1. **Introduction**

Few individuals have more of an impact over the direction of our country than those elected to serve in Congress.

1. **Data Sources**

We primarily used data from the “Cosponsorship Network Data” dataset, collected by James H. Fowler, Andrew Scott Waugh, and Yunkyu Sohn. From this dataset we were able to get the cosponsorship matrix for each house of Congress in each sessions. From this matrix we were able to create the adjacency matrix, connecting the legislators. It is important here to note the characteristics of the data. Because of the directional nature of cosponsorship and the number of bills proposed in each session of congress, it is essential to represent the congressional relationships as a weighted, directional graph. If these key features are overlooked, an over-connected clique-like graph emerges, on which no meaningful analysis can be done.

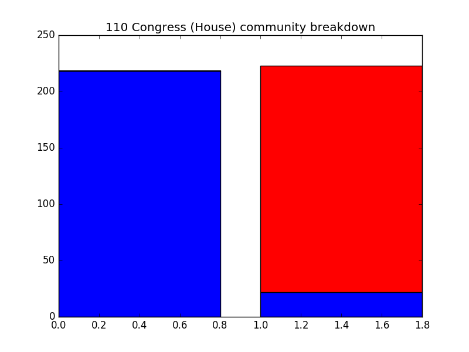
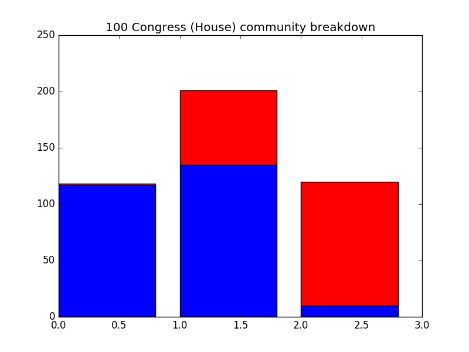
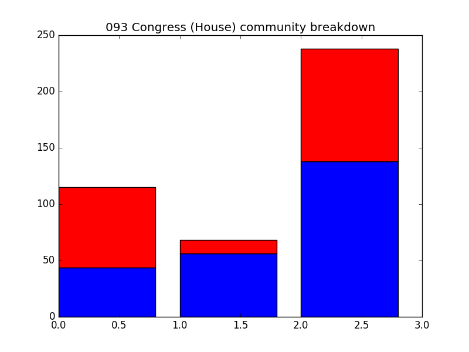
The graph data was also put into iGraph graphs, and then serialized using Python’s pickle protocol, and the adjacency matrices saved in the hdf5 format. This allows for a near 400x speedup on data loading, avoiding the csv and text file processing associated with the original data.

We also utilized the ICPSR numbers associated with the pair of (congress person, congress number) to identify the party of each congressperson, and provide a rough value of their seniority. These values were collected both through the Fowler dataset and the work of the VoteView reference website.

1. **Community Detection**

One of the primary goals of this project was to attempt to identify the groups within Congress and determine the political party breakdown within the groups. We used the walktrap algorithm, as provided by the iGraph python package to do this partitioning, and used the optimal modularity to determine the number of communities, to avoid artificially forcing the partitions to the two parties we have today. The optimal modularity comes from a maximizing the number of intra-community edges, when summed over all communities. Figure 1 has the community breakdown for the 93rd, 100th, and 110th House.

Figure 1: Left to right, 93rd, 100th, 110th House



Even in this small subset of the data, we can see that in 110th House there is a closer alignment of political party with graph community. This corresponds to the session of Congress beginning in 2007, which aligns with our intuition. Unfortunately, the data from the 111th and 112th Congress is not yet part of the dataset, which would likely show even an even tighter alignment along party lines.

We can also see that in the 110th congress there are fewer subgroups when compared to other years. With the 100th Congress for example, we can see that there are two distinctly partisan groups, but there is also a distinct bipartisan group in the middle. This is not seen in the 110th. An interesting question for further work would be to look at data for the 112th Congress, after the formation of the Tea Party, to see if there is a visible split from the mainstream Republican Party in the cosponsorship data.

1. **Community Visualization**

The community detection by itself is quite interesting, but another goal of the project was to visualize these communities in a manner that communicated the changes in community composition. However, one of the main issues with visualizing this graph data was the high degree of connectedness, even among members of different communities. (This can be explained when considering generally non-partisan bills, such as defense authorizations, that involve a large majority of Congress.) Because of this, even when considering the weighted graph, there was not a good way to spread the communities out visually. After trying a number of difference graph visualization algorithms provided by Gephi, we came concluded to instead color the communities by party breakdown.

When outputting the graph data in the .gexf format, we specify the color of each community in RGB values, according to the following formula.  
   
We then loaded the files into Gephi, and used the Fruchterman–Reingold Force Directed graph algorithm to position the nodes. The nodes were then sized according to their relative weighted in-degree, which we used as a rough heuristic for importance. There will be more on importance later.

In figure 2, you can see the visualized graphs for the 93rd, 100th, and 110th House, to compare with the bar graph breakdown.

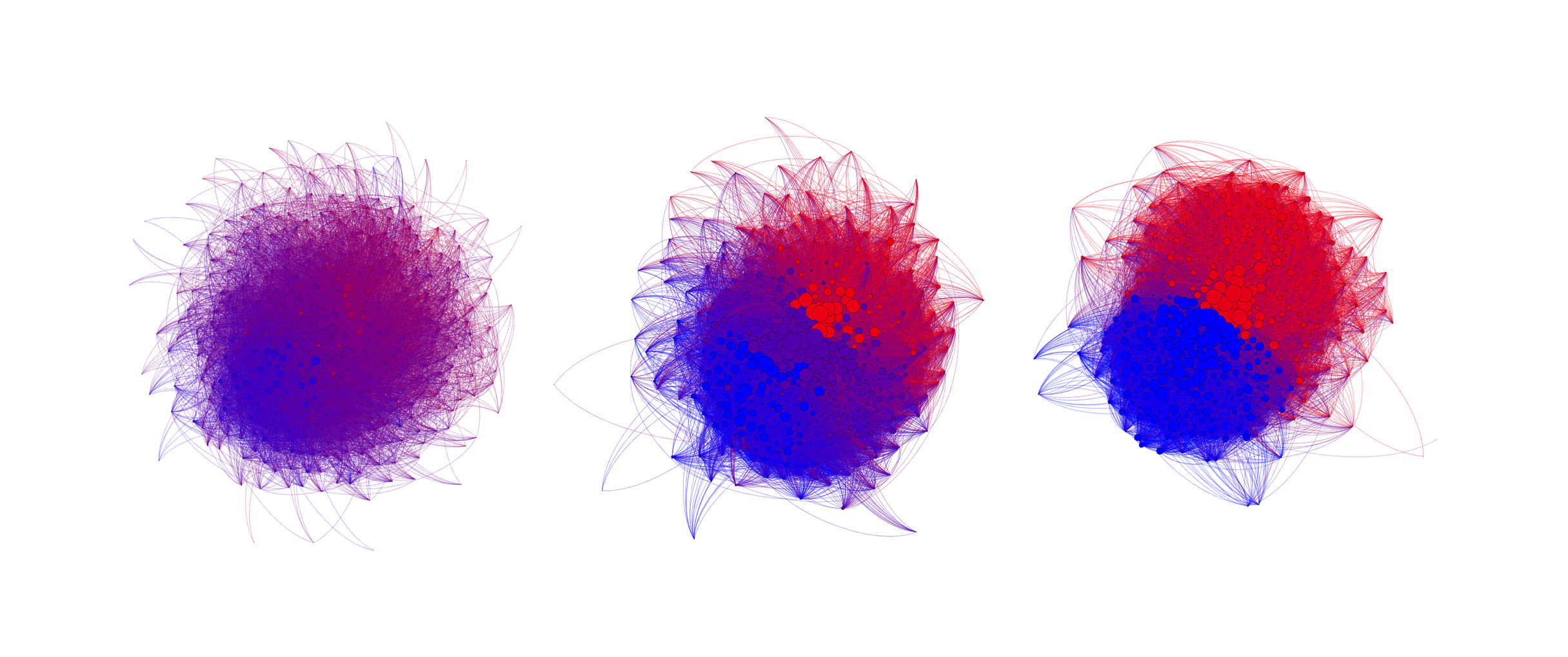


Figure 2: Left to right, 93rd, 100th, and 110th House

There are a number of notable things from this picture. The most immediately obvious conclusion is that just based on color, we can see that the “purple-ness” of the House has gone down, mirroring our results from the community detection. We can also see the bipartisan group present in the 100th House is visible in purple, and is also located between the “red” and “blue” groups. This was not planned, and is just a consequence of the graph structure and layout drawing, but makes intuitive sense. The size of the nodes has also increased since the 93rd House – this is actually due to a rule change concerning the number of cosponsors a bill could have – the restrictions were lifted in the 95th House. We also see our primitive measure of importance, the weighted in-degree, has roughly corresponded with the center of the visualized graphs, with the smaller nodes on the outside, or roughly less important.

1. **Degree Analysis**

Another goal of the project was to identify the key characteristics of the social network structure. We looked at the graph diameter, average clustering coefficient, and the degree distribution.

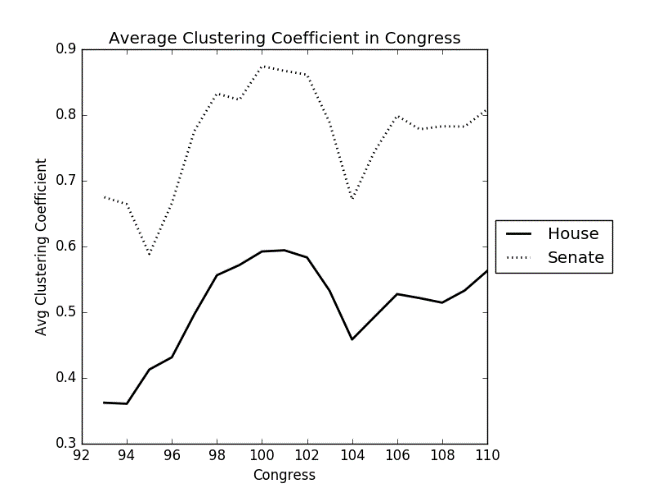
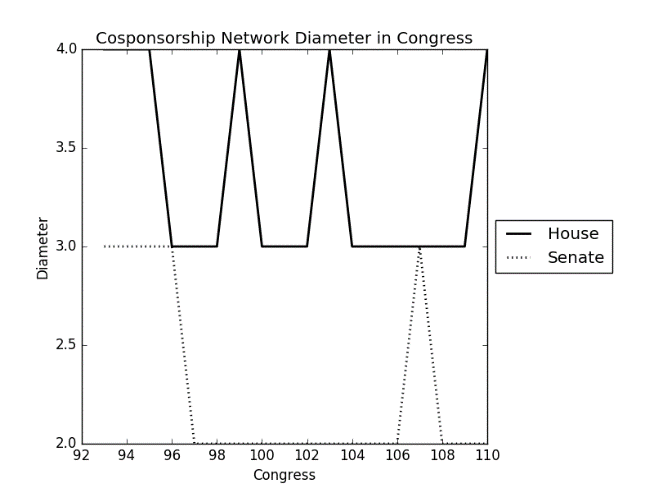
Figure 3 shows the graph diameter for both the House and Senate, and figure 4 details the clustering coefficient.  


Figure 3: Network Diameter. Figure 4: Avg Clustering Coefficient.  
  
We can clearly see from both graphs that Congress is a highly connected network. The Senate, with the smaller number of members and longer terms has an understandably smaller diameter and higher clustering coefficient. Both chambers are also clearly small world graphs, with higher clustering coefficients than a random graph and a small diameter.

The next logical question to ask is if Congress forms a power-law network. Because of how tightly knit the Senate is, we didn’t expect to find a power-law network there, but with the House elections happening every two year, we thought preferential attachment (to the influential members) might occur. To determine if this is true, we plotted log(degree count) vs log(number of nodes with that degree), and found the line of best fit. Figure 5 has an example of the degree distribution and corresponding log-log plot, and table 1 has the results for all years in our dataset.

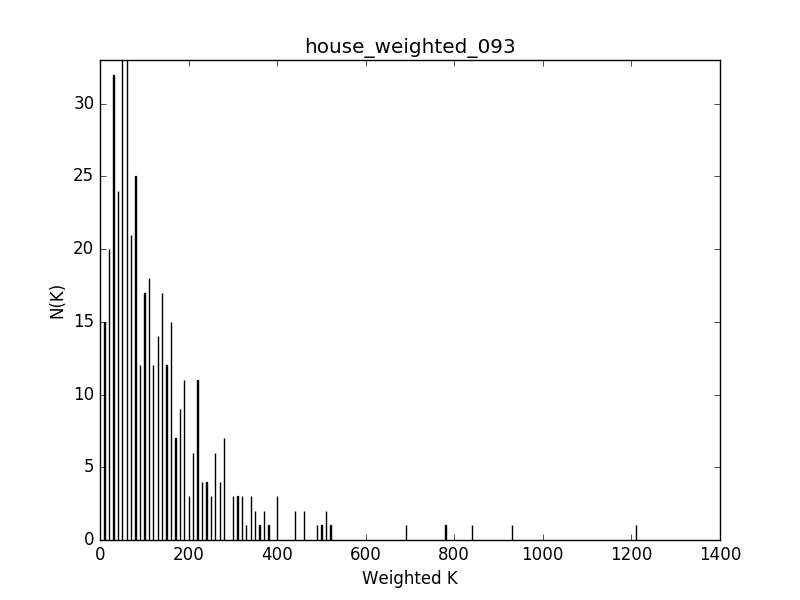
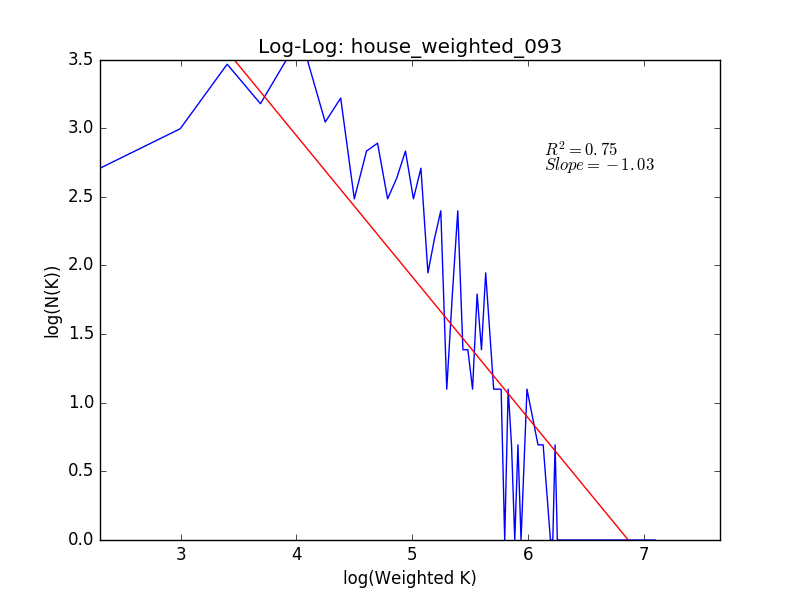
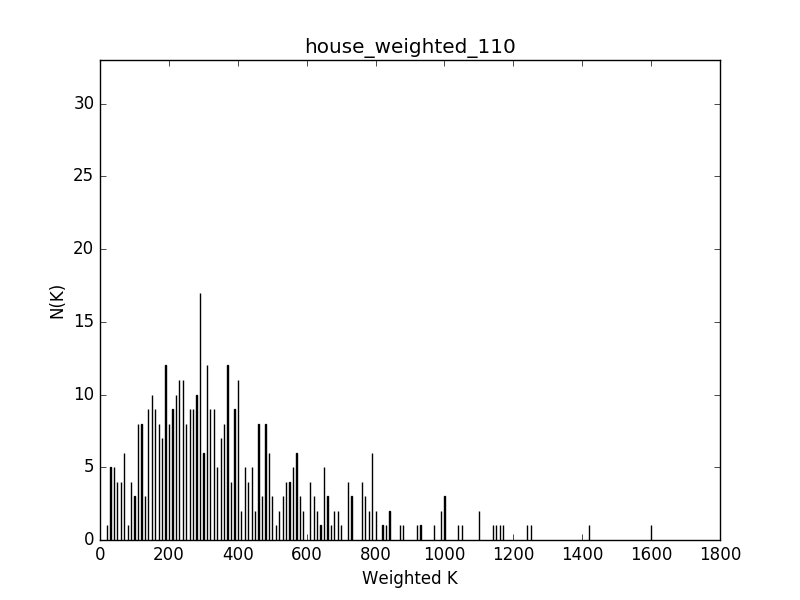


Figure 5: 93rd House degree distribution and log-log plot

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| : | | | | | | | |
| Chamber | Session | Slope | R^2 | Chamber | Session | Slope | R^2 |
| Senate | 93 | -0.31 | 0.13 | House | **93** | **-1.03** | **0.75** |
| Senate | 94 | -0.51 | 0.31 | House | **94** | **-0.99** | **0.71** |
| Senate | 95 | -0.30 | 0.15 | House | **95** | **-0.96** | **0.66** |
| Senate | 96 | -0.11 | 0.02 | House | 96 | -0.76 | 0.45 |
| Senate | 97 | 0.08 | 0.01 | House | 97 | -0.68 | 0.46 |
| Senate | 98 | 0.00 | 0.00 | House | 98 | -0.64 | 0.46 |
| Senate | 99 | -0.01 | 0.00 | House | 99 | -0.61 | 0.41 |
| Senate | 100 | -0.02 | 0.00 | House | 100 | -0.48 | 0.30 |
| Senate | 101 | -0.03 | 0.00 | House | 101 | -0.48 | 0.31 |
| Senate | 102 | 0.02 | 0.00 | House | 102 | -0.46 | 0.23 |
| Senate | 103 | -0.15 | 0.02 | House | 103 | -0.59 | 0.32 |
| Senate | 104 | -0.06 | 0.00 | House | 104 | -0.75 | 0.40 |
| Senate | 105 | -0.27 | 0.05 | House | 105 | -0.71 | 0.42 |
| Senate | 106 | 0.03 | 0.00 | House | 106 | -0.52 | 0.24 |
| Senate | 107 | -0.23 | 0.08 | House | 107 | -0.55 | 0.29 |
| Senate | 108 | -0.26 | 0.07 | House | 108 | -0.64 | 0.35 |
| Senate | 109 | -0.33 | 0.11 | House | 109 | -0.40 | 0.18 |
| Senate | 110 | -0.08 | 0.01 | House | 110 | -0.47 | 0.24 |

Table 1: Best fit lines for degree distribution log-log plots

In this case, the weighted degree distribution clearly seems to follow a power-law distribution. We can see a negative slope of -1, and an R^2 error value of .75, which is tolerable. However, while this gave us initially promising results, the rules change in the 95th House seems to have changed this distribution. By removing limits on the number of cosponsors a bill could have, popular bills saw a large increase in cosponsor volume. This increased the overall number of cosponsorships, shifting the entire distribution seen on the right in figure 5 to the right.

Above we can see the 110th House degree distribution, and we can see that the distribution has become more spread out and shifted right.

To the right we can see Table 1, detailing the slope and R^2 value for each chamber and session of Congress. Only the 93rd, 94th, and 95th have R^2 values that would even hint at a linear relationship for the log-log plot (bolded). After distributions look more like the 110th, and the Senate observed a similar pattern.

From this, we can fairly confidently make the claim that Congress no longer follows a power-law distribution, even if the 93rd and 94th House did.

1. **Seniority Prediction**

The final major goal of this project was to attempt to predict the seniority of members of Congress using a graph metric such as their in-degree, centrality, or some other metric. As we saw earlier, using the weighted in-degree to scale nodes resulted in the “center” nodes being larger, with nodes on the outskirts of the graph being smaller. Unfortunately, after collecting a rough estimate of seniority for the 110th Congress (by noting who was in previous Congress’ in the dataset), there was not a significant relationship between weighted in/out degree and seniority in the House or Senate.

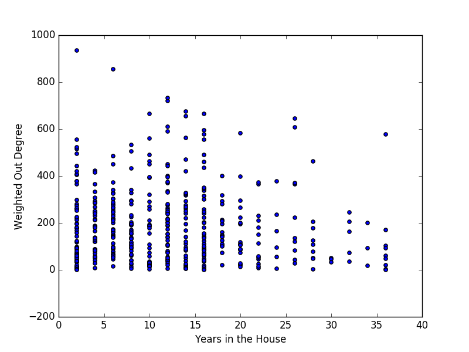
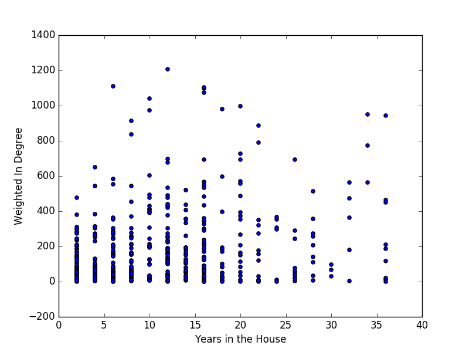
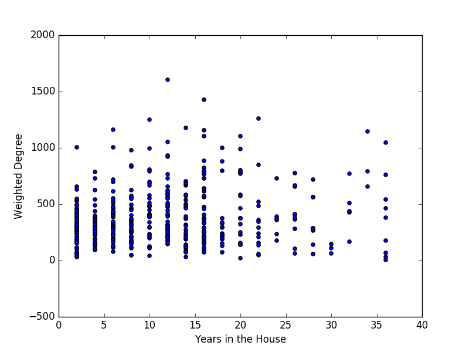


Figure 6: Seniority vs degree measures for the 110th House

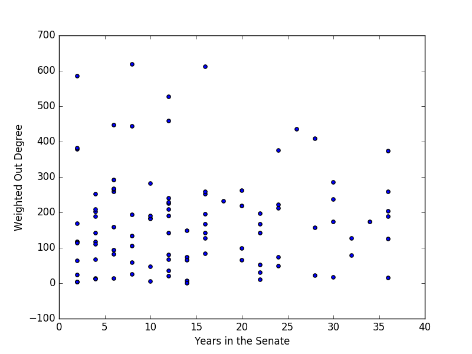
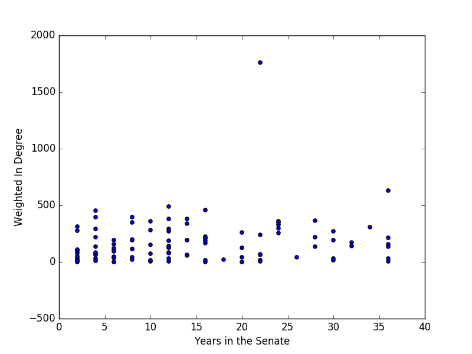
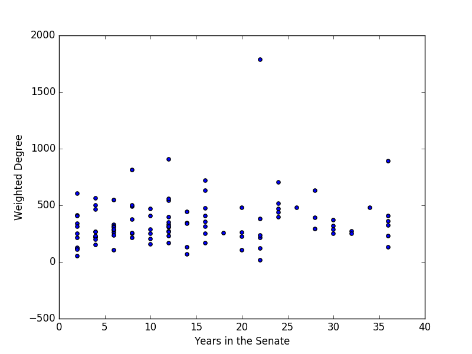


Figure 7: Seniority vs degree measures for the 110th Senate

Given the lack of clear relationship, we turned to more a sophisticated graph analysis technique, HITS, which attempts to find good authorities and hubs. A good authority is defined as knowing good hubs, while good hubs know good authorities. The iterative process of calculating these measure ends up being identical to the calculation of the primary eigenvector of the two forms of the covariance matrix of the graph adjacency matrix. After calculating those values, we again got disappointing results, as shown in figure 8. However, despite the setback, we still looked at the names of the authorities. This ended up being extremely positive. As detailed in table 2, both the authorities and hubs managed to identify some of the major influential Senators of 2007 – including future President, Barack Obama, as well as a number of familiar names.

Here it seems that HITS has done an extremely good job of identifying key players in the Senate, regardless of their seniority. This also makes sense in the context of the network – seniority is not the same as influence, and so it makes sense that there isn’t always a clear relationship with seniority. Unfortunately, the results from the House were not as strong. Key figures like the Speaker of the House were not identified through HITS, and while several committee chairs were, they weren’t the more important committees.

|  |  |
| --- | --- |
| Authorities | Hubs |
| Harry Reid | John F. Kerry |
| Edward M. Kennedy | Richard Durbin |
| Dianne Feinstein | Barbara Boxer |
| Richard Durbin | Hillary Rodham Clinton |
| Jr. Joseph R. Biden | Robert Menendez |
| Hillary Rodham Clinton | Sherrod Brown |
| Christopher J. Dodd | Charles E. Schumer |
| Jeff Bingaman | Barack Obama |
| John F. Kerry | Olympia J. Snowe |
| Russell D. Feingold | Joseph I. Lieberman |

Table 2

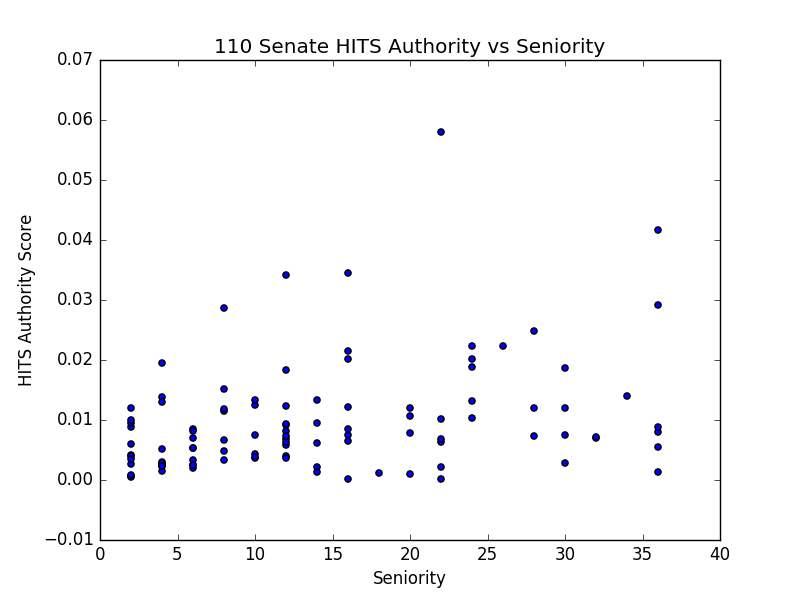


Figure 8: 110th Senate Authority vs Seniority

We hypothesize that this can be explained in part due to the turnover in the House, especially in the 110th House, when the Democrats took back control. When the 111th and 112th Congress data becomes available, it would be interesting to see if HITS is more relevant for those sessions.