Community Detection in Congressional Networks

Yan Zhang

Abstract

There has been considerable recent effort to study social networks in the context of graphs by analyzing real-world data. In this project, we investigate the community structure of social networks composed of Members of Congress. This entails finding subgraphs within Congressional networks that contain more connections than expected, indicating communities within which Congressmen like to collaborate. Analyzing the Congressional networks determined by legislation cosponsorship and committee memberships allow one to see significant political results, such as the increased polarization of Congress that has developed over the past 30 years.

1 Introduction

Graphs, defined as sets of nodes connected to each other by links, are integral structures in many disciplines, including mathematics, computer science, physics, biology, and the social sciences. In the graphs ("social networks") studied by social scientists, nodes commonly consist of individuals or groups of people, with links between them based on some specified relationship. For example, in Stanley Milgram's "small world" experiments, from which the phrase "Six Degrees of Separation" arose, each link connects a person who sent a letter and the person who received it [1].

An important example of a social network is the one formed by the members of the United States Congress. Because Members of Congress typically collaborate throughout the lawmaking process, finding the communities they form using methods without political bias is of great importance. By determining the composition of these cliques, one can find collaborative ties among Congressmen. This knowledge can help achieve a better understanding of the voting behavior of the Members of Congress, which is of great interest not only to the myriad interest groups and to political scientists, but also to the American public at large.

When constructing an abstract representation of the Congressional network, one must define nodes and create links between them. The nodes are the Congressmen themselves, but there are several ways to define the links, which should indicate levels of collaboration between Congressmen. One method of determining links arises from the fact that after legislation is proposed the work of framing

and phrasing the bills is done in committees. Therefore, one possible abstraction includes links between Congressmen if they served on the same committee or subcommittee. In addition, we can consider the committees as the nodes and add links between them if they share a Congressman [2], [3].

One can also measure the level of collaboration using the legislation itself. A piece of legislation has only a single sponsor but potentially many cosponsors. This allows one to connect Congressmen by adding to the link strength between two Congressmen if they both have served as a sponsor or cosponsor of the same bill; this ultimately results in a weighted network. Such a network was analyzed by Fowler to determine the "most central" Congressmen [4], [5], and legislation cosponsorship will be the primary subject of this study as well. Our objective is to find the community structure of the legislation cosponsorship network. The data extends from the 93rd Congress to the 108th Congress, spanning about 30 years, so we will also be able to do it over an extended period of time. In addition, we will analyze the committee network, so we can compare the networks' features.

2 Data and Methods

In this study, legislation is defined to include all resolutions, public and private bills, and amendments. When processing data, we first create a *bipartite* adjacency matrix M, in which the rows correspond to bills and the columns correspond to Congressmen. An entry is 1 if a Congressman sponsored or cosponsored a bill and 0 otherwise. We can use this matrix to examine the cumulative distribution of Congressmen who have sponsored a given number of bills (the "degree distribution").

We show the degree distributions for the House of Representatives from the 93rd Congress to the 107th in Figure 1. Observe that the 93rd and 94th curves are to the left of the others; this is due to a rule change in the House after the 94th Congress that lifted a limit on the number of cosponsors of a bill. Otherwise, the degree distributions of the other Congresses are similar.

When considering the problem of finding communities in a network, we note that communities should intuitively have more links between them than what is expected. This is embodied by the concept of *modularity* [6]. The modularity Q of a set of communities is defined as

$$Q = (\text{number of edges within communities}) - (\text{expected number of such edges}).$$

Our objective is to find a partition of the network into communities such that the modularity is maximized. First we consider the problem of the optimal division of a network into two subcommunities. This problem resembles a "canonical" graph partitioning problem, which is solvable by methods such as spectral partitioning [7],[8]. However, the fact that the sizes of the communities are unknown prevents the methods of graph partitioning from working [9], which forces us to rely on a different approach.

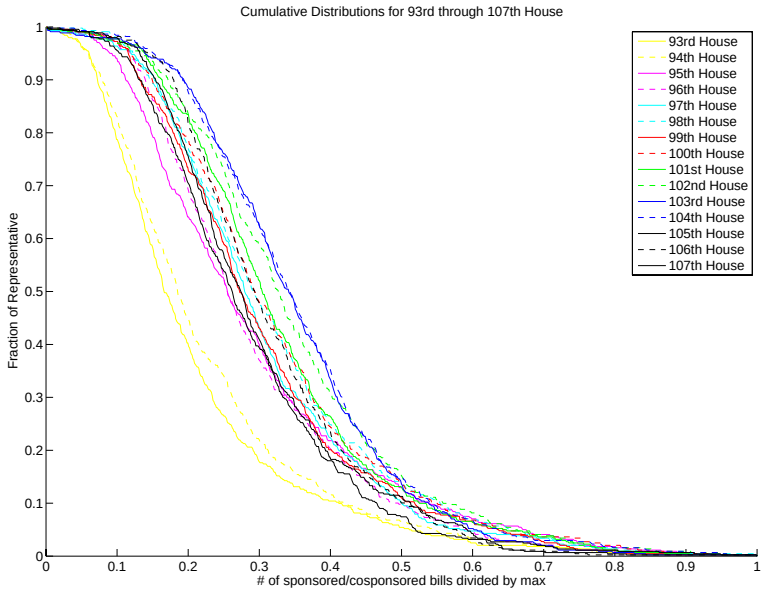

Fig. 1: Degree distribution of House from the 93rd to 107th Congress

We will formulate the modularity in matrix form [6]. Consider A, a unipartite *adjacency matrix*, and P, a "null model" matrix that gives the expected number of edges between any two vertices. For this study, the adjacency matrix $A = M^T M$, so that A_{ij} represents the number of bills sponsored or cosponsored by both Congressmen i and j, and P_{ij} is the expected number of bills sponsored by both Congressmen i and j. The modularity is then

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - P_{ij}] \delta(g_i, g_j), \tag{1}$$

where m is the total number of edges in the network, g_i is the community in which i is contained, and δ is the Kronecker delta function.

To proceed, one must find a suitable P. First, we assume that the sum of the entries in P are the same as the sum of the entries in A. That is

$$\sum_{ij} P_{ij} = \sum_{ij} A_{ij}. \tag{2}$$

Second, we assume the "null model" has approximately the same degree distribution as the real-world network. Therefore, we require that

$$\sum_{ij} P_{ij} = k_i, \tag{3}$$

where k_i, the degree of vertex i, is calculated by using

$$k_i = \sum_{j} A_{ij}. \tag{4}$$

Clearly, if (3) is satisfied, then (2) is automatically satisfied as well. If the edges in the null model are randomly placed subject to (3), then

$$P_{ij} = \frac{k_i k_j}{2m}. \tag{5}$$

Because we are considering the division of a network into just two subcommunities, we can define an index vector $\mathbf{s}$ with $|\mathbf{s}|$ = the number of nodes, and each entry has the value 1 if the corresponding node is in one community and the value -1 if it is in the other. Substituting for the Kronecker delta in (1) yields

$$Q = \frac{1}{4m} \mathbf{s}^T B \mathbf{s}, \tag{6}$$

where $B := A - P$ is the *modularity matrix*. If $\mathbf{s}$ were unconstrained, then the modularity would be maximized for $\mathbf{s}$ parallel to $\mathbf{u}_1$, the eigenvector of B with the largest eigenvalue (i.e., the leading eigenvector). However, the elements of $\mathbf{s}$ must be ± 1, so we must settle for an approximation. We want $\mathbf{s}$ to be as close to $\mathbf{u}_1$ as possible by assigning the value $+1$ to an element in $\mathbf{s}_1$ if the corresponding element in $\mathbf{u}_1$ is positive and assigning the value -1 if the corresponding element

is negative. If the corresponding element in $\mathbf{u}_1$ is 0, we subsequently assign the element to the group that would give the greatest modularity. This yields a division of the network into 2 subcommunities which maximizes the modularity.

If we apply this algorithm recursively (i.e., applying the algorithm to the two subcommunities that emerge after the algorithm finishes), and terminate the algorithm when there exists no subdivision that gives a positive modularity, we can divide the network into more than 2 subcommunities. In principle, this allows one to partition the network into a hierarchy of subcommunities. It is important to realize, however, that the formula for modularity changes as the algorithm proceeds. The change is necessary because the original formula results in an incorrect computation of the modularity of the full network when considering a subnetwork. Instead of merely taking the submatrix of the modularity matrix corresponding to a subnetwork, one applies the formula:

$$B_{ij}^{(G)} = B_{ij} - \delta_{ij} \sum_{l \in G} B_{il}, \tag{7}$$

where G is the set of nodes of the subnetwork in question [6]. Substituting (7) in place of the original in the algorithm yields the correct modularity change in the network rather than the modularity change of the subnetwork.

With the legislation cosponsorship data, this algorithm terminates after one stage. Each network is partitioned mostly according to political party, reaffirming the political truth that party is the most important determining factor of a Congressman's behavior. However, it is desirable to obtain additional hierarchical structure, so it is necessary to modify Newman's algorithm. Instead of terminating the algorithm when no division can be found, we apply the algorithm to the subnetwork while considering it as the full network. For the cosponsorship data, this modification will result in further subdivisions beyond the first split, which will allow us to see more hierarchical structure.

3 Results and Discussion

The above algorithm yields not only a set of subcommunities but also the way in which the network was partitioned to created them. This process can be represented by drawing a tree, or a *dendrogram*, that shows the hierarchical structure of the communities. We depict dendrograms in polar coordinates for visual clarity. The leaves can be colored several ways in order to reveal network features. In this study we use coloring schemes by party affiliation, state, and a rank-ordering from conservative to liberal created using the DW-NOMINATE scores of Poole and Rosenthal, which are based on the voting record [10], [11]. We have produced dendrograms for the House and Senate from the 93rd Congress (1973-1974) to the 108th Congress (2003-2004).

In addition to the legislation cosponsorship dendrograms, we also applied the algorithm to produce dendrograms of the committee membership data from the 101st to the 108th House. The dendrograms consist of the subcommittees and committees as leaves and are colored by parent standing committee.

We examine the networks at a given hierarchical level in a dendrogram by producing a pie chart with the composition of each community at that level. This chart is drawn using an algorithm from Kamada and Kawai [12]. The pie charts are useful for seeing how the party/state compositions of the communities differ at different levels, especially for the House dendrograms, where labeling by name is not practical. Because the compositions of the communities is not clear by reading the labels, the pie charts help to convey the community compositions at each level.

3.1 Legislation Cosponsorship

The first result that is evident from all of the dendrograms for both the House and Senate is that the parties are almost completely separated when the original network is partitioned into two subnetworks. However, each dendrogram includes some Congressmen who appears with members of the opposite party. We observe as well that the number of people identified with the incorrect party decreases as time passes. For example, for the 108th Senate (Figure 2), there are approximately 10 Senators who appear with the opposite party. However, in the 96th Senate (Figure 3), there are approximately 25 Senators who appear with the opposite party. These observations support the contention that Congress has become more polarized over the past 30 years.

The dendrograms also pick out some known moderate Senators who sometimes collaborate more with members of the opposite party. For example, in Figure 2, we can see that several liberal Republicans such as Lincoln Chafee [R-RI], Arlen Specter [R-PA], and (former Republican) James Jeffords [I-VT] appear to be closely connected to the Democrats, while several conservative Democrats such as Zell Miller [D-GA], John Breaux [D-LA], and Kent Conrad [D-ND] appear to be closely conected to the Republicans.

Coloring the dendrograms by state is also very insightful. First, we note that the communities show a positive correlation to state. For example, in Figure 4 several communities include similar and even identical colors. This is also reasonable, as many of the bills and amendments involve geographic-specific themes such as "pork." Examining the state and party dendrograms together reveals a group of Southern Democrats that consistently sides with the Republicans. In addition, we observe that this group starts as a very large bloc in the earlier Congresses (it is almost the same size as the entire body of Republicans) but decreases to a much smaller group by the later Congresses. In Figure 4, for example, a small group of Southern Democrats appears among the Republicans around the 10 o'clock position.

This change has been postulated to be concomitant with the 1994 Congressional elections, the so-called "Republican Revolution," in which the Republicans gained control of the House (for the first time since 1954) and the Senate. However, our analysis suggests that this process has occurred gradually from the 93rd Congress (1973) to the 108th Congress (2003), rather than as a sudden change from the 103rd Congress to the 104th.

By coloring the dendrogram according to DW-NOMINATE rank-ordering,

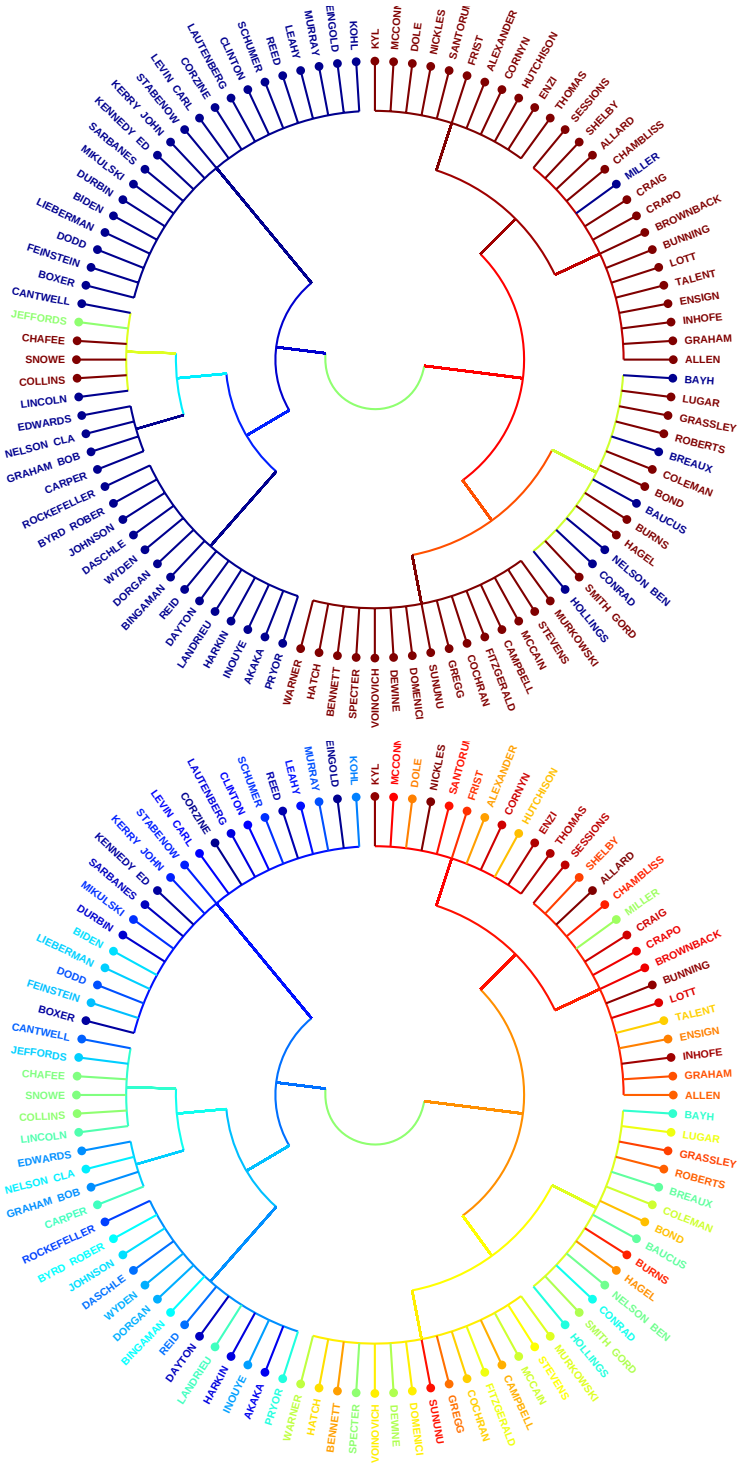

Fig. 2: Dendrograms of the 108th Senate (2003-2004) colored by party (top) and DW-NOMINATE rank-ordering (bottom). Democrats are blue, Republicans are red, and the independent Jeffords is green. Deep blue is most liberal and deep red is most conservative in the DW-NOMINATE dendrogram.

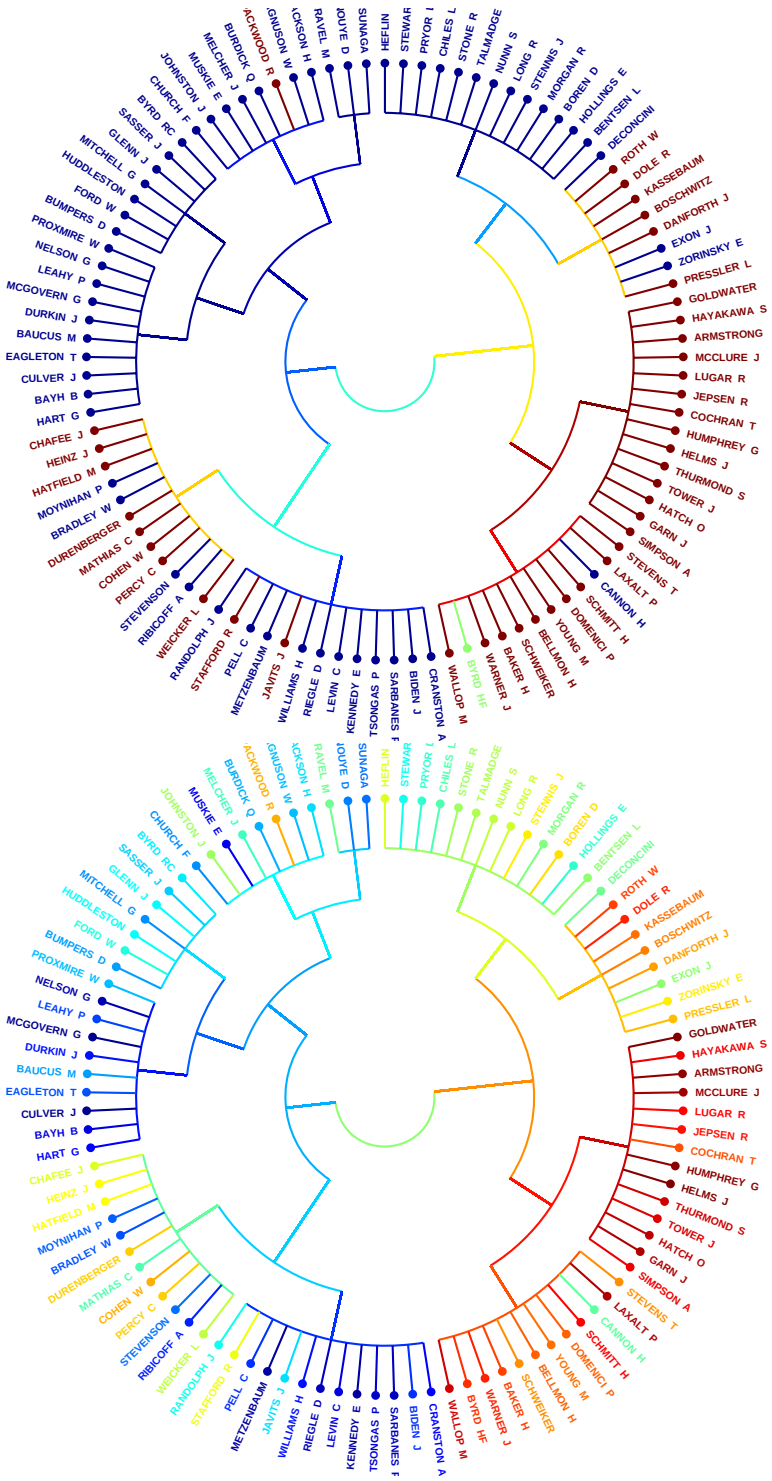

Fig. 3: Dendrograms of the 96th Senate (1979-1980) colored by party (top) and DW-NOMINATE rank-ordering (bottom).

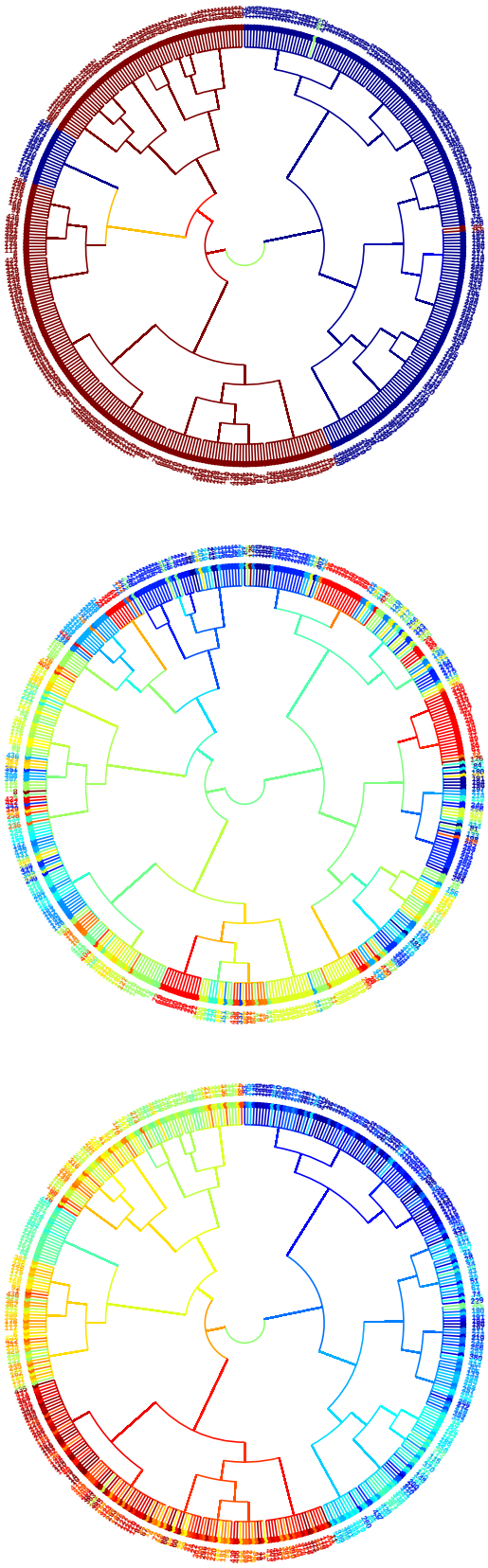

Fig. 4: Dendrograms of the 108th House colored by party (top), state (middle), and DW-NOMINATE rank-ordering (bottom).

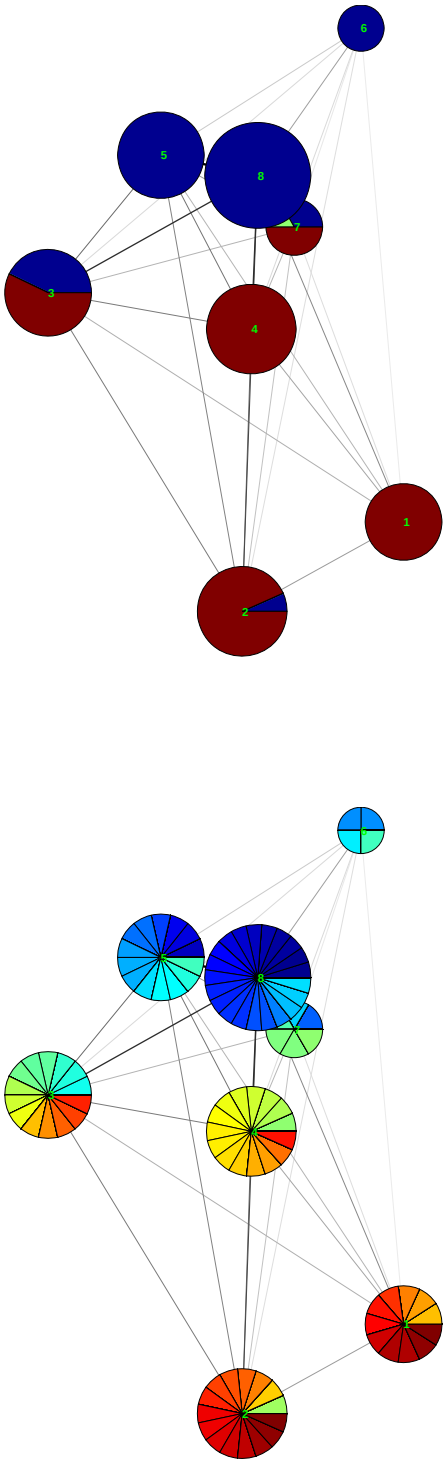

Fig. 5: Pie chart of the 108th Senate colored by party (top) and by DW-Nominate rank-ordering (bottom). Coloring schemes are the same as in the corresponding dendrogram, and numerical labels are used to determine the actual members of the communities.

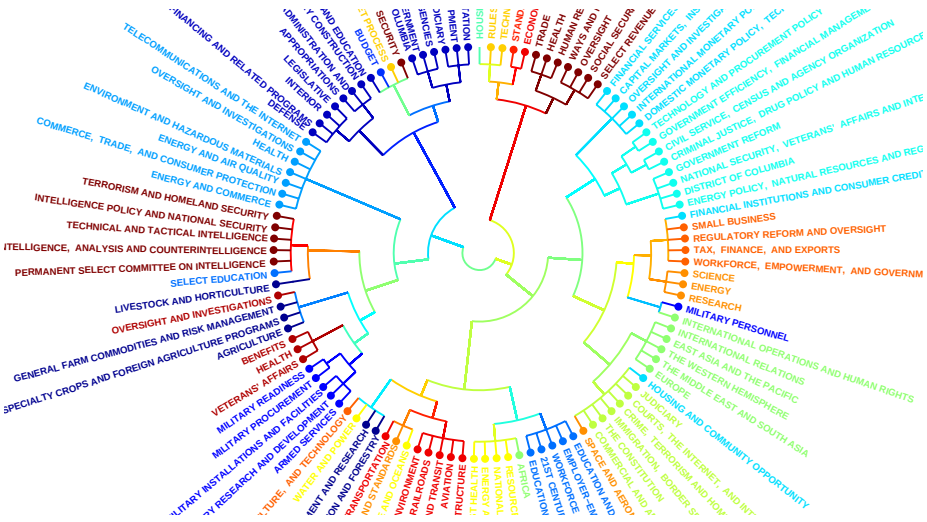

Fig. 6: Dendrogram of the 107th House committee assignment network (colored by parent committee.)

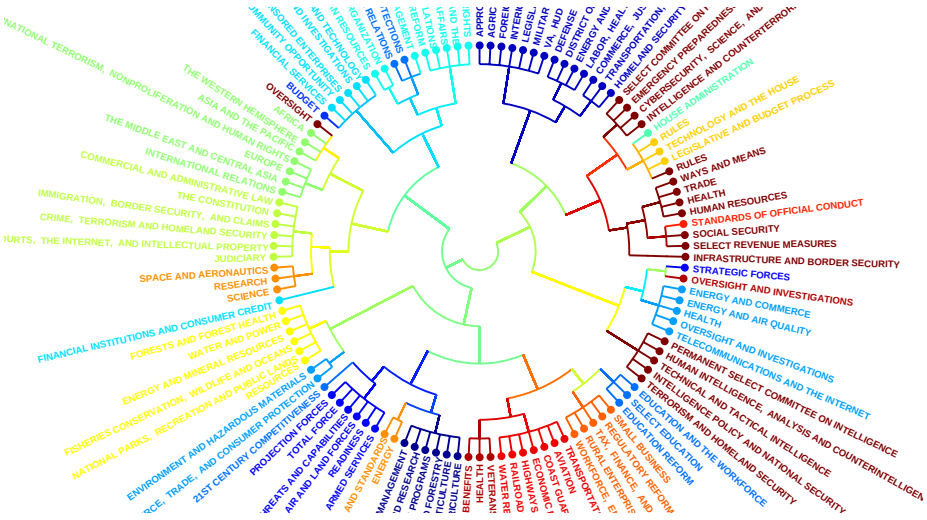

Fig. 7: Dendrogram of the 108th House committee assignment network (colored by parent committee.)

we observe a strong correlation with people of similar political ideology sponsoring legislation together. For example, in the 108th Senate, by examining the DW-NOMINATE pie chart (Figure 5), we can observe that most of the communities consist of Congressmen in a small ideological range, because the colors in the pies are very similar. In this case, the Southern Democrats and Northeastern Republicans who appear with members of the opposite party appear as moderates. This is because their DW-NOMINATE scores tend to be close to the median, as they vote with their own party on party-line legislation but vote against their party on many other issues.

3.2 Committee Membership

By examining the committee network dendrograms (Figures 6 and 7), we observe that most of the communities consist of groups of subcommittees of the same parent committee. This is consistent with the partitions found using other methods of community detection [3]. However, we note that in the 107th House, at approximately the 11 o'clock position of Figure 6 the Select Committee on Homeland Security appears in the same community as the Legislative and Budget Process Subcommittee of the Rules Committee. In the 108th House, a similar result can be observed around the 1 o'clock position of Figure 7, where the Select Committee on Homeland Security and its subcommittees appear as a community next to a community containing the Rules Committee and its subcommittees.

This connection between the Rules Committee and the Select Committee on Homeland Security was also observed by Porter using different algorithms [3], but in that study Homeland Security appeared with the entire Rules Committee. This set of results suggests that the strong connection between the committees is mostly due to the Legislative and Budget Process Subcommittee, and this possibility can be confirmed by examining a list of Committee memberships of the 107th House. We indeed see that the Rules Committee and the Select Committee on Homeland Security share two Congressmen, Martin Frost [D-TX] and Deborah Pryce [R-OH], and both of these Congressmen serve on the Legislative and Budget Process subcommittee, but not the other subcommittee of the Rules Committee.

There are several structural differences between the committee network and the legislation cosponsorship network. One is that the modularity values of the committee dendrograms (about .4) are significantly higher than those of the cosponsorship data (about .1). Because the number of committees is approximately the same as the number of Senators and is far less than the number of Representatives, we can eliminate the positive correlation of modularity with network size as a factor. Therefore, we conclude that the community structure in the committee network is stronger than in the cosponsorship network.

Another difference is that the dendrograms produced from the committee-assignment network have more hierarchical structure than those produced using the cosponsorship network. For the latter, Newman's algorithm produced only one partition, whereas for the former the algorithm typically terminates after 3

to 4 levels. This represents the difference in the maximum modularity level of the two data sets, which can be interpreted as the partition of the network which gives the strongest community structure. In the cosponsorship network, the maximum modularity level roughly corresponds to the party split. In the committee network, the maximum modularity level roughly corresponds to groups of standing committees.

The committee network also contains more hierarchical levels with our modification of Newman's algorithm, containing approximately 10-12 levels for the committee data, while the cosponsorship data contains approximately 5 levels for the Senate and 7-9 for the House. These observations could be a result of the fact that the cosponsorship network contains far more links than the committee network and have a similar number of nodes, as each Congressman is limited to being on two parent committees can sponsor any amount of legislation. It is possible that the many additional links in the cosponsorship network obscures the community structure, yielding lower modularity values and a less hierarchical structure.

4 Conclusions and Further Work

Employing a modification of Newman's modularity-maximizing algorithm, we produced dendrograms and pie charts of the legislation cosponsorship networks of the 93rd-108th Congresses. We used different coloring schemes of the dendrograms to make observations about the Congressional network, such as the identification of moderates, the increase of political polarization over the past thirty years, and the positive correlation of communities to state and partisanship rank-ordering. By examining the committee network, we were able to clarify Porter's results concerning the Homeland Security and Rules committees and compare the structure of the two networks.

Further work includes refining the algorithm so that other algorithms are used to further partition the network from its "preferred" level to its leaf structure. In addition, centrality measures can be applied to the communities to determine which communities play important roles in the network, and the roll-call vote network can also be analyzed to provide an additional comparison with the legislation cosponsorship and the committee networks.

Acknowledgements

I would like to acknowledge Caltech's SURF program and Robert F. Blinkenberg for funding my research. I would also like to acknowledge James Fowler for providing the legislation cosponsorship data. In addition, I would like to thank my mentor Mason Porter, and other members of his collaboration, Peter Mucha, Mark Newman, AJ Friend, and Fowler for their advice and help.

References

[1] Milgram, S. The small world problem. **1967**. *Psychology Today* **1**(1): 60-67.

[2] Porter, M. A., Mucha, P. J., Newman, M. E. J., & Warbrand, C. M. A network analysis of committees in the United States House of Representatives. **2005**. *Proceedings of the National Academy of Sciences* **102**(20), 7057-7062.

[3] Porter, M. A, Mucha, P. J, Newman, M. E. J, & Friend, A. J. Community Structure in the United States House of Representatives. **2006**. physics/0602033.

[4] Fowler, J. H. Legislative cosponsorship networks in the US House and Senate. **2006**. *Social Networks* **28**(4), 454-465.

[5] Fowler, J. H. Connecting the Congress: A Study of Cosponsorship Networks. **2006**. *Political Analysis* **14**(4), 456-487.

[6] Newman, M. E. J. Finding community structure in networs using the eigenvectors of matrices. **2006**. *Physical Review E* **74**, 036104.

[7] Fiedler, M. Algebraic connectivity of graphs. **1973**. *Czech. Math. J.* **23**, 298-305.

[8] Pothen, A., Simon, H., & Lou, K. P. Partitioning sparse matrices with eigenvectors of graphs. **1990**. *SIAM J. Matrix Anal. Appl.* **11**, 430-452.

[9] Newman, M. E. J. Modularity and community structure in networks. **2006**. *Proceedings of the National Academy of Sciences* **103**(23), 8577-8582.

[10] Poole, K. T. Voteview Website. http://www.voteview.com.

[11] Poole, K. T. and Rosenthal, H. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press, 1997.

[12] Kamada, T. and Kawai, S. An Algorithm for drawing general undirected graphs. **1989**. *Information Processing Letters* 31, 7-15.