly indicate bill success, we will use these as features in a machine learning model to develop a prognosis for any given bill. Our method for testing this model will be similar to how the matrix.

factorization for individual voting prediction was tested. We will train our model with a portion of the congressional bill data, and test our accuracy on the remaining portion.

REFERENCES

(1) Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." *Computer* 42.8 (2009): 30-37.

(2) http://www.fec.gov/finance/disclosure/metadata/DataDictionaryContributionstoCandidates.shtml

(3) GovTrack.us. 2013. Ideology Analysis of Members of Congress. Accessed at https://www.govtrack.us/about/analysis.

(4) GovTrack.us. 2013. Leadership Analysis of Members of Congress. Accessed at https://www.govtrack.us/about/analysis.

(5) GovTrack.us. 2013. Bill Prognosis Analysis. Accessed at https://www.govtrack.us/about/analysis.

(6) Jure Leskovec, Daniel Huttenlocher, Jon Kleinberg. “Predicting positive and negative links in online social networks.” Proceedings of the 19th international conference on World wide web, April 26-30, 2010, Raleigh, North Carolina, USA

(7) R. Guha, Ravi Kumar, Prabhakar Raghavan, Andrew Tomkins. “Propagation of trust and distrust.” Proceedings of the 13th international conference on World Wide Web, May 17-20, 2004, New York, NY, USA

(8) Paterek, Arkadiusz. “Improving regularized singular value decomposition for collaborative filtering.” Proceedings of KDD cup and workshop. Vol. 2007. 2007.

(9) Terkel, Amanda. "112th Congress Set To Become Most Unproductive Since 1940s." The Huffington Post. TheHuffingtonPost.com, 28 Dec. 2012. Web. 16 Oct. 2014.

(10) "How Congress Became the Most Polarized and Unproductive It’s Ever Been." Washington Post. The Washington Post. Web. 16 Oct. 2014.