Identifying Factors which Lead to Community Development in Complex Networks Via Ontological Structure: Cosponsorship Communities in the U.S. Senate

Jean "Betweenness" Hausser, Mark Rivera and Skyler Cranmer

Abstract

Substantive Background The production of legislation in the U.S. Senate is a complex phenomenon shaped by a combination of formal rules, party politics, and personal maneuvering. We examine the dynamics of legislative cosponsorship – when legislators write legislation together – in the Senate using data from Senates 106-108.

Questions We seek to understand the processes by which "cosponsorship communities" emerge. What factors contribute to clusters of Senators writing bills with one another? Is party the only driving force in community formation or do personal or network characteristics matter?

Contributions We build upon previous analyses in several ways. First, we have expanded the breadth of the data available for network analysis on the Senate. More importantly, we have synthesized methods for modularity analysis and specification searching to create a means by which the factors leading to community formation can be identified with reasonable confidence. Previous methods and analyses have only been able to identify that communities exist within networks, but have not been able to identify the factors which predict them.

Methods We expand upon Leicht and Newman's [4] method for modularity analysis in weighted/directed networks by introducing to it the method of ontological structure specification. Our hybrid technique allows us to identify factors that are disproportionately associated with each community in the network. This methodological extension augments standard community detection algorithms by identifying the defining characteristics of communities within the network; something previous techniques have not been capable of doing.

Findings We find, not surprisingly, that party affiliation plays a strong role in shaping cosponsorship networks. In each Senate analyzed the Democratic and Republican Parties constitute distinct communities. Interestingly, however, we also find that, when the Democratic Party lost their Senate majority coming into the 108^{th} Senate, the party fractioned while the Republican party remained unified. The Democrats split into a moderate faction that cooperated more frequently with the Republican majority and a faction with cooperated much less with the majority.

☐ Affiliations
Jean Hausser is a Ph.D. Candidate in the RNA regulatory networks group at the Biozentrum at the **Universität Basel**. Mark Rivera is a Ph.D. Candidate at the Kellogg School of Management at **Northwestern University**. Skyler Cranmer is an Assistant Professor of Political Science at the **University of North Carolina, Chapel Hill**.

☐ Support
This work was partially supported by the Santa Fe Institute whose research and education programs are supported by core funding from the National Science Foundation and by gifts and grants from individuals, corporations, other foundations, and members of the Institute's Business Network for Complex Systems Research.

"We were born to unite with our fellow men, and to join in community with the human race."

- Cicero

In the U.S. Senate – as with most legislative bodies around the world – the giving and receiving of political support is key to legislative productivity at the Senate level as well as success and influence at the individual level. A nascent literature on the subject [1][2][11] confirms what we know anecdotally to be true: that sub-communities of support form within the greater body of the Senate and the communities' members help each other achieve their political goals.

There are a number of ways in which legislators express support for one another. Ideally, we would have an index of support which included backroom meetings and closed-door conversations. Unfortunately, such data are not available and what Senators say publicly about each other is not necessarily indicative of the way they feel. Early analyses of the Congress have suggested that a good measure of support from one legislator to another is whether legislators vote for each others' bills [9][10], but the more recent consensus in the literature is that agreement on roll call votes is indicative of ideological agreement more than interpersonal support [8].

We think voting support is suboptimal in its ability to reflect signals of support: many votes are non-controversial and voting is generally considered a cheap-talk signal [1]. For this reason, we follow the recent trend [1][2][11] of using legislative cosponsorship as a measure of interpersonal support within the Congress. Cosponsorship, the writing of a bill *with* another legislator, represents a much stronger signal of support for the primary sponsor (author) of the bill than simply voting for the bill because the cosponsor's name becomes inextricably tied to the legislation, its policy objectives, and the sponsor of the bill.[1]

[1]The U.S. Senate has rather idiosyncratic rules about sponsorship and cosponsorship. Each bill may have only one sponsor (the primary author of the bill) and may then have any number of cosponsors; it is not possible to have two sponsors of any bill.

Senate	n	Mean In-Degree
106	102	99.07
107	101	98.36
108	100	106.7

Table 1: Descriptive characteristics of cosponsorship networks in the 106^{th} through 108^{th} Senates

The cosponsoring of multiple bills by multiple legislators over time creates a complex and dynamic network which we believe holds insights to the power structure and legislative process of the Senate. We hypothesize that *the development of subgroups within the cosponsorship network, or "cosponsorship communities" is driven by a number of factors including party identification, the individual's place in the larger network, committee assignments, and the legislator's personal characteristics.*

Our methods and analysis build upon previous work in several ways. First, we have expanded the breadth of the data available for network analysis on the Senate. More importantly, we have synthesized methods for modularity analysis and specification searching to create a means by which the factors leading to community formation can be identified with reasonably confidence. Previous methods and analyses have only been able to identify that communities exist within networks, but have not been able to identify the factors which predict them.

Network Properties

We expanded a dataset originally collected by Fowler [1][2] in which, for each Senate, every vertex represents a Senator and each edge is a directed edge representing the cosponsorship of a bill written by Senator A by Senator B. In other words, there are 100 vertices in each Senate[2] and each bill that is considered by the Senate – regardless of whether it was passed into law or not – can generate a non-

[2]50 states, two senators per state. Sometimes the number of vertices in a given Senate can be fewer than 100. This occurs when seat is left vacant before the beginning of a senate cycle (i.e. due to death or scandal) and no replacement is appointed. A given Senate may have more than 100 vertices if a Senator leaves the Senate once the legislative cycle has started and a replacement is appointed. This will look like more than two Senators from a given state in a given cycle, but the *extra* Senator did not serve alongside the one who left.

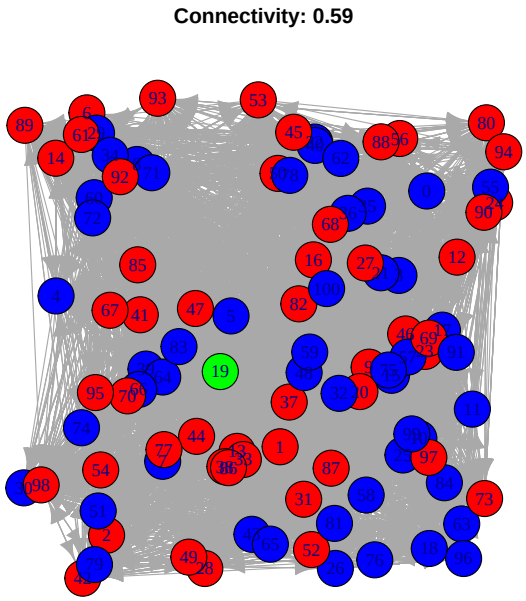

Figure 1: Two Dimension graph of the cosponsorship network from the 107^{th} Senate. This figure is representative of all Senate cycles under consideration. The network contains 101 nodes, and has an average incoming degree of 98.36 edges per vertex. The blue vertices represent Democrats, the red vertices represent Republicans, and the single green vertex is former Republican Jim Jeffords of Vermont who declared himself an Independent part way through the 107^{th} Senate.

negative number of cosponsorship edges directed at the sponsor of the bill. Our date cover three legislative cycles, the 106^{th}, 107^{th} and 108^{th} Senates corresponding to the years 1999 through 2005. Table 1 presents descriptive statistics for the three Senates under consideration.

The cosponsorship networks are dense. As can be seen in Table 1, the average in and out degrees are higher than is typical in social networks. This makes cosponsorship communities – clusters of Senators frequently cosponsoring each others bills – difficult to identify. As can be seen in Figure 1, the density of the network, even when vertices are colored by party, precludes any obvious visual segmentations of the network into sub communities. For this reason, the modularity analysis we discuss and perform below is necessary.

Vertex Characteristics

For each node, we expanded Fowler's [1][2] data by collecting additional personal details on each Senator in the data set. In addition to party, race and gender, we gathered data on each Senator's age (year of birth), place of birth (often different from the state they represent), their highest level of educational attainment, whether they served in the military, whether they fought in a war, whether they received the Medal of Honor, and what their pre-politics profession was. Enhancing the detail of vertex-specific data allowed us to conduct an ontological study as described below and identify the factors which drive cosponsorship community formation.

Methods

To accomplish both the division of the greater network into cosponsorship communities *and* the identification of the factors which drive community clustering, we have hybridized a modularity algorithm created by Leicht and Newman [4] and a method of model identification from genetics research [3]. To best understand our hybrid, let us first examine the component parts.

Modularity Detection in Directed and Weighted Graphs. Modularity is a simple and intuitive, yet powerful concept developed by Newman an his collaborators [5][6][7][4]. The intuitive premise is that modularity can be measured by the extent to which there are fewer edges between divided groups than would be expected by chance alone. As Newman [7] writes "true community structure in a network corresponds to a statistically surprising arrangement of edges" (p.4).

Until recently modularity analysis was usually conducted on undirected networks, or by treating directed networks as undirected. In a natural extension of previously established modularity methods, Leicht and Newman [4] develop a powerful method of analysis which can accommodate directed and weighted edges. This extension still looks for divisions of the network in which there are more edges within communities than would be expected, but takes the direction of edges into account. Intuitively, if one Senator A has a high in-degree but a low out degree and Senator B has a low in-degree but high out-degree, it should be more "statistically surprising" for an edge to run from B to A than from A to B.

Formally, if k_i^{in} and k_j^{out} are the in-degree and out-degrees of vertices i and j respectively, then $k_i^{in}k_j^{out}/m$, where m is the total number of edges, is the probability of a vertex from j to i. Now, directed modularity can be defined as

$$Q = \frac{1}{m}\sum_{ij}\left[A_{ij} - \frac{k_i^{in}k_j^{out}}{m}\right]\delta_{c_i,c_j},$$

where A_{ij} is an adjacency matrix, δ_{ij} is Kronecker's delta, and c_i is the label of the community to which i is assigned. Leicht and Newman further note that – as one would hope – when k_i^{in} and/or k_j^{out} is small, $j \to i$ edges make larger contributions to Q. The algorithm then searches for the division of the network into communities c_i that maximize Q. While simulated annealing is the most thorough way of conducting this search, the problem is NP hard [7] so we use Newman's [7] spectral optimization method for its greater efficiency.

The problem we face, is that we can now divided a given network into $c_i, i = 1, \ldots, N$ communities, but we have no idea what those communities signify! While dividing a network into many subgroups is illustrative of the structure of the network, the ability to do this is not particularly helpful when seeking to answer substantive questions.

Future Methods

In order to empirically examine the generative social processes that lead to the establishment of cosponsorship communities, we propose an application of gene ontology analysis [3]. In the study of gene expression, gene ontology analysis is frequently used to ascertain the unique functions of a particular gene. In much the same way, we propose that a hybrid gene-ontology analysis can effectively identify the factors that differentiate cosponsorship communities. That is, community detection algorithms decompose the network into discrepant groups; gene-ontology analysis highlights the differences between these communities.

We identify three meta-factors that may serve as the basis for community formation: personal characteristics, political connections, and network characteristics. These three meta-categories are subsequently used as a basis for the ontological structure (see figure 2.). The ontological structure is in fact its own directed network, though unlike the cosponsorship network, all ontological networks are strictly acyclic. Each vertex is a characteristic with directed edges leading to more specific characteristics which, in turn, can have directed edges to more specific characteristics. In other words, the specificity of the characteristic captured by a vertex is strictly increasing in the number of steps taken away from the three most general domains.

The ontology we specified for the Senate is drawn in Figure 2.

The majority of ontological analyses make use of the hypergeometric distribution to measure overrepresentation of a particular characteristics [3]. Rather than examining each term in an ontology individually, Grossmann et al developed an alternative approach that examines each term in the context of its parent terms.

Following Grossmann et al [3], let P represent the population and S the study set with sizes m and n respectively. If the term we want to measure overrepresentation on is t, we can write P_t for the set of characteristics related to to t with some number of elements m_t. The S_t and n_t are analogous to these quantities but in the study set rather than the population.

We can write Σ to represent a random sample from P of size n and let σ_t be the number of vertices in Σ that are related to t. Write $pa(t)$ to represent

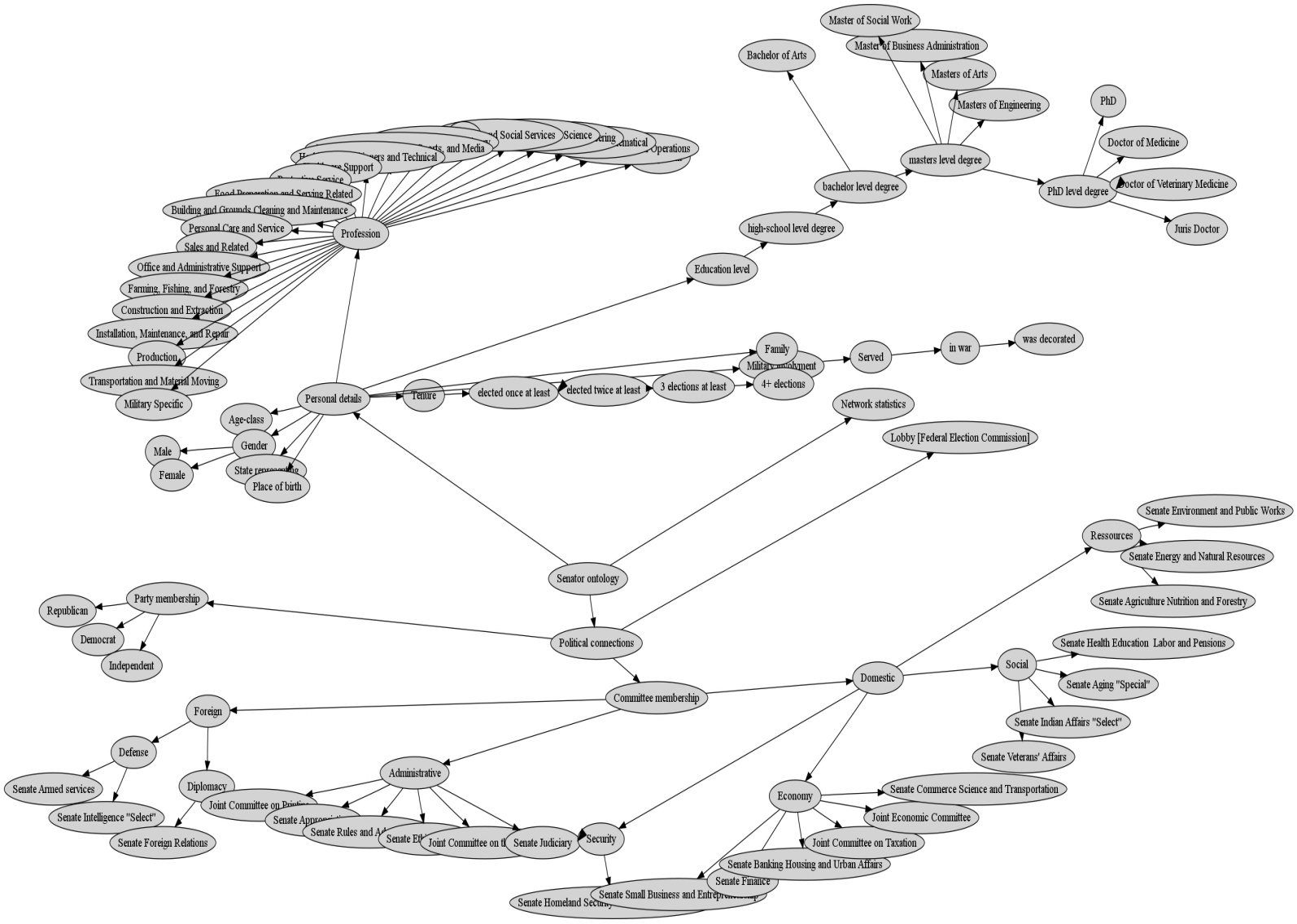

Figure 1: The Senate Ontology. This figure shows the directed acyclic graph we use to structure the variables characterizing the senators. Notice that the more steps one moves from the top of the ontology, the more specific the vertices become.

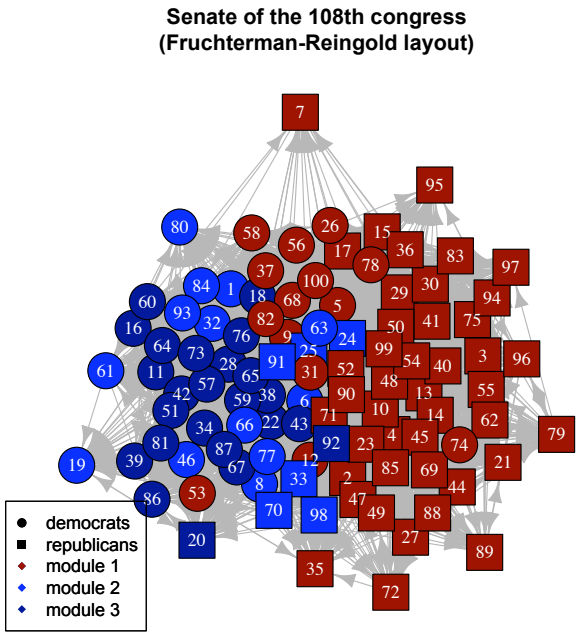

Figure 3: Cosponsorship communities in the 108^{th} Senate. This figure shows the three cosponsorship communities which developed when the Democratic party seems to have split after losing its Senate majority coming out the 107^{th} Senate. Democratic Senators are drawn as circles while Republican Senators are drawn as squares. While the first community, colored in red, is not exclusively Republican, it is easy to see that it is predominantly Republican. Of the two Democratic communities we identified, the one we colored in light blue cooperated more frequently with the Republican majority than the community colored in dark blue.

the parent (less specific) nodes of t. It is now possible to compute the probability of observing exactly σ_t characteristics using the hypergeometric distribution conditioning on the event that the overlap of $P_{pa(t)}$ and Σ is observed :

$$\mathbb{P}(\sigma_t = k|\sigma_{pa(t)} = n_{pa(t)}) = \frac{\binom{m_t}{k}\binom{m_{pa(t)}-m_t}{n_{pa(t)}-k}}{\binom{m_{pa(t)}}{n_{pa(t)}}}.$$

It is now straightforward to calculate the significance of the overrepresentation by summing over the probabilities for $n_{pa(t)} < \mathbb{P} < \min(m_t, n_{pa(t)})$.[3]

It should be clear that using the ontological method in conjunction with directed modularity analysis will allow us not only to identify cosponsorship communities via spectral optimization, but to identify the factors which lead to community formation if the relevant variables are present in our dataset.[4]

Results

The directed modularity analysis above revealed an interesting pattern. One expects to find that partisanship plays a major role in cosponsorship community formation; and our study is no exception. We see the identification of a party cleavage as further validation of the directed modularity method. We were further encouraged by the fact that the Senate does not break down into a large number of very small communities, but rather sorts itself into just a few.

The most interesting result was that the 108^{th} Senate splits three ways. One community is comprised predominantly of republicans, and two communities consisting largely of democrats. The Republican party controlled both houses of Congress during the 108^{th} Senate and seems to have functioned as a fairly cohesive cosponsorship community. While the Democrats, relegated to the minority, fractured and split into two communities, one which cooperated more closely with the Republican majority and one which cooperated much less with the majority. The cosponsorship communities of the 108^{th} Senate are shown in Figure 2.

The bifurcation of the minority party into two cosponsorship communities does not appear to be a general pattern. When the republicans were the minority party in the prior Senates, they none-the-less maintained a single cosponsorship community. Thus, the splitting of the Democratic Party into two com-

[3] The significance calculation is somewhat more involved if t has more than one parent. See Grossmann et all [3] for details.

[4] Indeed, most any statistical model assumes no omitted variable bias. Though this assumption is never demonstrably true or untrue, it is ubiquitously made and we make it here.

munities in the 107th Senate is particular interesting. The Senate term began with the Republicans controlling both chambers; the Senate by a single vote. However, on June 6, 2001 Senator Jim Jeffords, a Republican from Vermont, broke with the party, declared himself an Independent, and announced that he would vote with the Democrats. This gave the Democrats a majority in the Senate and Tom Daschle became the majority leader. Because of the close margin of the majority, holding the party line became very important and both parties managed to maintain single cosponsorship communities: the Senate split down the middle by party lines.

As it stands, the source of the cleavage within the Democratic party during the 108^{th} Senate is unidentified. However, the ontological method will identify the source if the assumption that the relevant variables are included in our dataset holds. We leave this for future research.

References

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

[2] Fowler, J. H. "Legislative Cosponsorship Networks in the U.S. House and Senate" *Social Networks*, 28(4): 454-465 (2006).

[3] Grossmann, S., S. Bauer, P. N. Robinson, M. Vingron "Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis" *Bioinformatics*, 23(22): 3024-3031 (2007).

[4] Leicht, E.A. and M. E. J. Newman "Community Structure in Directed Networks." *Physics Review Letters*. 100: 118703 (2007).

[5] Newman, M. E. J. "The Structure and Function of Complex Networks." *SIAM Review*. 45: 167-256 (2003).

[6] Newman, M. E. J. and M. Girvan "Finding and Evaluating Community Structure in Networks" *Physics Review E*, 69: 026113 (2004).

[7] Newman, M. E. J. "Modularity and Community Structure in Networks." *PNAS*. 103: 8577-8582 (2006).

[8] Poole, K. T. and H. Rosenthal "Patterns of Congressional Voting" *American Journal of Political Science*, 35(1): 339-278 (1991).

[9] Rice, Stuart A. "The Identification of Blocks in Small Political Bodies" *The American Political Science Review*, 21(3): 619-627 (1927).

[10] Truman, David *The Congressional Party: A Case Study* New York: Wiley (1957).

[11] Zhang, Yan, A.J. Friend, Amanda L. Traud, Mason A. Porter, James H. Fowler, and Peter J. Mucha "Community Structure in Congressional Cosponsorship Networks" *Physica A*, 387(7): 1705-1712 (2007).

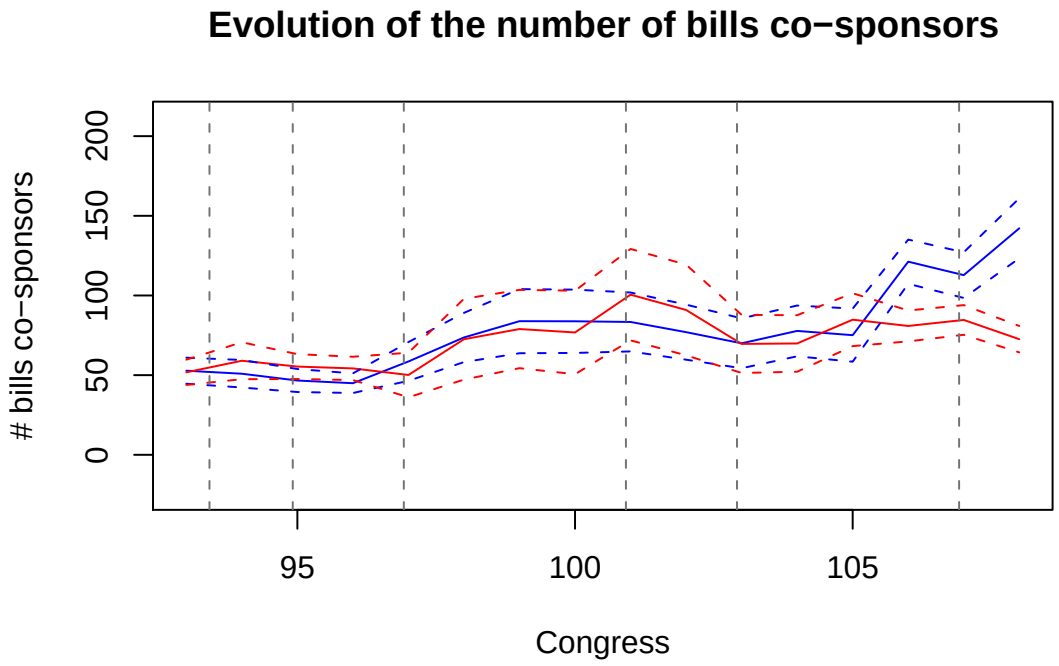

Figure 2: The mean number of bills cosponsored over multiple Senates. The dashed vertical lines represent transitions in the White House. Presidents: Nixon (tail end), Ford, Carter, Reagan, Bush, Clinton, and Bush respectively. The divergence occurs in the 106^{th} Senate when the "New Republicans" lead by Newt Gingeritch took over both chambers of Congress [11].

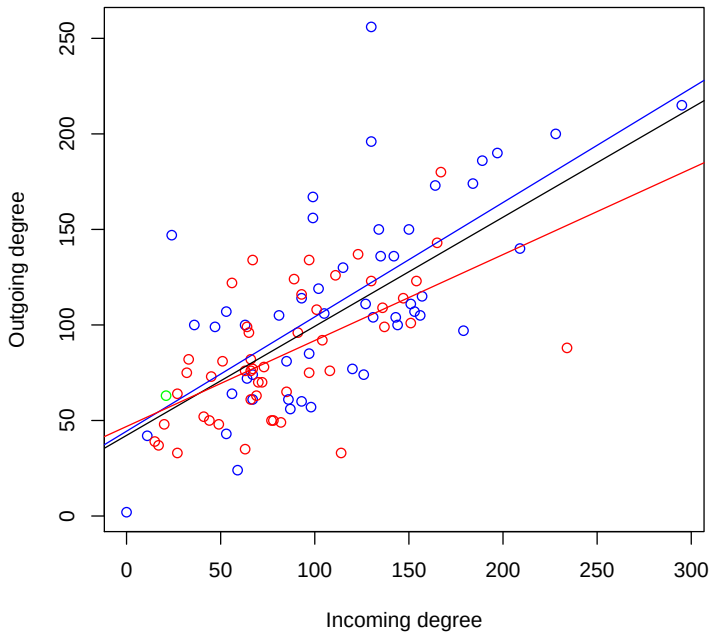

Figure 3: This figure plots in-degree against out-degree for the 107^{th} Senate. The lines are Least Squares regression lines. The red line is fit only to Republicans, the blue line is fit only to Democrats, and the black line is fit to the entire Senate.

Democratic Community #1

	Name	Thomas	Seniority	Party
1	Akaka Daniel K.	14400	14	D
2	Bayh Evan	49901	3	D
3	Biden Joseph R. Jr.	14101	16	D
4	Carper Thomas R.	15015	11	D
5	Coleman Norm	40302	1	R
6	Collins Susan M.	49703	4	R
7	Dayton Mark	40101	2	D
8	DeWine Mike	15020	11	R
9	Graham Bob	15503	9	D
10	Kohl Herb	15703	8	D
11	Landrieu Mary L.	49702	4	D
12	Levin Carl	14709	13	D
13	Lugar Richard G.	14506	14	R
14	Nelson Bill	14651	13	D
15	Pryor Mark L.	40301	1	D
16	Rockefeller John D. IV	14922	10	D
17	Snowe Olympia J.	14661	13	R
18	Stabenow Debbie	29732	4	D
19	Voinovich George V.	49903	3	R

Democratic Community #2

	Name	Thomas	Seniority	Party
1	Boxer Barbara	15011	11	D
2	Byrd Robert C.	1366	26	D
3	Cantwell Maria	39310	6	D
4	Chafee Lincoln	49905	3	R
5	Clinton Hillary Rodham	40105	2	D
6	Corzine Jon S.	40104	2	D
7	Dodd Christopher J.	14213	15	D
8	Durbin Richard	15021	11	D
9	Edwards John	49902	3	D
10	Feingold Russell D.	49309	6	D
11	Feinstein Dianne	49300	6	D
12	Harkin Tom	14230	15	D
13	Jeffords James M.	94240	2	I
14	Kennedy Edward M.	10808	21	D
15	Kerry John F.	14920	10	D
16	Lautenberg Frank R.	14914	11	D
17	Leahy Patrick J.	14307	15	D
18	Lieberman Joseph I.	15704	8	D
19	Mikulski Barbara A.	14440	14	D
20	Murray Patty	49308	6	D
21	Reed Jack	29142	7	D
22	Sarbanes Paul S.	13039	17	D
23	Schumer Charles E.	14858	12	D
24	Specter Arlen	14910	12	R

Republican Community

	Name	Thomas	Seniority	Party
1	Alexander Lamar	40304	1	R
2	Allard Wayne	29108	7	R
3	Allen George	29148	7	R
4	Baucus Max	14203	15	D
5	Bennett Robert F.	49307	6	R
6	Bingaman Jeff	14912	11	D
7	Bond Christopher S.	15501	9	R
8	Breaux John B.	13056	16	D
9	Brownback Sam	29523	5	R
10	Bunning Jim	15406	9	R
11	Burns Conrad R.	15701	8	R
12	Campbell Ben Nighthorse	95407	5	R
13	Chambliss Saxby	29512	5	R
14	Cochran Thad	14009	16	R
15	Conrad Kent	15502	9	D
16	Cornyn John	40305	1	R
17	Craig Larry E.	14809	12	R
18	Crapo Mike	29345	6	R
19	Daschle Thomas A.	14617	13	D
20	Dole Elizabeth	40303	1	R
21	Domenici Pete V.	14103	16	R
22	Dorgan Byron L.	14812	12	D
23	Ensign John	29537	5	R
24	Enzi Michael B.	49706	4	R
25	Fitzgerald Peter	49900	3	R
26	Frist William H.	49502	5	R
27	Graham Lindsey	29566	5	R
28	Grassley Chuck	14226	15	R
29	Gregg Judd	14826	12	R
30	Hagel Chuck	49704	4	R
31	Hatch Orrin G.	14503	14	R
32	Hollings Ernest F.	11204	19	D
33	Hutchison Kay Bailey	49306	6	R
34	Inhofe James M.	15424	9	R
35	Inouye Daniel K.	4812	23	D
36	Johnson Tim	15425	9	D
37	Kyl Jon	15429	9	R
38	Lincoln Blanche L.	29305	6	D
39	Lott Trent	14031	16	R
40	McCain John	15039	11	R
41	McConnell Mitch	14921	10	R
42	Miller Zell	49904	3	D
43	Murkowski Lisa	40300	1	R
44	Nelson E. Benjamin	40103	2	D
45	Nickles Don	14908	12	R
46	Reid Harry	15054	11	D
47	Roberts Pat	14852	12	R
48	Santorum Rick	29141	7	R
49	Sessions Jeff	49700	4	R
50	Shelby Richard C.	94659	5	R
51	Smith Gordon H.	49705	4	R
52	Stevens Ted	12109	18	R
53	Sununu John E.	29740	4	R
54	Talent Jim	29369	6	R
55	Thomas Craig	15633	8	R
56	Warner John	14712	13	R
57	Wyden Ron	14871	12	D

Table 2: The three cosponsorship communities detected in the 108^{th} Senate and their membership. Aside from their names, this table shows each Senator's Thomas ID (unique identification), the number of Senates they have served in (notice that this number divided by 3 is the number of times they have been elected), and their party affiliation.