**Network and Predictive Analysis on Proposed Legislation in**

**the United States Congress**

Spencer Woodfork [Leader] G00800343  
 Applied Information Technology  
 George Mason University   
 swoodfo4@gmu.edu

Dilpreet Singh  
G00973687  
 Applied Information Technology  
George Mason University   
 dsingh14@gmu.edu

Meena Rapaka   
G01101759  
 Applied Information Technology  
George Mason University   
 mrapaka@gmu.edu

***Abstract - In this project, we perform network analysis legislators from the past 50 decades and try to create a model to predict the legislative outcome of some bill introduced to the US Congress. Significant literature already exists on the subject, including studies which involve analysis of Roll-call data [2, 12], text analysis of bill content [15, 6, 4], and studies which employ Machine Learning and other Big Data techniques [3, 7, 5, 13]. In our study, we extracted data from historical legislative records [11] and transferred these to a RDBMS via our Extract-Transform-Load (ETL) process. The ETL process used employs Python’s Element Tree library to extract Legislative Metadata and placed this in a SQL database. Once there, we will begin building a predictive model using the data returned in our SQL queries, combined with historical political information [9] and network analysis from Gephi. Utilizing rapidly growing technologies of the 2000s millennia which store high volumes of volatile data, one can receive real time, population-based views of what people are going through on a day by day basis. ProPublica Data Store is one of the largest government related data repository platforms which provides detailed information as such, and even has a supported platform known as ProPublica Data Store API which enables those with expertise to perform statistical analytics and aggregations on the volatile data available. Due to the available of this free resource we can continually improve our model by testing the current model against present-day congressional data. New Legislative data is made available by US Congressional Organizations constantly and patterns are bound to be presents for many of the information saved daily***.

***Keywords***

*Data Collection, Legislators, Democrats, Republicans, Congress, United States, Network Analysis*

1. **Introduction**

For a given bill being considered in congress, reliable metrics of how likely the bill will pass and become legislation are not easily accessible to everyday people. While making an educated guess might not be a problem for those working on Capitol Hill, everyday people must resort to written and TV journalism to get the information they need to make this kind of prediction.

While Journalism is an essential part of a free society, it isn’t always a good source for predictive analytics. News Outlets are subject to bias, and their vested interest in delivering information as quickly as possible can make what they report sometimes unreliable. This becomes a problem when a bill’s passage has significant and short-term and long-term effects on the livelihoods of everyday people. These people likely want to know, in plain English as possible, how much they should count on this legislative proposal becoming law. In the course of determining this information via traditional journalism, they may be exposed to a lot of information they don’t need and may be misled into thinking a bill will likely pass when the opposite is true (or vice-versa).

Examples of bills whose passage carries this level of gravity are widespread, including in recent history. Legislative proposals to end the Affordable Healthcare Act which were brought to congress in 2017. Americans already receiving benefits from the AHCA could only watch the news and bite their nails for fear of losing coverage until it was finally revealed that the legislation would not pass. The Farm Bill presents a similar case. It must be renewed every 5 years, and access to government subsidies on which farmers depend for their livelihood depends entirely on this bill’s passage. The most dramatic example came this year in which failure to come to an agreement on legislation concerning a border wall put the government into partial shutdown for 35 days [10]. Government workers furloughed during this time missed not one but two paychecks. It was impossible for them to tell when they would be paid again. To make matters worse, the uncertainty they were facing made a temporary job even more complicated, as Temp Agencies felt that furloughed workers were “too temporary” even for temp jobs [8].

Big Data Analysis gives us an opportunity to drastically reduce this uncertainty. If historical records of bills handled by the US Congress are analyzed in the right way, a predictive analytical model could potentially be created. This model could then be applied to real bills, and it could yield a simple percentage likelihood for a bill’s Congressional approval. Even better, the algorithmic process which yields these predictions could be improved over time as new bills are brought into congress, as this adds new data for the algorithm to consider. If done right, predicting whether a bill becomes law could be like predicting the weather. This could bring certainty, stability, and relief to millions of Americans whose lives depend on what happens in Congress.

Social media plays a large role in expressing the struggles many Americans suffered during the congress bills. ProPublica Data Store along with many other social media platforms, were able to provide an outlet for the American public, which during the time of the bills provided significant amounts of information on monthly basis in real time.

By performing Networking Analysis, Average Degree Report, Modularity, Graph Density, Betweenness Centrality, Microsoft Power BI, or any other techniques which we deem necessary, we can predict the legislative outcome of some bill introduced to the US Congress. With this information we could answer various key questions such as:

*1. Can measurable relationships between legislators in United States congress be observed based on co-sponsor support of legislation proposed?*

*2. What type of communities/cluster are existent based on the co-sponsor support of legislation?*

*3. How does proposed legislation differ from the communities discovered?*

*4. Can a predictive model be created using various indicators which can predict the outcome of legislation currently being proposed in congress?*

1. **Related Work**

Our planned contribution to this area of study is a model of the full lifecycle of a bill, from inception to final vote. The existing research in this area includes statistical analysis of congressional behavior, theoretical models of ideology and text analysis of the contents of bills.

1. *Exiting work in Roll Call*

The most common topic for statistical analysis papers focused on United States Government legislative processes is an analysis of roll call data. Roll call data contains a record of how each congressperson voted for a piece of legislation but not all legislative motions or bills go through the roll call process. Roll call votes and teller votes are the only records of how specific congress people voted that are available to the American public; other types of votes include voice votes and division or standing votes where the final tally is recorded but not the behavior of an individual senator or house representative. An analysis of roll call votes is therefore useful for analyzing the behavior of legislatures but does encompass all the ways a bill can become a law.

Examples of these studies include one by Clinton, Jackman, and Rivers developed “a Bayesian procedure for estimation and inference for spatial models of roll call voting,” and sought to make the model flexible with an option to include ideological background information of each Congressperson. The overall goal of the paper is to provide a static framework for the behavior of individuals based on a measurement of their legislative preference. Their findings were that Bayesian simulations are useful in legislative predictions because they allow the analyst to incorporative substantive information about each proposal, the ideological behavior of legislatures, and theoretically implied constraints on the standard model [2]. It has been demonstrated in other studies, however, that the behavior of Congresspersons isn’t necessarily static.

For instance, in Roberts’ article The Statistical Analysis of Roll-Call Data: A Cautionary Tale, the effects of congressional procedure on the behavior of a Congressperson are examined. Here, he focuses “on recent efforts to measure party effects and ideological alignments, and that the composition of the roll call record can affect these measures.” His findings were party affiliations and ideologies change over time and that seemingly small procedural changes—such as allowing a motion to reconsider—have a large impact on voting behavior [12]. Roll-Call data is not the only perspective-of-interest when predicting the outcome of some bill, however.

In analyzing Roll-Call data, the question “Who voted for this bill?” is examined thoroughly. Another natural question to ask is “will this bill pass?” Recent efforts thoroughly examine this question by predicting the success of bills based on predictive models. GovTrack, an open-government data site, provides open source analytics on historical and ongoing legislation, including predictors as to whether a bill will pass. In an interview with the Washington Post [14], GovTrack founder Joshua Tauberer spoke about general predictors relating to the legislator introducing the bill and the history of the bill itself. Specifically, when the legislator introducing the bill “is a member of the chamber’s majority or the chair of the assigned committee, it is more likely to pass.” However, if the bill itself had already been introduced in a previous congress, this “actually correlated negatively with making it out of Congress this time,” says Tauberer.

1. *Bill Text Analysis*

GovTrack’s findings illustrate how the progress of legislation – from proposal, through committee, to debate and vote - affect the outcome of a bill. Another perspective is considered by Adler et al [1]. They propose that (in addition to congressional process) the content of a bill is a critical predictor for understanding congressional research. The authors include a summary of past studies in “keys to legislative success” including if legislators concentrate on specific issues or if they adopt a “shotgun approach of introducing lots of bills.” Controlling for bill type is particularly significant in assessing the importance and the urgency of a bill. The authors complete an analysis of the 102nd Congress and categorize bills as narrow in scope or “trivial”, particularly “urgent”, and all others as “discretionary.” While “trivial” and “urgent” bills make up only 12 percent of all bills introduced to Congress, they make up 45% of all the bills enacted into law [1].

Other studies in analyzing bill content include one by Yano, Smith, and Wilkinson. They created a predictive model to assess if a bill will pass through committee by applying text analysis to the contents of the bill. Their body of data included “each bill - with its title, text, committee referral(s), and a binary value indicating whether or not the committee reported the bill to the chamber... [and metadata including the] sponsor’s name, from each bill’s summary page provided by the Library of Congress” [15]. Of the existing literature, this paper is the closest to the project we plan to complete; the primary difference is that the focus of this paper was on one part of the bill’s lifecycle and our project will attempt an end to end analysis.

Continued research in this area shows that text analysis of bill content can in turn help predict roll call data. Kraft, Jain, and Rush developed a predictive model for legislative roll-call votes using the bill text as input and multidimensional ideal vectors for legislators. Their model is a bilinear with low-dimensional embedding and it captures the past voting behavior of congress people as predictors for future votes [6].

In a similar study, Gerrish and Blei developed several predictive models that use legislative text to predict sentiment. These models included ideal point estimations, topic models, and voting patterns of the sponsors. The ultimate result of their models is that the authors were able to predict specific voting patterns of Congress with high accuracy. The scope of the study was limited to text analysis and bill sponsors; 4,743 unique samples for training and analysis were pulled from 12 years of Congressional data history [4].

1. *Machine Learning*

Advances in machine learning and big data analytics offer yet another source of predictive models for congressional behavior and law. Explorations in this area move beyond roll call voting and explore topics such as ideological affiliation, the behavior of politicians, and how those factors can be measured and predicted.

Some studies in machine learning have already been applied to Congressional behavior and Legislative outcomes specifically. Examples of these studies include Diermeier, Godbout, Yu, and Kaufmann, who use the super learning Support Vector Machines text classification algorithm to predict senators’ ideological positions with a 92 percent level of accuracy [3]. They draw a clear distinction between party affiliation and ideological alignment, demonstrating how a political party can go through ideological shifts over time. We can take their findings of separating party and ideology and apply it to the predictors we generate for our own project.

Other studies using Big Data Analytics which are specific to this area of interest include one by Laver, Benoit, and Garry. They extracted a policy dimension from the text of proposed bills by treating words as merely data points, not as contextual words to be understood or interpreted. They created policy estimates from text of proposed bills in Ireland, the United Kingdom, and Germany—proving that their method had consistent results across different languages. The results of their computerized word scoring method performed comparably against the more time intensive human analysis [7]. Other studies done in Machine Learning may not apply to legislative outcomes of congressional behavior specifically, but their findings can nonetheless be utilized in our own investigation. For instance, Phillips explores the current and future applications of artificial intelligence with respect to the legal industry, interpreting laws and generating legal advice. He argues that machine learning applied to a legal corpus of documents can give companies a competitive advantage by improving their ability to make informed legal decisions and form outlines for effective legal arguments [5]. This study provides more evidence on the feasibility of predicting some legal outcome using the text of legal documents, which is one possible avenue we can consider in our own study.

A more specific example of Text Analysis which could be useful in our own investigation is a study by Vargo and Hopp. They identify and quantify incivility in Twitter posts using a Python-based semantic analysis involving wordlists. Through their analysis, they were able to create a scoring system that ranged “from 0 (completely civil) to 23.00 (highly in civil).” In addition, they were able to correlate this quantified incivility metric information with characteristics of the region in which the Tweet author came from. These characteristics include socioeconomic Status, social capital, and partisan polarity. In the same way that a Tweet may be mostly civil or uncivil, the language in a Bill may be mostly positive or negative. The analysis used here may serve as useful inspiration when computing the "sentiment score” for the language in some bill and correlating this with likelihood of becoming law [13].

1. **Proposed Approaches**

First, our team performed extensive network analysis on every legislator which resides in the ProPublica database. Once we extracted them all into our MySql database we were able to create a node file which contained important attributes corresponding to a legislator such as the party affiliation, elected position, etc. However what ProPublica does not have is a predefined structure to relating legislators to each other (edges) like most social media platforms which use likes or follows as edges. This however ties into the network model we design which uses the cosponsor ship as an edge between legislator nodes. This lead to the development of a recursive SQL query which would construct a temporary holding the edge for all legislators based on a unique identifier for each legislator. We quickly noticed that the number of edges was immense and would lead to a very dense graph. This is tied to the fact that despite different political viewpoints, it is not uncommon for one individual in a political party to cosponsor with someone outside of their own party if the legislation proposed aligns with their ideals. However, it takes just one cosponsor ship to create the edge, and we felt that it did not truly represent and put emphasis on legislators who cosponsor the same legislator frequently. This led us to create 3 different models in our network analysis.

* Model 1: At least one co-sponsorship equals a relationship
* Model 2: At least 10 co-sponsorships equals a relationship
* Model 3:  At least 20 co-sponsorships equals a relationship

Secondly, our team built a predictive model to assign a probability of legislation being introduced in the United States Congress to pass or fail. The input for this model will include the joint data set called “US Congress: Bulk Data on Bills” compiled by ProPublica and the Sunlight Foundation [11]. The data set includes a history of all bills and resolutions that were introduced to congress from 1974-2019 and is updated twice daily so that new information will be available to model in near real time. Additional information about each bill can be derived, such as the number of co-sponsors for each bill. Additional areas of interest may include an indicator to display if the bill sponsor’s party aligned with the current executive branch or with the speaker of the house.

There are many data points which can be gathered from ProPublica’s Data Repository [11]. The following key data fields from the primary data set will make up the initial predictors considered for our model.

Primary Legislation Metadata Fields

* Bill Name: name of legislation being discussed among the house and senate
* Action: key events in the legislative process for a bill such as introduction to Congress, referral to committee, debate, and votes.
* Action Date: The date that actions are taken on legislation.
* Sponsor: The Congressperson who submitted the bill for consideration. A bill has exactly one sponsor.
* Sponsor Affiliation: the stated political affiliation of a sponsor as a member of the Democratic party, the Republican party, or as an independent.
* Co-Sponsor: Congresspeople who support the bill and agree to lend their support in a documented role. A bill may have zero, one, or many cosponsors.
* Co-Sponsor Affiliation: the stated political affiliation of a co-sponsor as a member of the Democratic party, the Republican party, or as an independent.
* Bill Tag: the subject tag associated with a bill such as “military” or “health care.” A bill may have one or many subject tags.
* Originating House: a categorical variable that identifies a bill as originating in either the Senate or House of Representatives.

With the above details provided for every bill starting from 1974 to present, we believe patterns and trends can be discovered, leading to statistically significant hypotheses which can provide predictions on the outcome of legislation being discussed in the house and senate.

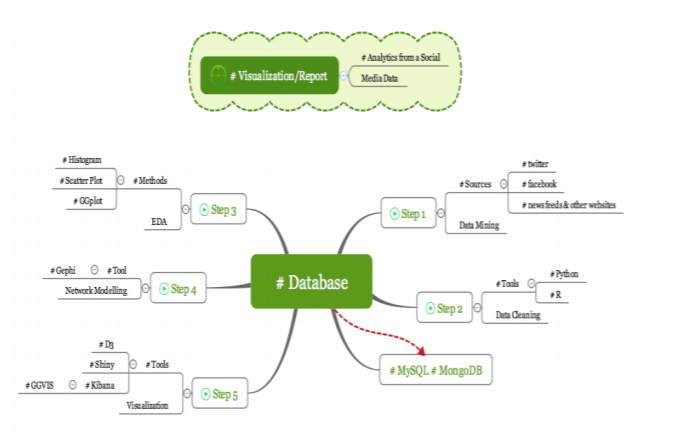
The types of legislations in the Bulk Data on Bills data sets include: House Concurrent Resolution (hconres), House Joint Resolution (hjres), Bills starting from the House of Representatives (hr), House Simple Resolution (hres), Bills starting from the Senate (s), Senate Concurrent Resolution (sconres), Senate Joint Resolutions (sjres), and Senate Simple Resolution (sres). See Appendix A for additional details about each type of legislation.

In addition to the fields available in the primary data set, the team supplemented the legislative data with additional historic information. This information was used in conjunction with the primary data set to derive indicators.

Additional data fields include:

* Executive Party: the political affiliation of the executive branch at the time a bill is proposed
* Majority: The majority party of the originating house
* Membership Change: The net change of membership in the originating house. In election years, each party may have a net gain, loss, or stagnation with respect to the number of like-party representatives they have in Congress. In non-elective years, the net change of party distribution will almost always be zero, barring any off-cycle elections due to resignation or other end of term activities.

1. **System Design**



***Figure 1: System Architecture Visual Map***

* *Data Cleaning:*

Cleaning the data from the ProPublica Data Store was the most challenging part and that’s where we spent 70% of the project time. We cleaned the unstructured data collected from the ProPublica in the form of JSON files or API calls. Since a majority of the data was in JSON format there was no guarantee certain attributes would be consistent with each file. Certain identifiers and attributes changed with different years, so our Data Cleaning process had to be optimized to account for them.

* *Network Modelling:*

We created three custom made network models in order to figuring out the various relationships and patterns between legislators. These allowed us to dive deeper into our goal to analyze various relationships/patterns for legislators using public data on ProPublica. The network model contained nodes (legislators) and edges (cosponsor ship) helped us detect to the patterns.

* *Programming Languages*

Programming was done in R, Python, and MySQL. Python was primarily used for acquiring the ProPublica data for network and predictive analysis. Python libraries which made connections to MySql databases were used to transfer the ProPublica data directly into our local database. This was done using both Rest API calls for more recent data or JSON extraction libraries for archived data. MySql was used for maintaining all the ProPublica data in a database and creating edges and nodes for network visualization.

* *Software Tools:*

PyCharm IDE was used for Python development. MySQL Workbench was used for MySQL development and data transfer processes to other applications such as RStudio and Gephi. RStudio was used for Programming R and MySQL database. We also used Microsoft Power BI for visualizing the data result sets.

* *Visualization*

Gephi was used for the final presentation of the data. The ease of use, multidimensionality of the visualizations, interactive nature of the platform, and quality of graphics produced provided the best presentation. Visualizations such as maps, histograms, clusters, word clouds, heat maps, and others were used to represent result data sets.

1. **Data Sources**
2. *ProPublica Data Store*

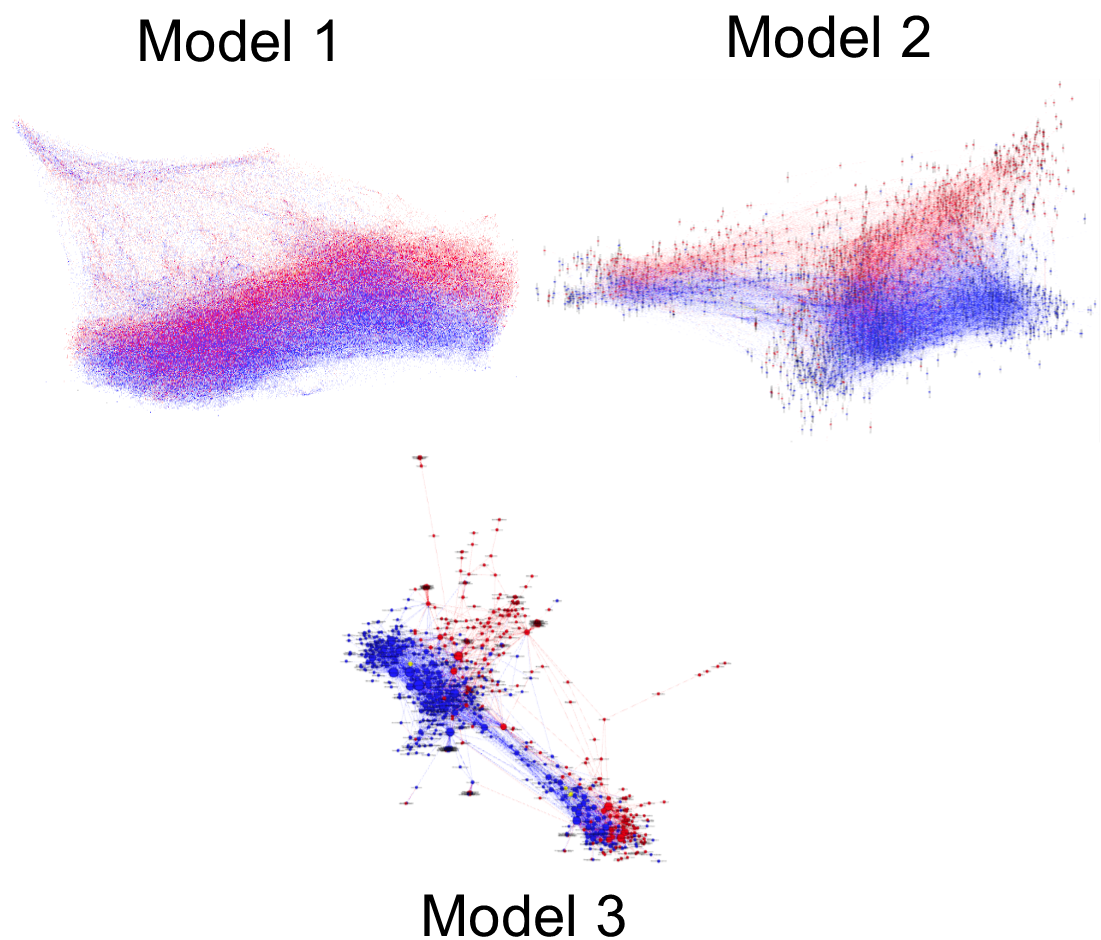
ProPublica Data Store was the primary source for the data. Data was gathered from the ProPublica using ProPublica Data Store API via Python in the form of JSON objects. Data gathered contained various key attributes that helped us answer the key questions mentioned earlier and it helped us to do the network analysis and various other reports. The data gathered was stored in the .csv files and it was filtered out accordingly. After that the data was moved to MySQL database and from there we used the structured data for the visualization.

1. *Congressional Websites*

Reputable websites pertaining to congressional processes played a pivotal role in gathering information which helped explain results we found from our datasets.

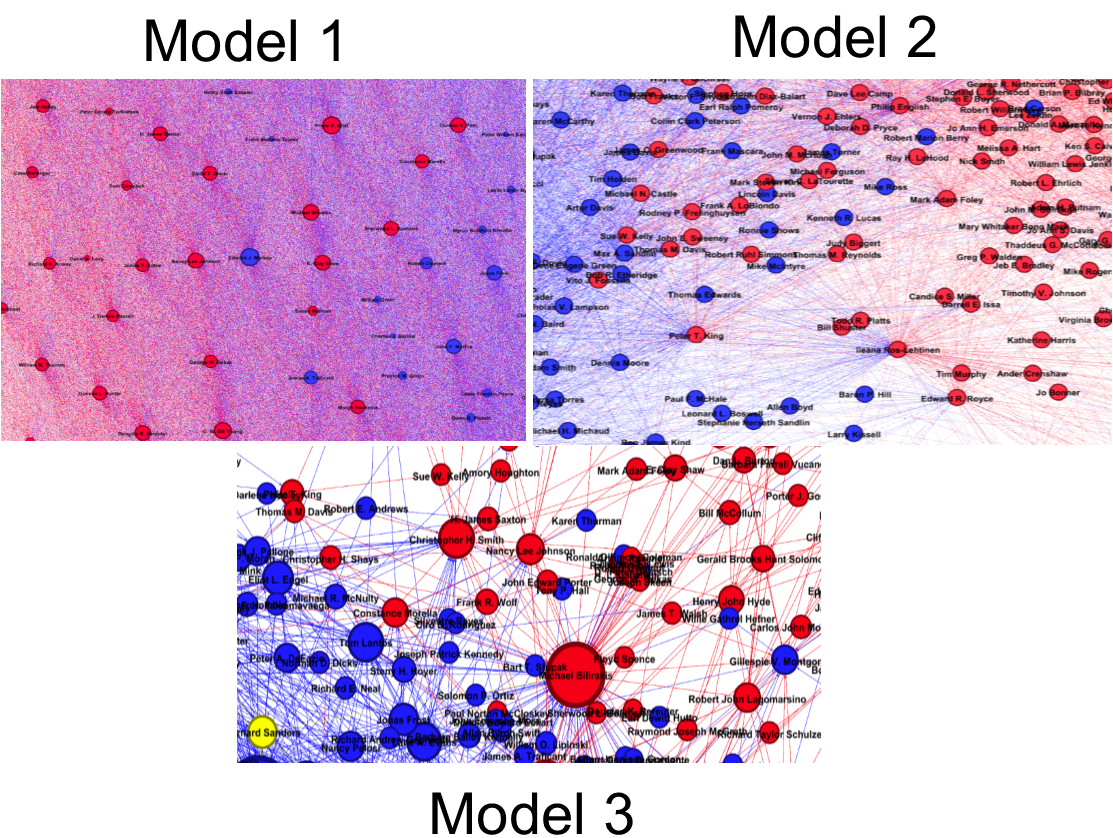
1. **Experiment Results**

With many of the key questions in this project being built around analyzing an established network which represents legislators supporting(cosponsoring) other legislators’ work, significant detail was put in to ensure such a network was possible. The first step in our experimental process was creating our three network models as described in the proposed approach.



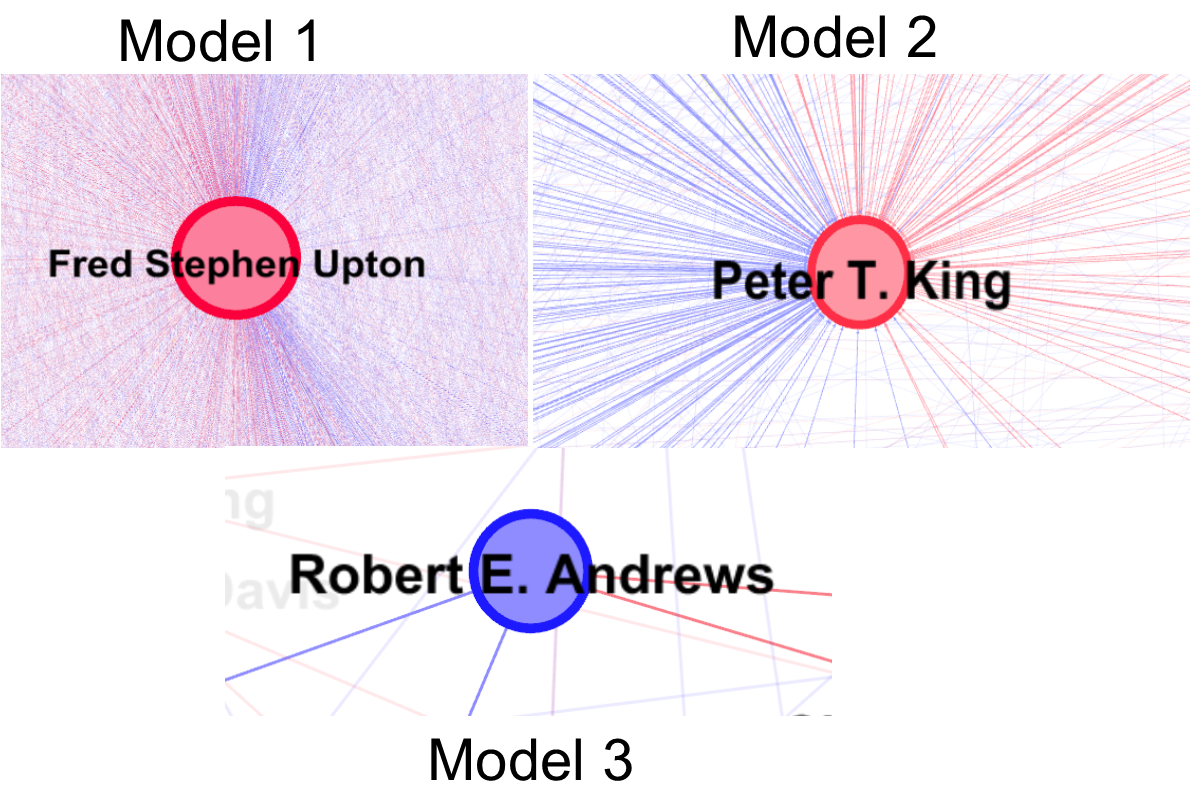
***Figure 2: Network Models***

Based on *figure 2*, the difference in relationship constraints placed on each model played a significant role in their overall number of edges per node.



***Figure 3: Network Models Close Up***

Viewing the models from different depth levels provides an interesting contrast to the amount of edges which can be seen in more detail in *figure 3 and figure 4*. The most reassuring factor however was the fact that similarities did exist amongst all three models, which were visible clusters and communities of legislators that were partitioned by political affiliations.



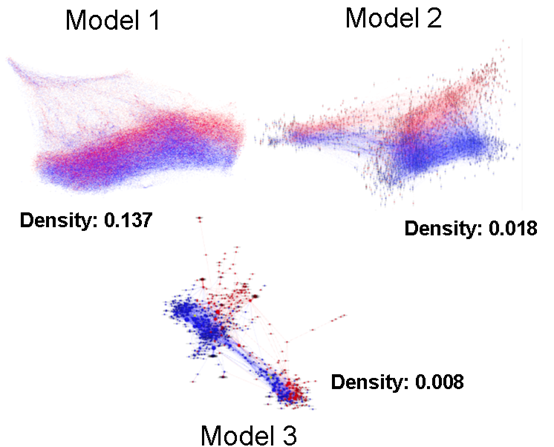
***Figure 4: Network Models Unit Level***

* *Network Density*



***Figure 5: Network Density Equation***

When evaluating the network models, density had some of the more expected yet still noteworthy outcomes. Model 1 contained the densest of the network models with a density of **0.137**. This make sense considering the formula shown in *figure 5* which divides actual connections by potential connections. Model 1 has the most connections of any of the models even when accounting for the potential orphaned nodes which are created due to the network models’ relationship constraints. Model 2 came roughly 7 times less dense with a density of **0.018** while model 3 was roughly 2 times less dense than model 2 with a density of **0.008**.

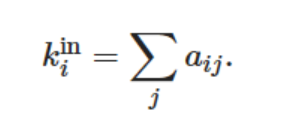


***Figure 6: Network Density of Models***

* *Degree Centrality*

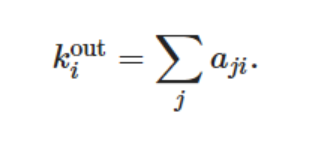
Degree centrality of a legislator refers to the number of cosponsors or times the legislator co-sponsored with another legislator. Degree centrality was one of the simplest centrality measures to compute and takes less computational time to compute in comparison to many of the other formulas to calculate. For degree centrality, higher values mean that the legislator is cosponsored more heavily than others or is cosponsoring other legislators frequently. Degree centrality essentially shows how many connections a legislator has and therefore was highest for model 1 with an average degree of **306.347**. Model 2 had an average degree of **31.802** and model 3 had an average of **5.638**.

*In Degree Formula:*



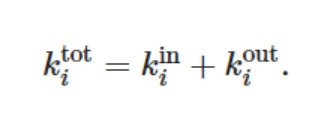
With our models, one cosponsor to another legislator counts towards the in degree of the legislator.

*Out Degree Formula:*

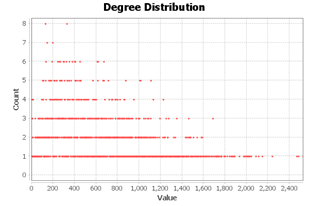


With our models, one cosponsor to another legislator counts towards the out degree of the co-sponsor

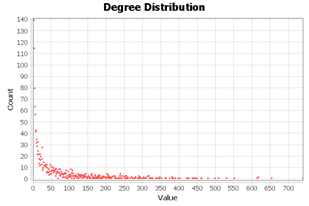
*Degree Centrality Formula:*



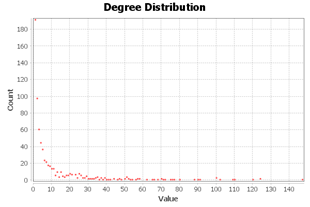
The summation of the in-degree and out-degree of a legislator equates to the degree of centrality for the legislator.



***Figure 7: Model 1 - Average Degree 306.347***

****

***Figure 8: Model 2 - Average Degree 31.802***

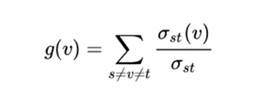
****

***Figure 9: Model 3 - Average Degree 5.638***

* *Betweenness Centrality*

Betweenness centrality is a way of detecting the amount of influence a legislator has over the flow of information in a graph. It can be used to find legislators that serve as a bridge from one cluster to another. Betweenness centrality measures the extent to which one individual lies on paths between other legislators. Legislators with high betweenness may have considerable influence within a network as many relationships are existent only by passing through themselves.

*Betweenness Centrality Formula:*

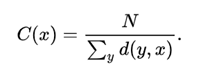
****

* *Closeness Centrality*

Closeness centrality (or closeness) of a legislator is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the legislator and all other legislators in the graph. Thus, the more central a legislator is, the closer it is to all other legislators. Closeness centrality is a way of detecting legislators that are placed in a very efficiently location on a graph.

The closeness centrality of a legislator measures its average farness (inverse distance) to all other legislators. Legislators with a high closeness score have the shortest distances to all other nodes.

*Closeness Centrality Formula:*



With the average path of Model 3 being higher than the other Models (**4.1983**), the average closeness and betweenness centrality of legislators in this graph will be higher than the other models. Model 2 would follow second with an average path of **3.0650** and then model 3 with an average path of **2.0344**.

**Model 1**

Diameter = **5**

Average Path Length = **2.0344**

**Model 2**

Diameter = **9**

Average Path Length = **3.0650**

**Model 3**

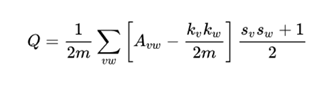
Diameter = **11**

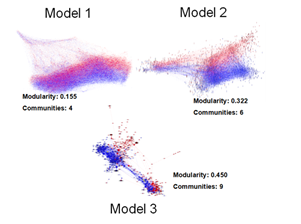
Average Path Length = **4.1983**

* *Modularity*

The modularity of the three network models under experimentation will provide very useful insights in terms of which model represents groups (clusters or communities) best. This is because modularity is designed to measure the strength of division in a network. For a network to have high modularity there must be dense connections between legislators within a specific group but sparse connections with legislators in different modules. The Model 1 network was the densest of the three networks by a significant factor, however the density of Model 1 is not enough alone to make its modularity higher than the other models. This is because Model 1 lacked enough sparse connections with legislators in different modules. Hence why Model 1 only has a modularity rating of **0.155** while Model 2 (**0.322**) and Model 3 (**0.450**) excel at this calculation.

*Modularity Formula:*



***Figure 9: Modularity of Models***

* *Predictive Modeling*

With predictive analytics on ProPublica data not being a commonly ventured practice, very little was preparable in our planning stages as to what indicators we would use. We therefore decided to wait until our data extraction and network analysis processes were completed as we figured our knowledge of the data itself would be stronger. This would equate to more practical and reasonable indicator being selected.

Based on our observations working with the data for some time, the following predictive models were established and have the following prediction indicators.

*Actions & Partisan Model:*

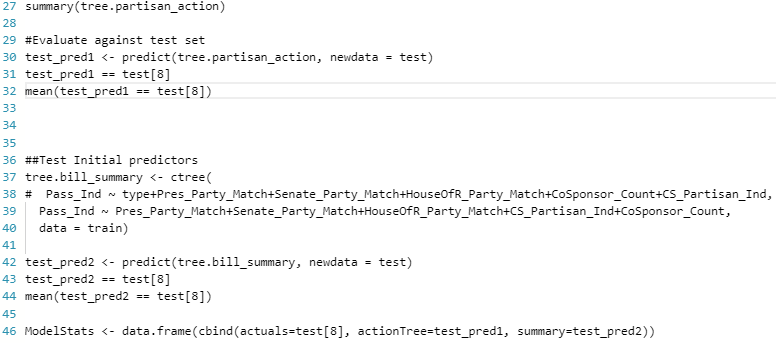
* the cosponsor partisan indication
* the action count

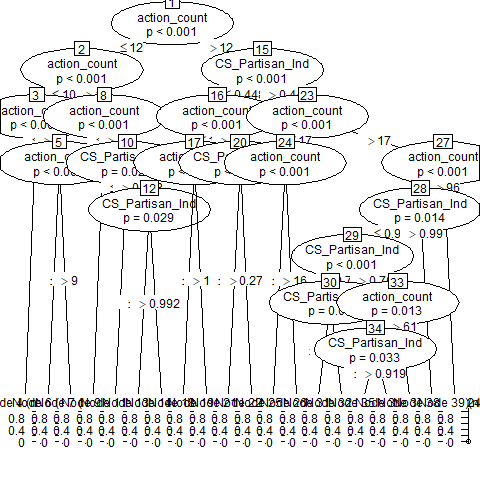
*Bill Summary Results:*

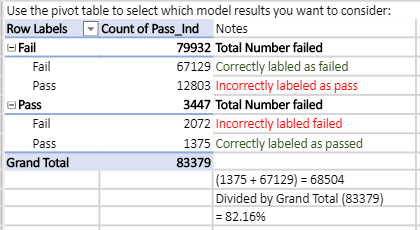
* if the sponsor’s political party matches the president
* if the majority of the senate matches the cosponsors party
* if the majority of the House of Representatives matches the cosponsors party
* the cosponsor partisan indicator
* the count of cosponsors

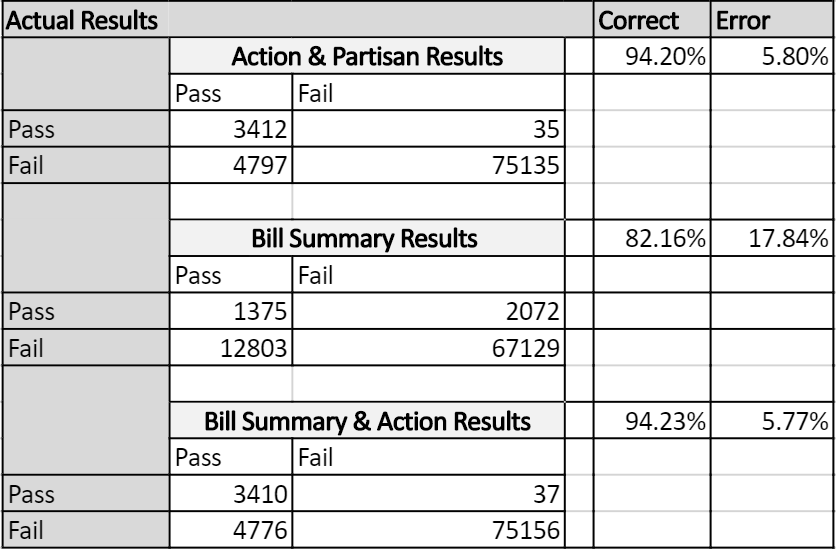
*Bill Summary & Action Results:* (Prediction based on)

* if the sponsor’s political party matches the president
* if the majority of the senate matches the cosponsors party
* if the majority of the House of Representatives matches the cosponsors party
* the cosponsor partisan indicator
* the count of cosponsors
* the action count

***Figure 10: Sample Decision Tree Logic written in R***

***Figure 11: Decision Tree used in Predictive Model***

***Figure 12: Example calculating the accuracy of the Bill Summary Results predictive model***

***Figure 13: Results of the 3 most significant predictive models we discovered***

Based on these results we found that our models’ accuracy is increased significantly if action counts are a factor in the model.

1. **Discussions**

With all of our experiments finished we found that the key questions we aimed to answer at the start of our research had more dependencies and similarities to each other than we had originally expected.

*1. Can measurable relationships between legislators in United States congress be observed based on co-sponsor support of legislation proposed?*

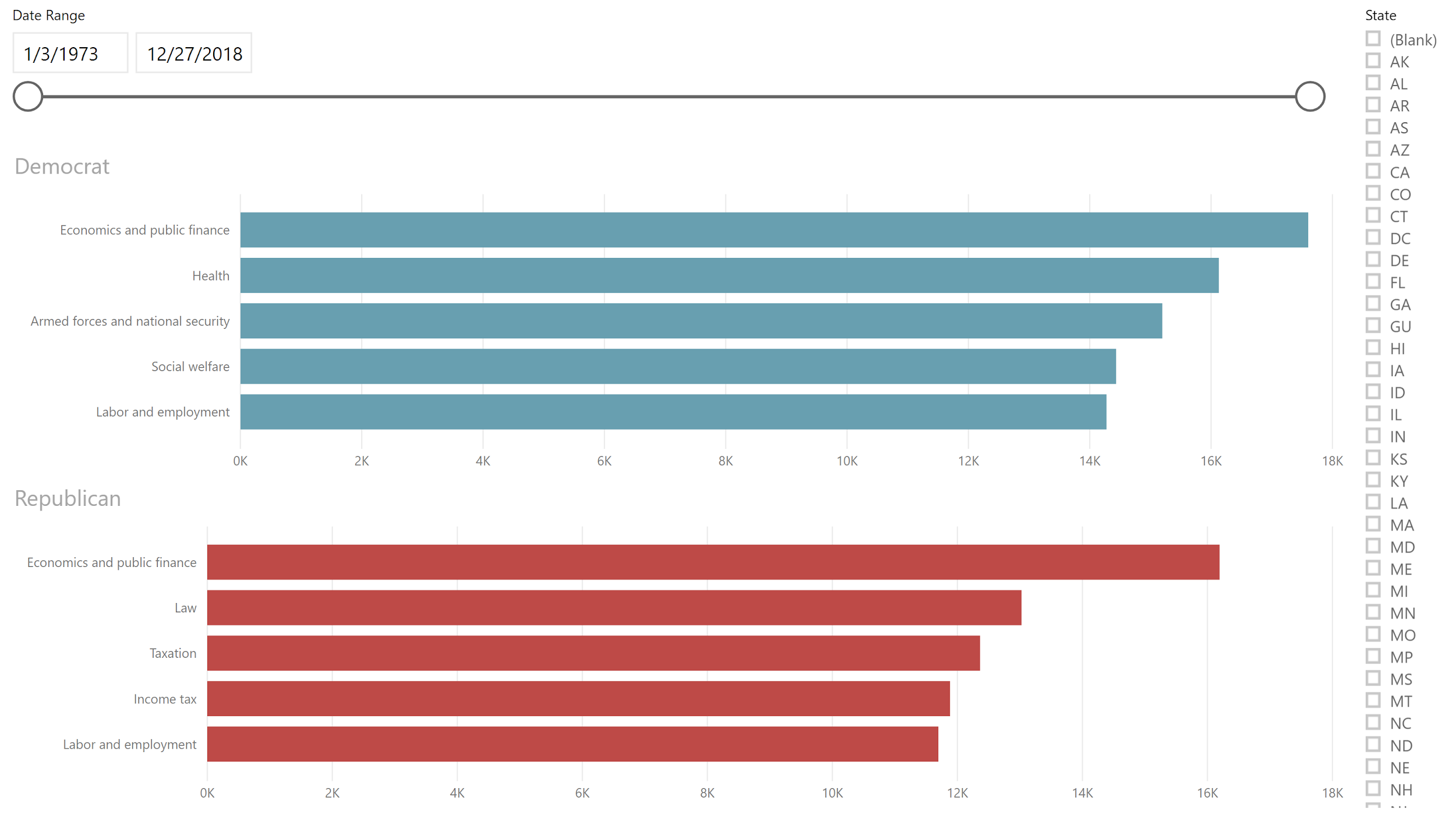
Even though we created three separate network models in the hopes that at least one of them would represent a network of political allegiances, we ended up proving that all three of our designed networks represented a similar structure of the bipartisan government system the United States is established on.

*2. What type of communities/cluster are existent based on the co-sponsor support of legislation?*

When looking at our network analysis experimental results it is clear that the number and quality of communities/clusters differ from model to model. Specifically, Model 1 had the lowest modularity which in terms equates to Model 1 having the lowest number of clusters in comparison to the other models. One key observation that could be observed amongst all networks however was a minimum of four communities which ended up being Senate Democrats, Senate Republicans, House of Representatives Democrats, and House of Representatives Republicans. Moving from Model 1 to Model 3 greatly increased the modularity which allowed for these distinguished communities to become more apparent, along with many other smaller communities which arose.

*3. How does proposed legislation differ from the communities discovered?*

With the extensive network analysis and strategic structured queries which we were able to incorporate into Microsoft Power BI, we were able to create a dynamic report which shows the top five subjects of legislation proposed by legislators based on political party, state, and date. Averaging from 1973 to 2018, Democrats proposed legislation centered around Economics, Health, Armed Forces/National Security, Social Welfare, Labor/Employment. During the same time period Republicans proposed legislation centered around Economics, Law, Taxation, Income Tax, Labor and Employment.

***Figure 14: Proposed Legislation Subjects by Sponsor***

*4. Can a predictive model be created using various indicators which can predict the outcome of legislation currently being proposed in congress?*

With the use of R Studio and the Decision Tree package which come with it, we were able to create three predictive models which had a higher than 80 percent success rate at least. Though we did not expect to make three predictive models, it's important to note that the predictive models which are built around the action date are highly dependent on a bill being predicted on having current actions on it. This means that these models are not effective until the bill has been on the floor of congress for a while. The good thing is that the model will re-compute its probability of success with each action, making it a living model.

1. **Conclusions**

Each day, new Legislative data is made available by US Congressional Organizations constantly and patterns are bound to be present for many of the information saved daily. Just like other applications in our society where big data is utilized, the more data that is turned into information equates to the higher quality of knowledge and further insight we can make in almost any topic area. We have experimented this ideology with both network and predictive analytics throughout the extent of this research. What we found was that without artificial identifies which link legislators to others, relationships are present using organic identifiers such as cosponsor ship on certain legislation. This essentially revealed various communities which share similar interests which can be observed when looking at the network as a whole. Taking things, a step further, predictive analysis can then be made on the actions made in these communities based on historical patterns and recurring trends. With the right design and rational, connections can be made across many mediums, and the United States Congressional system is not an exception to this phenomenon.

**References**

[1] Adler, S. and Wilkerson, J. (2005). The scope and urgency of legislation: Reconsidering bill success in the house of representatives. [online] Washington D.C.: American Political Science Association. Available at: <https://faculty.washington.edu/jwilker/353/adler.pdf> [Accessed 12 Apr. 2019].

[2] Clinton, J., Jackman, S. and Rivers, D. (2004). The Statistical Analysis of Roll Call Data. American Political Science Review, 98(2), pp.355-370.

[3] Diermeier, D., Godbout, J., Yu, B. and Kaufmann, S. (2012). Language and Ideology in Congress. British Journal of Political Science, [online] 42(1), pp.31-55. Available at: <https://www.jstor.org.mutex.gmu.edu/stable/41485863> [Accessed 12 Apr. 2019].

[4] Gerrish, S. and Blei, D. (2011). Predicting Legislative Roll Calls from Text. Princeton: Princeton University. Available at: <https://www.cs.columbia.edu/~blei/papers/GerrishBlei2011.pdf> [Accessed 12 Apr. 2019].

[5] J. Mark Phillips, The Infinite Legal Acumen of an Artificial Mind: How Machine Learning Can Permanently Capture Legal Expertise and Optimize the Law Firm Pyramid, 11 J. Bus. Entrepreneurship & L. 301 (2018) Available at: <https://digitalcommons.pepperdine.edu/jbel/vol11/iss2/3>

[6] Kraft, P., Jain, H. and Rush, A. (2016). An embedding model for predicting roll-call votes. In: Conference on Empirical Methods in Natural Language Processing.

[7] Laver, M., Benoit, K. and Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. American Political Science Review, [online] 97(02). Available at: <http://faculty.washington.edu/jwilker/559/Laver.pdf> [Accessed 12 Apr. 2019].

[8] Morning Edition, (2019). [Radio programme] WAMU: National Public Radio.

[9] Ourcampaigns.com. (2019). Our Campaigns Home Page. [online] Available at: <https://www.ourcampaigns.com/> [Accessed 12 Apr. 2019].

[10] Pramuk, J. (2019). Trump signs bill to temporarily reopen government after longest shutdown in history. CNBC. [online] Available at: <https://www.cnbc.com/2019/01/25/senate-votes-to-reopen-governmentand-end-shutdown-without-border-wall.html> [Accessed 12 Apr. 2019]

[11] ProPublic and Sunlight Foundation (2019). US Congress: Bulk Data on Bills. ProPublic Data Stores. [online] Available at: <https://www.propublica.org/datastore/dataset/congressional-data-bulklegislation-bills> [Accessed 12 Apr. 2019].

[12] Roberts, J. (2007). The Statistical Analysis of Roll-Call Data: A Cautionary Tale. Legislative Studies Quarterly, [online] 32(3), pp.341-360. Available at: <http://www.jstor.org.mutex.gmu.edu/stable/40263426> [Accessed 12 Apr. 2019].

[13] Vargo, C. and Hopp, T. (2016). Socioeconomic Status, Social Capital, and Partisan Polarity as Predictors of Political Incivility on Twitter. Social Science Computer Review, [online] 35(1), pp.10-32. Available at: <https://doi.org/10.1177/0894439315602858> [Accessed 12 Apr. 2019].

[14] The Washington Post (2018). Can big data predict which bills will pass Congress?. [online] Available at: <https://www.washingtonpost.com/lifestyle/magazine/can-big-data-predict-which-bills-will-passcongress/2018/01/31/ffe6c162-f7c3-11e7-b34a-b85626af34ef_story.html?utm_term=.a7b58074d133> [Accessed 12 Apr. 2019].

[15] Yano, T., Smith, N. and Wilkerson, J. (2012). Textual Predictors of Bill Survival in Congressional Committees. [ebook] Carnegie Mellon University School of Computer Science. Available at: <http://www.cs.cmu.edu/~nasmith/papers/yano+smith+wilkerson.naacl12.pdf> [Accessed 12 Apr. 2019]. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

**APPENDIX**

1. *Assignments*

- Spencer Woodfork: Team Leader, Developer

* + Dilpreet Singh: Developer/Analyst
  + Meena Rapaka: Developer/Analyst

1. *Tasks*

Data Crawling – Dilpreet Singh  
Data Preprocessing – Dilpreet Singh  
Databases - Meena Rapaka  
Network Analysis - Spencer Woodfork  
Predictive Modelling - Spencer Woodfork  
Visualization – Meena Rapaka

1. *Schedule*

As this project is to be done in a fixed period of time, certain deadlines are vital to meet:

Weekly team meetings with previous, current and next week deadlines.

|  |  |
| --- | --- |
| I). | *February 25th, 2019*: Project Proposal |
| II). | *April 1st, 2019*: Project Checkpoint |
| III). | *May 6th,* 2019: Project Presentation |

