Mapping Political Communities: A Statistical Analysis of Lobbying Networks in Legislative Politics

In Song Kim†  Dmitriy Kunisky‡

First Draft: August 25, 2017
This Draft: March 12, 2019

Abstract

Relationships between special interest groups and politicians play a significant role in policymaking. Yet empirical studies of interest group politics have been limited by the difficulty of observing these ties directly. We address this by constructing an original dataset that combines the politicians who sponsor congressional bills with the interest groups that lobby on those bills. We then develop a methodological framework to examine the political networks these data describe. Unlike networks in electoral politics, whose structure has been found to reflect ideology, we find distinct political communities in the lobbying network that are organized according to industry interests and jurisdictional committee memberships. Furthermore, we observe that important actors belong to multiple political communities, capturing their simultaneous commitments to diverse political interests. Our findings provide evidence for the existence of powerful networks in U.S. legislative politics that are distinct from electoral networks and that exhibit numerous unique structural features.

Keywords: Network analysis, lobbying, ideal point estimation, scaling, stochastic block model, link community model, community detection

Word Count: 12,256 (abstract: 149)
1 Introduction

Special interest groups engage in lobbying to promote their political objectives (e.g., Wright 1996; Grossman and Helpman 2001). The nature of the connection between politicians and interest groups, however, is much debated by political scientists. One popular view holds that interest groups take part in such costly political activity to retain access to policymakers, and in return policymakers gain an effective means of informing their legislative decisions (Bauer, Dexter, and Poll 1972; Potters and Van Winden 1992; Austen-Smith and Wright 1992; Wright 1996). Others consider whether policymakers are compensated more directly through the mechanism of campaign contributions (Austen-Smith 1995; Ansolabehere, Snyder, and Tripathi 2002). Yet another theory is that lobbying serves as a “legislative subsidy” through which interest groups help allied politicians pursue common objectives (Hall and Deardorff 2006). Though it is important for both politics and economics to assess these possibilities (Khwaja and Mian 2005; Faccio, Masulis, and McConnell 2006; Faccio 2006; Kang and You 2017), empirical studies of interest group politics have been limited by the difficulty of directly observing the ties between interest groups and politicians, let alone the actual policy context in which they are formed.

Our first contribution in this paper is to construct a large political network database tracing the universe of lobbying activities related to 108,086 congressional bills introduced between the 106th and the 114th Congress, including not only politicians and interest groups but also lobbying firms, lobbyists, government organizations, and foreign entities. We then identify a type of political connection between members of the U.S. Congress and the interest groups that engage in lobbying, by combining two types of political behavior related to each bill: (1) sponsorship by a politician, and (2) lobbying by interest groups. Although lobbying on a single bill does not necessarily imply political ties to its sponsor, recurring instances of lobbying that involve the same interest group and sponsor across several bills do reliably indicate close political relationships. Therefore, analyzing how often various politicians and interest groups interact in this way lets us infer the structure of political networks underlying the legislative process.

Our second contribution is to develop statistical network models to infer the unobserved underlying factors that might drive political connections between interest groups and politicians. We begin by introducing a Latent Space Network Model (LSNM). Like in Item-Response Theory models, we estimate for each political actor a spatial location summarizing their legislative behavior (Section 3). This approach has been widely used to study various legislative bodies, including state

---

1 We use the term special interest group to refer to any political actor who has a specific policy objective, including firms, trade associations, labor unions, business associations, and professional associations. See https://lobbyingdisclosure.house.gov/amended_ida_guide.html for the definition of lobbying.

2 There is ample empirical evidence that lobbyists help to draft bills or even write entire bills on behalf of legislators with whom they have political connections (Nourse and Schacter 2002; Hertel-Fernandez 2014).
legislatures (Shor and McCarty, 2011), the U.S. Congress (Fowler, 2006), and the United Nations (Bailey, Strezhnev, and Voeten, 2017). Unlike existing studies that rely on roll call votes for the few bills that are voted on the floor (Poole and Rosenthal, 2011; Clinton, Jackman, and Rivers, 2004), however, we analyze all Congressional bills in order to infer the latent structure at the time of their introduction.

Our model locates legislators and interest groups in a common space, in which proximity implies a closer alignment of interests. We find that, in this space, there is clustering of interest groups and legislators according to their industry affiliations and memberships in committees with jurisdiction over those industries, respectively. This finding differs from those based on similar models that analyze other political contexts, such as campaign contributions and social media (Bonica, 2013; Barberá, 2014; Bond and Messing, 2015), where the revealed underlying preferences of politicians tend to align strongly with measures of ideology, such as DW-NOMINATE scores (Poole and Rosenthal, 2011). Instead, our finding is consistent with the thesis of Ansolabehere, Snyder, and Tripathi (2002), who argue that interest groups that act mainly through lobbying are more bipartisan and less ideological than those that act mainly through campaign contributions. It also shows that interest groups make use of the power that committees have over policy outcomes during the legislative process, as described by Shepsle and Weingast (1987) and Fourranaies and Hall (2017). In short, we observe that interest groups target legislators with influence on relevant committees in interactions localized to specific policy domains, forming communities in the lobbying network.

We then propose a new methodological framework that incorporates the community structure we identify in the spatial model (Section 4). We begin by applying the stochastic block model, a sociological tool (Snijders and Nowicki, 1997) more recently studied in the physics and machine learning literature (Fortunato, 2010). We verify that the communities identified by the stochastic block model largely agree with those identified ex post in the latent space model, confirming that the lobbying network exhibits significant community structure. However, we also observe an important and qualitatively different pattern in the lobbying network, namely that many interest groups and politicians tend to represent varied interests across several political communities.

Motivated by this empirical regularity, we develop a new statistical model, the Bipartite Link Community Model (biLCM), that allows actors not only to belong to multiple political communities (so-called “mixed membership”), but also to interact with a given partner simultaneously in multiple policy domains. The proposed model finds that interest groups such as the CHAMBER OF COMMERCE that represent the heterogeneous interests of many members engage in aggregate lobbying by targeting politicians who are members of broad “power committees” such as the HOUSE COMMITTEE ON WAYS AND MEANS and the SENATE COMMITTEE ON APPROPRIATIONS.
This is consistent with Wright (1990) who, based on numerous personal interviews and a survey of lobbyists, found that lobbying by powerful interest groups affected politicians’ voting especially strongly within the House Committee on Ways and Means. Our model also captures the focused lobbying of industry-specific organizations and firms that we observed previously with the LSNM. In this regard, the proposed methodology improves our understanding of the organization of lobbying in the U.S. Congress, accurately capturing the full spectrum of behaviors from very targeted to very broad lobbying. The proposed methods for network analysis will be implemented in the open-source software package “polnet: Statistical Analysis of Political Networks.”

Taken together, our results illustrate a nuanced community structure in the lobbying network. First, we find that the key to understanding interactions in the lobbying network is actors’ overlapping topical interests rather than their shared ideology. The community models we propose are also well-suited to identifying both direct and indirect political connections (both “friends” and “friends of friends”), the latter of which have been notoriously difficult to examine in interest group politics previously. Our models accomplish this task by inferring actors’ shared community memberships even when researchers do not observe their direct interactions.[3] Second, our results call for more careful consideration of the ways in which interest groups’ policy preferences are expressed through committee action (Shepsle and Weingast 1987; Schickler 2001; Sinclair 2016). In particular, these mechanisms cannot be understood merely by analogy with public-facing politics: our results show that the interactions of legislators with interest groups are different from their position-taking interactions with the public at large, which are driven mainly by electoral motivations (Mayhew 1974). Finally, based on mixed membership modeling, we find that “specialist” and “generalist” interest groups, respectively representing narrow and aggregate political interests, coexist in U.S. legislative politics.

Our analyses also provide technical guidance for further research on political networks. To the best of our knowledge, ours is the first statistical study of lobbying networks in legislative politics that examines both politicians and interest groups using micro data. More broadly, the proposed community detection approach is a novel empirical framework for applied political scientists to systematically identify political communities and characterize their members, which will help leverage the rich network data that have become increasingly available in various fields of political science (Keck and Sikkink 1998; Hoff, Raftery, and Handcock 2002; Maoz et al. 2006; Hafner-Burton, Kahler, and Montgomery 2009; Clark and Lauderdale 2010; Ward, Stovel, and Sacks 2011).

The network data, all pairwise measurements of political interactions between politicians and

---
[3] For instance, if Interest Group A frequently lobbies on bills that Politician 1 and Politician 2 sponsor, while Interest Group B lobbies on bills that Politician 2 and Politician 3 sponsor, then our models would find it likely that Interest Group A belongs to the same community as Politician 3, though they do not interact directly.
interest groups, the estimated latent spatial positions and their posterior distributions, the estimated community memberships of political actors, and the visualization tools used in preparing this paper will also be made publicly available at [https://www.lobbyview.org](https://www.lobbyview.org).

2 The Lobbying Network Database

The Lobbying Disclosure Act of 1995 (later amended by the Honest Leadership and Open Government Act of 2007) requires mandatory quarterly electronic filing for any organization (“client”) that “actively participates” in lobbying. Filings must disclose the lobbying issue area, the congressional and federal agencies contacted, and the specific issues on which lobbyists (“registrants”) have engaged in political activities. Based on the information available from this disclosure of political activities, researchers have studied various aspects of the politics of lobbying, including the sources of and returns on decisions to lobby [Richter, Samphantharak, and Timmons 2009; de Figueiredo and Richter 2014; Kang 2015; You 2017], the similarities and differences between campaign contributions and lobbying [Ansolabehere, Snyder, and Tripathi 2002], characteristics of lobbyists [Baumgartner et al. 2009; Vidal, Draca, and Fons-Rosen 2012; Bertrand, Bombardini, and Trebbi 2014], and the implications of lobbying for trade politics [Bombardini and Trebbi 2012; Kim 2017]. However, the available lobbying data has the important limitation that interest groups are not required to report the identities of their political contacts. This is unfortunate, given that almost 90% of lobbying reports indicate that at least one member of Congress or a member of their staff was contacted. In order to study the connections between interest groups and politicians, there is therefore no choice but to infer those unobserved connections statistically from the available data.

To that end, we construct an original database of the congressional lobbying network that captures the connections among lobbyists, interest groups, bills, and politicians. The database is generated from lobbying reports filed between 1999 and 2017, which are available from the Senate Office of Public Records. A key component of the database is a suite of automated systems to (1) detect congressional bills reported to have been lobbied, (2) identify the session of Congress those bills occurred in, and (3) relate each bill to its sponsor. Below, we give a brief outline of the process used to construct the lobbying database.

First, we identify all lobbying activities related to congressional bills, which is possible due to 2 U.S.C. S. 1604(b)(2)(A), a law requiring interest groups to disclose the list of bill numbers on which they have lobbied. For example, a report filed by Intel Corporation in 2013 mentions 4 Compliance with disclosure requirements is closely monitored, strictly enforced, and annually audited by the Government Accountability Office (GAO). According to the 2014 GAO audit report, available from [http://www.gao.gov/products/GAO-15-310](http://www.gao.gov/products/GAO-15-310) 90% of organizations filed reports as required, and 93% could provide documentation related to expenses.
Figure 1: Descriptive Statistics of Lobbying on Congressional Bills. The left panel compares the numbers of bills introduced, lobbied, and voted on in each congressional session from the 106th through the 114th Congress. The number of bills that are identified as lobbied increases significantly beginning in the 110th Congress, when lobbying reports became digitized per the Honest Leadership and Open Government Act of 2007. The right panel shows the distribution of the number of interest groups that lobbied on each bill introduced in the 113th Congress. The distribution is highly skewed: the median is three, the maximum is 946, and 22.9% of the lobbied bills are lobbied by only one interest group.

H.R. 289, “National STEM Education Tax Incentive or Teachers Act of 2011—Science and Math Education Legislation.” In practice, bills are often referenced within the text of a lobbying report, and sometimes only by alternative names or subtitles, so the entire report text must be processed and all bill references algorithmically identified. Second, we identify the session of Congress that each bill belongs to, which the report often does not mention explicitly, as in the above example. We use text data mining techniques to determine whether information such as the bill title or phrases like “Act of [YEAR]” occur in the report text. Based on our algorithm, described in greater detail in Appendix A.1.1 we correctly determine that, for instance, the above bill, H.R. 289, is from the 112th Congress. If the year that the lobbying report was filed had been used to determine the Congress instead, then this bill would have been identified as H.R. 289 “Value Our Time Elections Act” from the 113th Congress, since the report was filed in 2013. Finally, we identify the sponsor of each bill, completing the connection between interest groups and politicians. We repeat this process for each of 1,111,859 lobbying reports, which in total link 20,092 special interest groups with 1,164 (current and former) members of Congress.

The left panel of Figure 1 shows that about 12,000 bills are introduced in each Congress and the majority of them are lobbied by at least one interest group. Very few of these bills make it to a vote as written, however. Instead, they tend to be merged into larger bills that are voted

---

5 In fact, the OpenSecrets.org database from the Center for Responsive Politics, which is often used in academic research, erroneously finds that Intel lobbied on the latter. See http://www.opensecrets.org/lobby/billsum.php?id=hr289-113. Such examples call into question the validity of previous findings based on these data.
on, following a complex process of amendment and political compromise. Our dataset allows us to observe political connections before this noisy process takes place at a finer level of granularity. Indeed, as the right panel shows, most bills are lobbied by very few interest groups, implying that each individual instance of lobbying tends to reflect narrow interests. A typical example: on “A bill to exempt the aging process of distilled spirits from the production period for purposes of capitalization of interest costs” (113th S. 1457), sponsored by Senator Mitch McConnell (R-KY), the DISTILLED SPIRITS COUNCIL OF THE U.S. was the only interest group to lobby.

Both lobbying and bill sponsorship are relatively rare activities: most interest groups lobby on only a small number of bills and most members of Congress introduce only a small number of bills, while a very small number of actors from both categories are highly active. Not surprisingly, then, most politician-interest group pairs experience no lobbying interactions on sponsored bills while a small number interact frequently. For instance, 97.96% of pairs of actors in the 113th Congress do not interact on any bills, while the pair with the greatest number of interactions is Senator John “Jay” Rockefeller (D-WV) and the NATIONAL CABLE TELECOMMUNICATIONS ASSOCIATION, which interacted on 26 bills.

We will later see that a group of firms and associations in the telecommunications industry in fact forms one of the significant clusters in our models. More details on our datasets are given in Appendix A.1.

**Notations** Before describing our models, we introduce some mathematical notation and shorthand terminology that will be useful. We index a set of interest groups by \( m = \{1, \ldots, m\} \) and a set of politicians by \( n \). When a bill is reported as lobbied on by interest group \( i \) and sponsored by politician \( j \), we say that the interest group and politician interact on that bill (interacting pairs will also sometimes be called “partners”), and each lobbying report interest group \( i \) writes that mentions the bill is a political interaction. We denote the number of interactions (over some period of time, usually a single session of Congress) between the pair \( i, j \) by \( A_{i,j} \), and organize these numbers as the entries of the interaction matrix \( A \in \mathbb{R}^{m \times n} \). This matrix may be viewed as the adjacency matrix of a bipartite graph \( G \) with weighted edges, where the politicians and interest groups lie on opposite sides of the partition. The degree (a term from network theory) of an interest group or politician is the total number of interactions they have with every actor of the other type. Therefore, high-degree politicians are those whose bills are often lobbied, and high-degree interest groups are those who lobby many bills.

---

6See Kim (2017) for examples of similar patterns in trade bills.
7See Figure A.3 in the Appendix for more precise statistics.
8Senator Rockefeller’s close relationship with the telecommunications industry is well-known; see e.g. [https://www.nytimes.com/2007/10/23/washington/23nsa.html](https://www.nytimes.com/2007/10/23/washington/23nsa.html).
9We will use this terminology throughout, though the Lobbying Disclosure Act calls the organizations that file lobbying reports “lobbying clients.”
3 Latent Space Modeling of the Lobbying Network

In this section, we develop a Bayesian Latent Space Network Model (LSNM). Following the introduction of this class of model by Hoff, Raftery, and Handcock (2002), there has been a proliferation of variants, of which we mention just a few that are useful for understanding our methods. Our model is formally similar to the “Wordfish” model of Slapin and Proksch (2008) for text analysis, which estimates ideal points of a single group of political agents based on text data, whereas we estimate the underlying preference structure of two distinct groups of interacting political actors. Our model is also closely related to that of Barberá (2014), which relates ideal points of politicians and social media users, but instead of modeling whether a binary political interaction exists (e.g., Twitter follows), we consider how frequently political contacts occur. Thus our interaction measurements have an associated magnitude.

The Model We model the connection between interest group $i$ and legislator $j$ as a function of the two actors’ latent spatial positions in a $d$-dimensional Euclidean space, $\theta_i \in \mathbb{R}^d$ and $\psi_j \in \mathbb{R}^d$ respectively, and assume that the number of their interactions $A_{i,j}$ has a Poisson distribution, with its mean parametrized by the Euclidean distance $\| \theta_i - \psi_j \|_2$.

To account for the differences in agents’ baseline propensities to sponsor or lobby (see Figure A.3), we also include interest group- and legislator-specific “popularity” terms, $\alpha_i$ and $\beta_j$ respectively, in the mean of the modeled $A_{i,j}$.

We then take the probabilistic model

$$A_{i,j} \sim \text{Poisson}(\exp(\alpha_i + \beta_j - \| \theta_i - \psi_j \|_2^2))$$ (1)

for the interaction matrix entries. To implement inference, it is more convenient to reparametrize in terms of $\tilde{\alpha}_i = \alpha_i - \| \theta_i \|_2^2$ and $\tilde{\beta}_j = \beta_j - \| \psi_j \|_2^2$, which yields the equivalent model

$$A_{i,j} \sim \text{Poisson} \left( \exp \left( \tilde{\alpha}_i + \tilde{\beta}_j + 2 \theta_i^T \psi_j \right) \right).$$ (2)

We then model the entries as independent conditional on the parameters, to obtain the following joint distribution:

$$\mathbb{P}(A | \alpha, \beta, \theta, \psi) = \prod_{i=1}^{m} \prod_{j=1}^{n} \text{Poisson} \left( A_{i,j} | \exp \left( \tilde{\alpha}_i + \tilde{\beta}_j + 2 \theta_i^T \psi_j \right) \right).$$ (3)

To impose parameter identification and improve the numerical behavior of Markov Chain Monte

---

10 We do not impose an ex ante interpretation of the latent spatial locations; similar analyses of political actors other than interest groups sometimes call the locations “ideal points” or “preferred policies.”

11 In the literature, popularity terms are also called “gregariousness” or “idiosyncratic factors.” A similar construction to ours was tested on synthetic datasets by Krivitsky et al. (2009).
Figure 2: **DW-NOMINATE Ideology vs. LSNM.** We illustrate the lack of correlation between politicians’ latent space positions obtained from one- and two-dimensional LSNM estimates and the DW-NOMINATE ideology dimension. It shows that the ideological structures common in electoral politics are not prevalent in the lobbying network. Points are colored on a linear scale from blue to red according to the DW-NOMINATE dimension as well, to introduce a convention we will use throughout the paper.

Computation We perform inference for our model with the Hamiltonian Monte Carlo No-U-Turn Sampler and Automatic Differentiation Variational Inference implementations in the Stan software package [Carpenter et al., 2017]. In the results presented here, we use empirical posterior means of our parameters taken as point estimates, and we perform further analysis of these point estimates so as to focus on politically meaningful findings. We include the Stan code describing our model and provide a more technical evaluation of our implementation, discussing how parameter identification is achieved, the choice of dimensionality, the concentration of the posterior distribution, and MCMC sampling diagnostics, in Appendix A.2.
Empirical Findings  We apply the proposed model to the lobbying network dataset for the 113th Congress. We first consider a one-dimensional LSNM, the simplest model our framework admits. To compare with the situation in electoral politics, we compare estimated latent spatial positions of politicians ($\psi_j \in \mathbb{R}$) with DW-NOMINATE ideology measures. As the left panel of Figure 2 shows, we do not find evidence that shared ideology drives the interaction between interest groups and politicians (Pearson’s $\rho = 0.15$ between our latent dimension and the DW-NOMINATE ideology dimension). This lack of correlation also holds for restrictions to only Democratic or Republican members of Congress. Our finding is consistent when we increase the dimensionality of the latent positions as well, as the right panel shows.

Increasing the dimensionality of the LSNM facilitates the interpretation of the estimated latent space positions. For the two-dimensional LSNM, Figure 3 presents the posterior means of the estimated latent spatial positions, $\theta_i$ and $\psi_j$, for all interest groups $i$ and politicians $j$. Each interest group is represented by a black circle and each politician is represented by a colored circle according to their DW-NOMINATE ideology dimension. The size of each circle is proportional to the exponential of their popularity factor ($\alpha_i$ and $\beta_j$), so that the mean of the Poisson interaction between two agents at fixed distance is proportional to the product of their circles’ sizes.

Figure 3 shows that the key structure of the lobbying network in this two-dimensional model revolves around specific issue areas and industries. As the shaded regions indicate, interest groups are clustered in the latent space according to their industry affiliation (such as telecommunications) or issue area of interest (such as veterans’ affairs). Importantly, these clusters emerge even though the interest groups involved usually do not lobby on the same bills (recall that most bills are lobbied by only a few groups). Instead, related groups are “pulled together” through their connections to the same politicians, who sponsor separate bills on which the groups separately lobby. As in the one-dimensional model, there is no clear partisan divide either across or within industry clusters. Rather, we find in each large cluster of interest groups a concentration of politicians from committees involved with the policy issues that affect those groups. For instance, Figure 4 shows that the legislators belonging to the “Finance & Insurance” cluster are likely to serve in the HOUSE FINANCIAL SERVICES COMMITTEE (31 out of 62), and those belonging to the “Veterans’ Affairs” cluster are likely to serve in the HOUSE COMMITTEE ON VETERANS’ AFFAIRS (10 out of 15).

To formally investigate the significance of industry and committee affiliations in determining an actor’s latent position in the lobbying network, we conduct a series of statistical tests. Specifically, we compare the performance of “Restricted” and “Full” linear models predicting latent positions

---

12 For more technical analysis of choice of dimensionality for the LSNM based on heuristics from spectral methods, the reader may consult Appendix A.2.2.

13 We analyze the posterior distributions further in Appendix A.2 with particular attention to the uncertainties of the point estimates given by the posterior means to illustrate the stability of these results.
Figure 3: **Estimated LSNM Positions: Full 113th Congress.** This figure presents the two-dimensional latent space positions and popularity factors inferred from the 113th Congress dataset. We indicate several significant clusters corresponding to specific industries and issue areas. Interest groups are represented with black dots while politicians are represented with colored dots according to their ideological score on the DW-NOMINATE scale. The size of each dot is proportional to the exponential of the agent’s popularity factor ($\exp(\alpha_i)$ or $\exp(\beta_j)$). We annotate the clusters with some representative members.

---

*One actor with outlying mean latent positions (right of the region shown), DIRECTV, INC., is omitted for the sake of visual clarity.*
Figure 4: **Committee Membership of Politicians by Cluster.** Histograms of the top 10 committee memberships for politicians in the larger clusters highlighted in Figure 3. Bars are divided by political party (red for Republican and blue for Democrat), and the horizontal bars at the top show the overall party distribution of each cluster. Senate committee and House committee labels are prefixed with (S) and (H), respectively.

by nesting the former in the latter. Table 1 presents the $F$-statistics and $p$-values summarizing the significance of obviating various sets of observable covariates from the “Full” model. We find that committee memberships are the only significantly informative covariates of politicians’ positions in both the one- and two-dimensional models. In contrast, ideology and geography play insignificant roles. Similarly, we find that industry affiliation, determined by the leading digits of an interest group’s NAICS industry code, is always significant in predicting the latent space position of an interest group, while the interest group’s home state only affects one dimension of the two-dimensional LSNM significantly. Statistics for the latter findings are included in the Appendix in Table A.1.

Not all regions in our spatial representation, however, are characterized by domain-specific political interests. We find that a region in the center of the latent space (marked by the dashed oval in Figure 3) is populated by politicians who are lobbied by groups with diverse interests. Furthermore, these politicians tend to sit not on committees that are directly related to the industries of the relevant groups, but rather on so-called “power committees” that concern government oversight, federal spending, and taxation, most prominently the House Committee on Ways and Means, the House Committee on Appropriations, and the Senate Committee on Appropriations, as shown in Figure 4. Several of the interest groups that are the most active in

---

14We thus find evidence that one dimension of the two-dimensional LSNM can partly be explained by geography. We believe that this is a result of politicians in Congress tending to support the interests of groups belonging to their own state [Bailey and Brady, 1998]. To examine such effects more formally, recent research has sought to incorporate observed covariates into network models directly (e.g., Minhas, Hoff, and Ward [2016]). We leave the challenging task of adapting such methods to our setting for future research.
Table 1: Regression Analysis of LSNM Covariates. Statistics of linear models of numerous covariates against the one- and two-dimensional LSNM latent dimensions for politicians. We compare full linear models with all available covariates against restricted models, omitting either committee-related covariates, geographic information, or ideology measures. In the bottom three rows, for each of the three latent dimensions, we give the $F$ statistics and $p$-values of the comparison $F$-test. The test illustrates that the committee membership covariates are the only covariates significantly explanatory of the LSNM latent space organization.

lobbying, such as the Chamber of Commerce (which lobbied more bills in the 113th Congress than any other group), the Specialty Equipment Market Association (SEMA), and the American Civil Liberties Union (ACLU), lie in this region. The region also includes many senior politicians and party leaders, such as Patty Murray (D-WA), Barbara Boxer (D-CA), Kirsten Gillibrand (D-NY), Orrin Hatch (R-UT), Lindsey Graham (R-SC), and Paul Ryan (R-WI). We defer further interpretation of the characteristics they have in common to Section 4.2, where we will introduce mixed-membership community models in part to answer this question.

The analysis presented in this section demonstrates the presence of political communities in the lobbying network, which are organized according to industry interests and jurisdictional committee memberships. This description contrasts with networks of electoral politics, which scholars have found are dominated by ideological factors. Although the identification of clusters and their visual demarcations might be seen as arbitrary, the spatial configuration of estimated latent positions exhibits substantively meaningful patterns. In fact, the geometry of the clusters is consistent with common intuitions about relationships among industries: for example, the “Technology” cluster is adjacent to both the “Telecommunications” and “Energy” clusters, while the “Universities” cluster is adjacent to both the “Industrial Research & Defense” and “Healthcare” clusters. Having bolstered our belief that political communities exist in the lobbying network, we next seek to study them in a more explicit and statistically principled manner, adapting our modeling techniques to better suit community structure.
4 Community Modeling of the Lobbying Network

In this section, we first introduce the canonical model of probabilistic community detection in bipartite networks—the Bipartite Stochastic Block Model (biSBM) of Larremore, Clauset, and Jacobs (2014)—to clarify the interpretation of the network structure that we identified in Section 3. Unlike the LSNM, the biSBM explicitly models the interaction between a politician and an interest group as a function of their community memberships, and allows us to verify that politicians and interest groups with shared group memberships interact more frequently. Although this model confirms the existence of communities in the lobbying network, it in turn provides us insights about the limitations of assigning a single community membership to each actor. To address this, we propose the Bipartite Link Community Model (biLCM) in Section 4.2. The proposed model allows for multiple community memberships as well as heterogeneous connection types, which are prevalent in most political networks. We derive an expectation-maximization (EM) algorithm to estimate the biLCM parameters, which characterize the mixed community memberships of interest groups and politicians in the 113th Congress. Our analysis illustrates a clear distinction between “specialist” and “generalist” interest groups and identifies the politicians with whom each kind is strongly connected.

4.1 Motivation: The Bipartite Stochastic Block Model

The Model  Following the formulation of the LSNM in Section 3, we model $A_{i,j}$ as having a Poisson distribution, whose mean now depends exclusively on the community memberships of interest group $i$ and politician $j$. Formally, we introduce a matrix parameter $B \in \mathbb{R}^{k \times \ell}$ with entries $B_{r,s}$, where $k$ is the number of interest group communities and $\ell$ the number of politician communities, viewed as hyperparameters. To fix the dimensions, we draw upon our exploratory analysis with the LSNM, which suggests roughly how many distinct communities we expect. Based on the analysis in Figure 3, we use eight politician and interest group communities. The membership parameters are denoted by $x_i \in [k]$ and $y_j \in [\ell]$ for each $i \in [m]$ and $j \in [n]$. As before, we include popularity factors $\alpha_i$ and $\beta_j$ for each interest group and politician, respectively. The entries $A_{i,j}$ are then modeled as

$$A_{i,j} \sim \text{Poisson}(\alpha_i \beta_j B_{x_i,y_j})$$

Independently, so that the joint probability is

$$P(A \mid B, x, y, \alpha, \beta) = \prod_{i=1}^{m} \prod_{j=1}^{n} \text{Poisson}(A_{i,j} \mid \alpha_i \beta_j B_{x_i,y_j}).$$

Our substantive findings are robust to adjustments of the number of communities in the community models we work with. Rigorous model selection for this parameter is an active area of research in block modeling, so we leave such quantitative justifications for future research.
In the block modeling literature, the inclusion of popularity factors is known as degree correction, because these variables account for the possibility of broad degree distribution in a network. The model described by equation (5) is therefore named the Degree-Corrected Bipartite Stochastic Block Model (dc-biSBM).\(^{16}\)

We note that, unlike our analysis of the LSNM, the dc-biSBM does not a priori assume that the community structure underlying the lobbying network is assortative. That is, we allow for the possibility that interest groups and politicians do not form shared communities within which they prefer to interact, but rather form groups that characterize their interactions in other ways. Nonetheless, we find that the former situation indeed obtains, and interest groups and legislators interact strongly only when they share a community membership. We support this claim quantitatively in Appendix A.3.2.\(^{17}\)

**Computation** We perform maximum likelihood parameter estimation for the dc-biSBM using the expectation-maximization algorithm implementation provided with the work of Larremore, Clauset, and Jacobs (2014). Since this algorithm is not guaranteed to converge to the parameters maximizing the likelihood, we take the parameters attaining the highest likelihood value of 50 randomly initialized runs.

**Empirical Findings** In the LSNM, we interpreted the proximity of actors’ latent positions as an indicator of their shared community membership. To validate this arbitrary interpretation, we begin by comparing the LSNM results from Section 3 with the community memberships found by the dc-biSBM.

Figure 5 presents strong evidence for the existence of community structure in the lobbying network. It overlays the finite cluster memberships (marked by different colors and shapes) identified by the dc-biSBM on the geometry obtained from the LSNM. To facilitate a direct comparison, we also include the grey boundaries previously used to classify interest groups and politicians. The figure shows that the LSNM and dc-biSBM community structures align remarkably well with each other. Specifically, the boundaries that we drew in the LSNM indeed correspond to distinct political communities or to mixtures of at most two or three political communities identified by the dc-biSBM. For example, we find that all of the interest groups and legislators in the LSNM’s “Veterans’ Affairs” community belong to a single dc-biSBM community (orange plus signs). Likewise, in the “Energy” cluster, 84% of the interest groups and 86% of the legislators in the LSNM community belong to a single dc-biSBM community (pink diamonds). Furthermore, two pairs of adjacent

---

\(^{16}\)We discuss the advantages of dc-biSBM over the Bipartite Stochastic Block Model (biSBM) without degree correction for analyzing the lobbying network in Appendices A.3.1 and A.3.\(^{17}\)

\(^{17}\)Thus we combine the dc-biSBM groupings of interest groups and legislators into eight joint communities when we present our findings in this section.
LSNM communities are merged into single dc-biSBM communities: “Universities” and “Industrial Research & Defense” (grey circles), and “Technology” and “Telecommunications” (green squares), which supports our earlier interpretation of adjacent LSNM communities as representing closely related topical interests.

However, the dc-biSBM provides further insights about community structures that were not legible in the LSNM. First, the dc-biSBM uncovers two distinct communities that consist of seemingly heterogenous members near the center of the latent space. The first, which we name “Civil Society,” includes major general civil rights associations (AMERICAN CIVIL LIBERTIES UNION), as well as lobby organizations for the rights of workers (AFL-CIO), senior citizens (AARP), women (NATIONAL WOMEN’S LAW CENTER), the disabled (EASTER SEALS), gender and sexual minorities (HUMAN RIGHTS CAMPAIGN), and ethnic minorities (AMERICAN JEWISH COMMITTEE). The second, which we name “Retail & Transportation,” includes agricultural groups (NATIONAL CORN GROWERS), manufacturing associations (NATIONAL ASSOCIATION OF MANUFACTURERS), automotive firms (NISSAN), retailers (HOME DEPOT), fuel manufacturers (BRITISH PETROLEUM), and environmental conservation groups (BLUE GREEN ALLIANCE).
Figure 6: Overlapping Communities in LSNM and dc-biSBM. We plot regions where pairs of communities discovered by the dc-biSBM overlap in their latent positions as discovered by the LSNM. These lie near the central region highlighted in Figure 3 and Figure 5 with a dashed oval, which we repeat in each plot for reference. For the sake of visual clarity, in each plot we include only the actors from the relevant pair of communities. We also highlight several illustrative examples of interest groups in each plot and name them below. (Markers and colors are the same as those used in Figure 5.)

Second, the dc-biSBM clarifies a key structural feature of the lobbying network: many significant political actors, previously found in the center region of the LSNM, have less specific industry affiliation and represent more complex collections of interlocking political interests than the single-industry actors discussed previously. Figure 6 illustrates this point. Specifically, it zooms in on different parts of the center region of the LSNM, focusing on three pairs of adjacent communities. We find that these communities overlap, and that actors in their intersections tend to lobby separately on bills related to the subject matters of either community. For instance, McAfee, Inc. (in the left panel), a security software firm (and subsidiary of INTEL CORPORATION), lobbied on 113th S. 1429 “Department of Defense Appropriations Act” together with defense industry firms such as GENERAL DYNAMICS and NORTHROP GRUMMAN, but also lobbied on 113th H.R. 756 “Cybersecurity Enhancement Act” together with technology firms including MICROSOFT, INC. and GOOGLE, INC. Accordingly, McAfee, Inc. in the LSNM lies near the intersection of the “Technology & Telecommunications” and “Universities & Research” communities. Similarly, BRITISH PETROLEUM (in the center panel), an oil energy firm, lobbied on 113th H.R. 1191 “Keep American Natural Gas Here Act” with utilities companies including NISOURCE, INC. and DOMINION RESOURCES, but also lobbied on 113th H.R. 1462 “RFS Reform Act” (an amendment to the “Clean Air Act” (42 U.S.C. 7545(o)(1)(J)) concerning renewable fuels) with retail organizations...
like the National Retail Federation and the Snack Food Association. Again, British Petroleum lies near the intersection of the “Energy” and “Retail & Transportation” communities. Finally, the Chamber of Commerce (in the right panel) is classified into the “Civil Society” community; however, being among the most active interest groups in the lobbying network, the Chamber of Commerce lobbied on a great variety of bills, ranging from 113th S. 601 “Water Resources Development Act of 2013,” to 113th S. 462 “United States-Israel Strategic Partnership Act of 2013,” to 113th S. 809 “Genetically Engineered Food Right-to-Know Act.”

Importantly, these findings suggest that the “hard” assignment of a single community to each actor by the dc-biSBM is severely limited, because many actors do participate in multiple major communities in the lobbying network simultaneously. Indeed, consistent with the analysis from the LSNM, we find that the actors in the center region are typically large interest groups with diverse interests and senior politicians with broad authority, both of whose legislative pursuits might plausibly be driven by varied interests. The dc-biSBM thus furnishes several additional pieces of evidence for the regions of the latent space without clear community structure being characterized by varied interests. In the following section, therefore, we propose a new community model that represents the varied interests of political actors in an explicit probabilistic framework.

4.2 Proposed Methodology: The Bipartite Link Community Model

In this section, we propose the Bipartite Link Community Model (biLCM). Our model combines three features which, to the best of our knowledge, have not been previously combined in the community detection literature. First, it explicitly models mixed membership through link communities (“links” being a generic term from network theory for what we call “interactions”); second, it does so in the bipartite setting; and finally, it uses a statistically principled generative model. Our model is most similar to that of Ball, Karrer, and Newman (2011), but borrows the ideas for specialization to bipartite graphs from Larremore, Clauset, and Jacobs (2014). The most similar work we are aware of is that of Li, Zhang, and Zhang (2015), who treat the link community detection task in bipartite graphs as an optimization problem and solve it with an ad hoc genetic algorithm. Unlike that work, our model provides an underlying probabilistic model and therefore a natural statistical interpretation.

Mixed Memberships  Mixed membership in multiple communities is particularly relevant for understanding political networks. As we have already seen in Section 4.1, many interest groups either have heterogeneous interests across various policy domains or represent diverse interests of their substituents. For instance, a cybersecurity firm might be interested in both national defense policy and internet regulation, and an oil energy firm might be interested in both restrictions on utilities providers and pollution regulation. A simpler instance of the same phenomenon occurs
with large holding companies, which lobby in multiple arenas on behalf of their diverse subsidiaries. For instance, the firm **PHILIPS NORTH AMERICA**, which specializes in medical equipment, home appliances, and lighting solutions, lobbies both on bills relating to healthcare and bills related to retail and transportation. Other lobbying organizations lobby on behalf of even broader collections of members, such as the aforementioned **CHAMBER OF COMMERCE** which represents business interests at large. Similarly, most politicians are interested in several policy domains [Lauderdale and Clark, 2014] and sit on several committees that have distinct legislative jurisdictions.

**Mixed Interactions** Besides the degree to which their interests are varied, interest groups also differ in which legislators they choose to interact with in expressing their varied interests. In some cases, an interest group lobbies on different subject matter with different legislators who serve on different topical committees. For instance, continuing an earlier example, **BRITISH PETROLEUM** lobbied on 113th H.R. 1191 “Keep American Natural Gas Here Act” sponsored by Senator Edward Markey (D-MA), a member of the Senate Committee on Environment and Public Works, as well as 113th H.R. 1462 “RFS Reform Act” sponsored by Representative Bob Goodlatte (R-KY), the vice chairman of the House Committee on Agriculture (a committee intimately related with renewable fuels, the subject of the legislation).

On the other hand, an even stronger version of varied interests arises when an interest group interacts with a single politician on bills that do not belong to the same political issue area. This occurs most frequently with aggregate interest groups or holding companies and senior politicians who tend to have seats on power committees. For example, as mentioned earlier, the **CHAMBER OF COMMERCE** lobbied on the three diverse bills 113th S. 601 “Water Resources Development Act of 2013,” 113th S. 462 “United States-Israel Strategic Partnership Act of 2013,” and 113th S. 809 “Genetically Engineered Food Right-to-Know Act.” In fact, all three of these bills were sponsored by the same legislator, Senator Barbara Boxer (D-CA). Similarly, the **SPECIALTY EQUIPMENT MARKET ASSOCIATION** lobbied on 113th S. 983 “Keep the IRS Off Your Health Care Act of 2013,” 113th S. 2635 “21st Century Endangered Species Transparency Act,” and 113th S. 725 “Small Business Taxpayer Bill of Rights Act of 2013.” Again, all three were sponsored by a single legislator, Senator John Cornyn (R-TX).

**Modeling Choices** These observations guide our choice of a model allowing political actors to have simultaneous memberships in several legislative communities. One model that incorporates this assumption is the **mixed-membership stochastic block model (mmSBM)**, in which each actor has a probability distribution over all communities [Airoldi et al., 2008]. For example, Airoldi et al. [2008] apply the mmSBM to the Sampson monastery dataset [Sampson, 1969], a popular example in the social networks literature. This dataset summarizes survey data on social relationships.
Figure 7: Schematic Comparison of Community Models. We illustrate three generative network models: the single-membership dc-biSBM model, the mixed-membership mmSBM model, and the mixed-membership biLCM model. For the sake of simplicity, the first two models are illustrated in the assortative case, where each politician community interacts strongly with just one interest group community and vice-versa. Colors indicate community memberships and, in the mixed-membership cases, interaction types. (a) In the single-membership model, political actors within a community interact more frequently. (b) The mmSBM allows for mixed membership distributions while all interactions involving a given partner are constrained to be of a single type. (c) The biLCM allows for mixed membership as well as mixed interaction types. This captures the common lobbying interactions in which interest groups and politicians with diverse community memberships interact for different political reasons.

among 18 novice monks joining a monastery, and the model identifies a group of “waverer” monks who do not commit to one of the primary social factions in the monastery but rather maintain mixed friendships with members of several other factions.

Although this model has some desirable features, the mmSBM has the property that an actor varies their community membership across their interactions with different partners but not with a specific single partner. We illustrate this important conceptual difference in Panels (b) and (c) of Figure 7. Panel (b) shows that a model in the mmSBM family would bizarrely insist that a single common community membership must account for each pair of interactions (i.e., “links”) involving a given politician and interest group. However, as we discussed before, the mixed membership we are interested in modeling is different: we are concerned with political actors whose community membership varies even across interactions with a single partner over the course of a session of Congress. Indeed, the bills that link a given politician and interest group often have different underlying motivations and subject matters, and thus should be modeled as such (Of course, as Panel (a) of Figure 7 shows, both models are far more flexible than the dc-biSBM, which makes

\[18\] We consider an extended version of the original mmSBM with assortative and weighted relationships that are closer to the data we observe in the lobbying network. Mathematically, the difference we are emphasizing is that an mmSBM adjusted to have Poisson-distributed weights on interactions would model each edge weight as a mixture of Poisson distributions, while the link community model we propose will model each edge weight as a sum of independent Poisson distributions, one for each link community in which the agents interact.
the assumption that only a single fixed community membership is allowed for each actor).

The Model  We suppose that there are \( k \) link communities, which we will always take as indexed by a variable \( z \in [k] \). To emphasize the political interpretation, we will subsequently refer to these as *legislation communities*. Each interest group and politician has a vector of parameters \( \alpha_{i,z} \) and \( \beta_{j,z} \), respectively, which represents their involvement in legislation belonging to community \( z \). The number of bills lobbied by interest group \( i \), sponsored by politician \( j \), and belonging to legislation community \( z \) is modeled as Poisson with a mean proportional to \( \alpha_{i,z} \beta_{j,z} \). To resolve the identification issue, we assume that for each fixed \( z \), \( \sum_{i=1}^{m} \alpha_{i,z} = \sum_{j=1}^{n} \beta_{j,z} = 1 \), and we introduce another parameter \( \kappa_{z} \) to capture the overall weight of group \( z \), so that the number of bills between interest group \( i \) and politician \( j \) in legislation community \( z \) has mean \( \kappa_{z} \alpha_{i,z} \beta_{j,z} \). We assume that these Poisson variables are independent, so that the model for the interaction matrix entries is given by

\[
A_{i,j} \sim \text{Poisson} \left( \sum_{z \in [k]} \kappa_{z} \alpha_{i,z} \beta_{j,z} \right),
\]

and their joint distribution is

\[
P(A \mid \alpha, \beta, \kappa) = \prod_{i=1}^{m} \prod_{j=1}^{n} \text{Poisson} \left( A_{i,j} \bigg| \sum_{z=1}^{k} \kappa_{z} \alpha_{i,z} \beta_{j,z} \right).
\]

Computation  We derive an expectation-maximization algorithm for this model in Appendix A.4.1, which involves alternating expectation and maximization update steps until the log-likelihood of the model converges. The update equations produced by our derivation are given below, including ancillary optimization parameters \( q_{i,j}(1), \ldots, q_{i,j}(k) \) (the first equation is the expectation step for the ancillary parameters, and the last three equations are maximization steps for the model parameters).

\[
q_{i,j}(z) = \frac{\kappa_{z} \alpha_{i,z} \beta_{j,z}}{\sum_{z=1}^{k} \kappa_{z} \alpha_{i,z} \beta_{j,z}},
\]

\[
\kappa_{z} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} q_{i,j}(z)}{\sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{i,z} \beta_{j,z}},
\]

\[
\alpha_{i,z} = \frac{\sum_{j=1}^{n} A_{i,j} q_{i,j}(z)}{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} q_{i,j}(z)},
\]

\[
\beta_{j,z} = \frac{\sum_{i=1}^{m} A_{i,j} q_{i,j}(z)}{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} q_{i,j}(z)}.
\]

As with the dc-biSBM, this procedure is not guaranteed to converge to the parameters maximizing the likelihood, so we take the parameters obtained from 50 randomly initialized runs and choose those that attain the highest likelihood value.
Empirical Findings We focus our analysis on how “spread out” agents’ distributions over legislation communities are: how many different legislation communities does an agent effectively belong to? We first describe a simple way to quantify this notion. From the model definition, the mean total number of times interest group $i$ lobbies in community $z$ equals $\kappa_z \alpha_{i,z}$, and likewise the mean total number of times politician $j$ sponsors in community $z$ equals $\kappa_z \beta_{j,z}$. A natural choice of measurement is then to normalize these quantities to form probability distributions $p_{i,z} = \frac{\kappa_z \alpha_{i,z}}{\sum_{i=1}^{k_z} \kappa_z \alpha_{i,z}}$ and $q_{j,z} = \frac{\kappa_z \beta_{j,z}}{\sum_{j=1}^{k_z} \kappa_z \beta_{j,z}}$, and then consider the (Shannon) entropies $H_i = H(p_{i,1}, \ldots, p_{i,k})$ and $H_j = H(q_{j,1}, \ldots, q_{j,k})$, where $H(c_1, \ldots, c_k) = -\sum c_i \log_2 c_i$\(^{19}\). In computing entropies, we take logarithms base 2 for the sake of interpretability: an entropy of $H$ may then be interpreted as, roughly speaking, an agent typically participating in $2^H$ legislation communities. Simply put, the higher an agent’s entropy, the more “spread out” their legislation community distribution is.

With this measure, the biLCM model can shed light on the properties of the region near the center of the LSNM latent space. Recall that we did not find legible community structure in this region based on the LSNM, while the dc-biSBM led us to suppose that this region might be characterized by varied interests. To evaluate this possibility, the left panel of Figure 8 repeats the latent space plot from Figure 3, but now divides the latent space into small hexagonal parcels and

\[^{19}\text{We reuse the letter } H \text{ for these quantities since it will always be clear from context whether we are discussing the entropy of interest groups or of politicians. A more principled quantity to consider would be the mean of the entropy of empirical link distributions drawn from our model with a given set of parameters. We do this and find in practice that this mean agrees almost perfectly (Pearson’s } \rho = 0.99 \text{) with the quantity we compute.}\]
shades them according to the mean legislation community membership entropy of actors within them. We clearly observe that actors with more memberships (darker hexagons) are clustered near the center of the latent space.

We provide a schematic visual representation of the same phenomenon in the right panel, in which we represent each actor with a pie chart showing their estimated legislation community membership distribution. Again, we find that actors in the LSNM center region tend to have mixed memberships in many different political communities (marked by pie charts split relatively evenly among many colors). We highlight three examples discussed previously: the CHAMBER OF COMMERCE, which as we suggested earlier has a very broad distribution over legislation communities; BRITISH PETROLEUM (BP), whose activities are mostly divided between the “Energy” (pink) and “Retail & Transportation” (purple) communities; and McAfee, Inc., whose activities are mostly divided between the “Universities & Research” (gray) and “Technology & Telecommunications” (green) communities. On the other hand, we consistently find many of the same industry-specific communities at the edges of the latent space as we have seen before, such as the energy, technology, and finance industries, whose members typically participate in only one or two legislation communities (indicated by pie charts with only one or two dominant colors).

We have already seen in Figure 4 that the politicians in the central LSNM cluster tend to belong to power committees that have broad responsibilities of government oversight or budgetary allocation. The interest groups in this cluster also admit a simple and politically meaningful description: many of them are large associations that represent numerous small businesses, professionals, or causes, such as the CHAMBER OF COMMERCE, the SPECIALTY EQUIPMENT MARKET ASSOCIATION, HERITAGE ACTION FOR AMERICA, and the NATIONAL TREASURY EMPLOYEES UNION. This analysis suggests that these actors participate in lobbying through aggregate action in the lobbying marketplace, either by lobbying politicians on power committees or by serving as a proxy for many individuals or smaller firms to lobby politicians who sit on committees with disparate jurisdictions. We believe that these findings indicate a unique aggregate mode of lobbying, wherein politicians with control over many policy areas are lobbied by numerous otherwise unrelated interest groups while interest groups representing the diverse interests of many constituents lobby numerous otherwise unrelated politicians.

In Table 2, we present the distributions over legislation communities of several interest groups, including some discussed previously. These distributions illustrate the full range of lobbying behaviors the biLCM identifies in the lobbying network. First, firms like MICROSOFT and ARCH COAL

---

20The communities discovered by the biLCM are well-aligned with those previously discovered by the dc-biSBM, so we reuse their labels, associated colors, and substantive interpretations. Quantitative results supporting this alignment are presented in Appendix A.5.
Table 2: Legislation Community Distribution Examples. We give examples of interest groups with focused, intermediate, and spread out legislation community distributions, along with examples of bills illustrating their interests. The distributions are represented as histograms, with eight bars corresponding to the eight legislation communities in the biLCM. We observe some overlap in the community memberships of these groups. For instance, Microsoft and McAfee, Inc. share an involvement in the “technology” community, colored green, while Arch Coal and British Petroleum share an involvement in the “energy” community, colored pink.

lobby exclusively on a single topic, each interacting in one community (“Technology & Telecommunications,” colored green, and “Energy,” colored pink, respectively) with probability at least 85%. Second, firms with broader interests like McAfee, Inc. or British Petroleum lobby primarily in two or three arenas (note that, for instance, Microsoft and McAfee, Inc. share interests in the “Technology & Telecommunications” community, colored green, while McAfee, Inc. has several other interests and interacts in the “Technology & Telecommunications” community with probability only 31% and in the “Universities & Research” community, colored gray, with probability 52%). Finally, large holding companies like Philips North America or lobbying organizations like the Chamber of Commerce lobby across essentially all categories of legislation, interacting in any particular community with probability no more than 25%.

How do we know whether a given political actor’s link community distribution is unusually concentrated or mixed? To provide an answer, we perform a permutation test on the lobbying network to obtain a null distribution of “mixedness” (measured by entropy) of the biLCM legislation community distributions for networks with similar connectivity to the lobbying network (see

<table>
<thead>
<tr>
<th>Interest Group</th>
<th>Distribution</th>
<th>Example Bills</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>. . . .</td>
<td>Startup Act 3.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Law Enforcement Access to Data Stored Abroad Act</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Satellite Television Access and Viewer Rights Act</td>
</tr>
<tr>
<td></td>
<td>. . . .</td>
<td>Cybersecurity Enhancement Act</td>
</tr>
<tr>
<td>McAfee, Inc.</td>
<td>. . . .</td>
<td>Department of Defense Appropriations Act</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Energy Efficient Government Technology Act</td>
</tr>
<tr>
<td>Arch Coal</td>
<td>. . . .</td>
<td>Climate Protection Act</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caring for Coal Miners Act</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Energy Consumers Relief Act</td>
</tr>
<tr>
<td>British Petroleum (BP)</td>
<td>. . . .</td>
<td>Keep American Natural Gas Here Act</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RFS Reform Act</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Corn Ethanol Mandate Elimination Act</td>
</tr>
<tr>
<td>Philips North America</td>
<td>. . . .</td>
<td>Protect Medical Innovation Act</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medicare DMEPOS Market Pricing Program Act</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Renewable Energy Parity Act</td>
</tr>
<tr>
<td>Chamber of Commerce</td>
<td>. . . . . . .</td>
<td>Water Resources Development Act</td>
</tr>
<tr>
<td></td>
<td></td>
<td>United States-Israel Strategic Partnership Act</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Genetically Engineered Food Right-to-Know Act</td>
</tr>
</tbody>
</table>
Appendix A.4.2 for technical details). Since the biLCM can capture both concentrated and mixed community membership, this is also a useful test of the statistical significance of the communities we have been discussing: if the lobbying network has strong community structure compared to a “typical” similar network, then the biLCM should find much more concentrated community memberships in the lobbying network than in such a generic network. Indeed, we find that this is the case, observing that the link community distributions of interest groups in the null model are typically much more mixed than those in the model of the actual lobbying network.

On the whole, we find that the biLCM successfully describes the multiple community memberships of political actors and how those interactions among actors occur in those communities. The proposed methodology allows researchers to quantitatively estimate the relative participation in various communities for each political actor. Furthermore, it allows researchers to relate interest groups with specialized and general interests to politicians with narrow and broad political connections. In sum, the proposed methods not only capture the community structure that we often cannot observe in political networks, but also provide a useful guide for gauging the varied interests and complex preference aggregation that drives lobbying in legislative politics.

5 Concluding Remarks

Lobbying is known to be an important channel through which special interest groups influence the U.S. legislative process. Even so, observable connections between interest groups and the politicians they target have proved elusive, since the individual politicians contacted in the course of lobbying need not be disclosed. Because interest groups must report the bills on which they lobby, however, and each bill has a unique sponsor, we are able to assemble a new lobbying network dataset that measures legislative interactions between interest groups and politicians for all bills introduced since the 106th Congress.

We show that this network can be usefully modeled with both latent space models and community models. Unlike previous applications of latent space models in political science, our models suggest that lobbying interactions depend not on an ideological spectrum, but rather on domain-specific community membership. Most of these domains are organized around industry interests and political issues, including energy, finance, and veterans’ affairs. Interest groups with concerns in these domains lobby politicians who sit on congressional committees with relevant jurisdiction, such as the House Committee on Energy and Commerce, the House Committee on Financial Services, and the House Committee on Veterans’ Affairs. Stochastic block

On the other hand, the link community distributions of legislators in the null model are comparably mixed to those in the model of the actual lobbying network. This agrees with our previous observation that varied interests are more typical for politicians than for interest groups: while many interest groups are focused on specific industries, most politicians are involved in a variety of policy domains.
models confirm and clarify this finding by explicitly modeling the community structure that we observe. Furthermore, our bipartite link community model illustrates that politicians who serve in power committees with broad responsibilities are active in multiple political communities and are linked to aggregate interest groups that represent heterogeneous political interests. The consistency between the latent space model and non-geometric community models in describing these patterns provides strong evidence that the interest group–politician network in U.S. legislative politics is not organized according to ideology. It is shared interests, not shared ideology, that drives lobbying.

By showing that it is helpful to view lobbying activity as occurring within a network of political actors, our findings suggest that other network analysis techniques may reveal further meaningful patterns in how lobbying is conducted in the U.S. political system. Of course, the lobbying network contains much more information than we have analyzed here, including bills and their contents, lobbyists and their lobbying firm affiliations, the money spent on lobbying, and the timing of lobbying and legislative events. Extending the network models in this work to describe new types of agents and interactions should further enhance the picture of lobbying and its political effects that we have obtained here.
References


A Supplementary Appendix

A.1 Dataset Construction

A.1.1 Identifying Bills and Missing Congress Numbers

Identifying congressional bills in lobbying reports is difficult because bill numbers are repeated across Congresses, and often do not appear directly annotated with Congress numbers in lobbying reports. Using the report filing year to guess the Congress often leads to erroneous matches, because reports filed at the beginning of a new Congress tend to include disclosures of lobbying activities from the previous year (and therefore, if a new Congress has begun recently, from the previous Congress as well). For example, consider the following lobbying report filed by GOOGLE, INC. in 2013. It reads:

Monitor legislation regarding online privacy including Safe Data Act (H.R. 2577, S. 1207) and Do not track proposals (H.R. 654). Monitor any Congressional or Administration efforts to impose privacy laws on search engines. Monitor Spectrum acts (S. 911, H.R. 2482).

Figure A.1: First Quarter Report by GOOGLE, INC. in 2013

A naive guess would be that the bill H.R. 2577 refers to a bill from the 113th Congress, because the report was filed in 2013. However, it is clear from the report that this is a bill from the 112th Congress, the “SAFE Data Act.” We use the following strategies to mitigate this problem and correctly identify Congress session numbers under various circumstances.

1. **Bill Number Search**: We first identify bill numbers (e.g., H.R. 2577 above) using regular expression search in the report text. In the above example in Figure A.1, our algorithm would identify bill numbers H.R. 2577, S. 1207, H.R. 654, S. 911, and H.R. 2482. Note that all of these bills are from the 112th Congress rather than the 113th.

2. **Congress Identification**: Given a bill number found in a specific issue text (a section of the lobbying report), we attempt to identify the most likely Congress to which that bill would belong using other text around the bill number. We consider a range of candidate Congresses extending backwards from the Congress containing the year that the lobbying report was filed. By default, we consider the three preceding Congresses; in the above example, therefore, we would consider the 113th, 112th, and 111th Congresses. We then retrieve the bills having the same number as the given bill from each of these Congresses (omitting the Congresses that do not have a bill of that number), and compute a bag-of-words representation (after a tokenization and stopword filtering pipeline) of each of those bills, producing vectors $v_1, \ldots, v_n$ representing the $n$ candidate bills. We also compute the same representation of the text around the mention of the bill number in the lobbying report, producing a vector...
\(\mathbf{w}\) representing that text. We then choose the Congress number by maximizing the cosine similarity between the \(\mathbf{v}_i\) and \(\mathbf{w}\), choosing bill \(i^*\) with index given by

\[
i^* = \arg\max_{1 \leq i \leq n} \frac{\mathbf{v}_i^\top \mathbf{w}}{\|\mathbf{v}_i\| \|\mathbf{w}\|}.
\] (13)

If no bill having the same number exists in the entire range of Congresses we consider, we simply guess that the bill comes from the Congress of the year the lobbying report was filed.

3. **Congress Propagation:** If we successfully find a match for a Congress, it may be propagated to the other bills mentioned in the lobbying report, since, being scheduled on a quarterly basis, lobbying report will almost always only mention legislation from a single Congress. If different bills in a lobbying report disagree on the best-matching Congress, a majority vote may be taken, but this rarely occurs in practice.

4. **Bill Title Search:** Bills are sometimes only referred to by titles or alternate names. To account for this, we clean and tokenize the specific issue sections of the lobbying report, and perform a text matching operation against a table of bill titles. For instance, this operation would identify “Safe Data Act” in our previous example, even if the bill number H.R. 2577 were not mentioned.

5. **Bill Range Expansion:** It is also common for bills with nearby numbers to be related, and for lobbying reports to refer to ranges of bills when lobbying all of them at once. For instance, a lobbying report filed by **MatTEL, INC.** in 2002 contains the following text:

> “H.R.3009, Trade Act of 2002. Certain miscellaneous tariff bills to suspend the rates of duty on certain toy-related articles (H.R.4182-4186; S.2099-2103). WTO market access negotiations for non-agricultural products Port and border security measures”

Figure A.2: Midyear Report by **MatTEL, INC.** in 2002

Therefore, if we find two bill numbers that are close (by default, we take this to mean that they share the same prefix and their numbers differ by at most 10), then we consider all other bills with numbers in between as also being lobbied in the same report. For instance, the pattern “H.R. 4182-4186” in the excerpt shown in Figure A.2 would be expanded into bills H.R. 4182, H.R. 4183, H.R. 4184, H.R. 4185, and H.R. 4186, all of which we would consider lobbied on by **MatTEL, INC.**

A.1.2 Filtering

Besides reducing the size of our data, the key task in filtering is to increase the minimum degree of the actors included for analysis in the lobbying network. This is especially important for the LSNM;
it is intuitive that there is no principled way to position an actor in the lobbying marketplace if the actor does not lobby or sponsor very frequently. Computationally, this issue is reflected in poorer convergence properties for actors with fewer interactions. Thus, while it may appear appealing at first to filter the dataset independently for interest groups and politicians by considering summary statistics like total numbers of bills sponsored or lobbying reports submitted during a Congress, such filtering is not necessarily aligned with the goal of finding a large submatrix of the interaction matrix $A$ (or an induced subgraph of the bipartite lobbying network graph) with only agents of sufficiently high degree. In particular, we must avoid politician filtering causing some interest groups who survived interest group filtering to have their degree reduced again, or vice-versa.

We find that the simplest way to avoid such problems is to build the full interaction matrix $A$ without filtering first, and then filter based only on interactions in a way that explicitly ensures that only high-degree agents remain. To that end, we use a filtering procedure defined by two thresholds, denoted $T_I$ and $T_P$, which alternates removing all interest groups with degree lower than $T_I$ and removing all politicians degree lower than $T_P$, until no more interest groups or politicians are removed in a full iteration. This is a simple greedy algorithm for finding an induced subgraph of the full lobbying network with only high-degree nodes, which we find to suffice for our purposes. Typically only a few iterations are required for the algorithm to terminate, but never does just one iteration suffice on the datasets we consider, showing that the algorithm is indeed performing additional filtering beyond the first thresholding step.

We construct our main dataset by the above procedure with $T_I = 30$ and $T_P = 5$. This produces a dataset with 676 interest groups and 523 legislators. While, as mentioned in the main text, in the unfiltered dataset 97.96% of pairs of actors do not interact, in the filtered dataset this proportion is reduced to 90.23%.

### A.2 LSNM Additional Information

#### A.2.1 Parameter Identification

Several transformations of the parameters $\alpha, \beta, \theta_i, \psi_j$ leave the means of the $A_{i,j}$ unchanged, so we cannot expect the posterior distribution not concentrate on a single collection of parameter values without imposing further constraints. There are four important families of symmetries under which our model is invariant, which we summarize below (for more detailed discussion, see e.g. Jackman (2001)). These are scaling, where for any invertible symmetric matrix $S \in \mathbb{R}^{d \times d}$, each $\theta_i$ may be multiplied by $S$ and each $\psi_j$ by $S^{-1}$; rotation, where for any orthogonal matrix $Q \in \mathbb{R}^{d \times d}$ each $\theta_i$ and $\psi_j$ may be multiplied by $Q$; popularity translation, where any constant may be added to each $\tilde{\alpha}_i$ and subtracted from each $\tilde{\beta}_j$; and mixed translation, where a vector $v \in \mathbb{R}^d$ may be added to each $\theta_i$ (resp. $\psi_j$) and $v^T \psi_j$ (resp. $v^T \theta_i$) subtracted from each $\tilde{\beta}_j$ (resp. $\tilde{\alpha}_i$).
We combine two strategies to accomplish identification, constraining the parameters to eliminate the above symmetries. Most of the task is accomplished by imposing hierarchical priors with constrained hyperparameters, as indicated by the form of the normal distributions appearing in equation (4). Specifically, the popularity translation invariance may be eliminated by assigning the $\tilde{\alpha}_i$ a normal prior with mean zero, and the mixed translation invariance by assigning the $\theta_i$ and $\psi_j$ a normal prior with mean zero. The scaling invariance is partly resolved by further setting the covariance matrices of the $\theta_i$ and $\psi_j$ to be equal; the only scaling transformations still admissible are those where $S$ is a symmetric orthogonal matrix. The rotation invariance is resolved by choosing this shared covariance matrix to furthermore be diagonal, which also constrains the above scaling invariances to those where $S$ is a permutation matrix, a diagonal matrix with diagonal entries $\pm 1$, or a product thereof.

The remaining task is to eliminate the discrete family of $2^d \cdot d!$ (e.g., two in the one-dimensional case) symmetries corresponding to applying arbitrary permutations and sign changes to vectors’ coordinates. Geometrically, these are compositions of reflections through certain hyperplanes in $\mathbb{R}^d$. We find that this task is most effectively accomplished not by assigning “symmetry-breaking” priors to the parameters of all agents, but rather by fixing the exact latent positions of a small number of distinguished agents. This generalizes the technique of Kubinec (2018) to multiple latent spatial dimensions. The main challenge in applying this technique is to choose the fixed agents carefully, so that it is not possible for one of the reflection symmetries to leave all of the fixed agents’ latent positions almost unchanged, in which case the constraints are said to “split the likelihood” (Bafumi et al., 2005). To address this, we first identify $2^d \cdot d!$ agents and fix their positions based on a fast variational estimation routine included in the Stan software package, and then run Monte Carlo sampling to perform the final inference. When $d = 2$, the agents to fix are chosen to maximize the number of samples drawn from the variational approximate posterior that lie in a single sector of the plane among the $2^2 \cdot 2! = 8$ sectors demarcated by the lines $y = 0$, $x = 0$, $y = x$, and $y = -x$ in the $(x,y)$ plane. Similarly, when $d = 1$, the agents to fix are chosen to maximize the number of samples lying in the positive or negative sides of the line. In both cases, ties are broken by maximizing the norm of the agent’s mean latent position (following the heuristic of Kubinec (2018)).

This technique appears effective both on our dataset and on synthetic data drawn from the true data-generating process described by our Bayesian model. We consistently obtain Gelman-Rubin $\hat{R}$ statistics smaller than 1.01 for all parameters (as reported by Stan diagnostic logging) for both $d = 1$ and $d = 2$, and Pearson’s $\rho$ greater than 0.95 between the collections of posterior means estimated from different chains. Comparable correlations are also achieved with the “ground
truth” values when data is drawn from the true data-generating process and when agent positions are fixed to their true values rather than the results of a variational approximation.

A.2.2 Dimensionality Selection

We present some heuristics on choosing a dimensionality for the LSNM based on a simpler modeling technique, the spectral biclustering of Kluger et al. (2003). In this method, one computes the singular value decomposition (SVD) of the bipartite adjacency matrix $A$, and searches for piecewise constant structure among the leading singular vectors. The idea we will borrow from this method is simply that the leading singular vectors contain most of the relevant information for the analysis of $A$, proportionally to the magnitude of their singular values, essentially the same concept as in PCA analyses. We then make a plot of the singular values, analogous to a scree plot, and search for an “elbow” indicating the suitable number of singular vectors to use. That number gives an estimate of the dimensionality of the low-rank structure in $A$, and thus also a heuristic estimate of the dimensionality to take in the LSNM. In Figure A.4, we show that a dimensionality of two is a reasonable choice if this heuristic is admitted.

A.2.3 Computation

The Stan code in [Listing A.1] was used to fit our latent space models, run through the PyStan interface. In practice, by far the most important optimization for this model is the vectorization of the declaration of the Poisson distribution for edge weights. Factorizing the latent space position prior covariance matrix into Cholesky factors and using them to transform standard normal variables is a common optimization for multivariate normal priors, but our covariance matrix is so small that we find this not to affect (or even to degrade, in some cases) sampling performance.

For each latent space model we run four MCMC chains, each drawing 10,000 samples, the first 4,000 of which are discarded as a burn-in period, leaving us with 6,000 usable samples per chain, for a total of 24,000 samples. These are used to compute estimates of posterior means of all of our models parameters as used in all visualizations in the main text.

A.2.4 MCMC Diagnostics

Position Variances We use a visualization that captures the variance structure of all of our latent space position estimates at once. For any agent, if we take $N$ samples of its $d$-dimensional position, we view these samples as the columns of a matrix $X \in \mathbb{R}^{d \times N}$. We let $\mu \in \mathbb{R}^d$ be the vector of means of the rows of $X$, and let $X_0$ be the centered matrix obtained by subtracting $\mu_i$ from each entry of the $i$th row of $X$. Then, $W = \frac{1}{N-1} X_0 X_0^\top \in \mathbb{R}^{d \times d}$ is the empirical covariance matrix of the position samples, which we diagonalize to obtain orthonormal eigenvectors $v_1, \ldots, v_d$ and associated non-negative eigenvalues $\lambda_1, \ldots, \lambda_d$. This is essentially a principal component analysis.
(PCA), except unlike the usual high-dimensional setting for PCA, we have many points in a low-dimensional space, $N \gg d$. Nonetheless, the computed quantities have the same interpretations: $\lambda_i$ is the amount of variance of the sampled positions in the direction $v_i$, and the shape of the set of sampled positions is approximately the ellipsoid with axes in the directions of the $v_i$ having lengths $\sqrt{\lambda_i}$.

For small $d$, in particular for $d = 2$ as we use for most of our models, these ellipses can be drawn for all embedded points at once. Small ellipses correspond to concentrated posterior distributions for latent positions, which justify the use of the mean position point estimate in our analysis in the main text. This visualization is given for the model inferred from our main dataset in Figure A.5.

**Popularity Factor Variances** For popularity factors $\alpha_i$ and $\beta_j$, it is more convenient to consider the variance of the estimates of the exponentials $\exp(\alpha_i)$ and $\exp(\beta_j)$: first, these are non-negative and so their statistics are easier to understand, and second they are the quantities we visualize and are ultimately more interested in, since the interaction mean scales linearly in each. We thus compute the standard error (standard deviation divided by the mean) of each of these quantities. As for the position variances, we visualize all of the standard errors at once, giving in Figure A.6 their histograms for our main dataset. We observe that the distributions of the standard errors are strongly concentrated on the low values; manual inspection reveals that the agents with the highest standard errors are typically those with the lowest degrees, which are less significant to our substantive analysis (these agents also tend to have smaller popularity factors, which can further inflate the standard error).

**Trace Plots** In Figure A.7 we show two representative sets of trace plots for latent space position and popularity factor parameters for one interest group and one politician. Sampling for all parameters regardless of the group appears to mix rapidly following the warmup period.

### A.3 biSBM Additional Information

#### A.3.1 Degree Correction

The uncorrected biSBM usually exhibits an undesirable property of grouping nodes by their degree, especially when the degree distribution of the network is not tightly concentrated, a feature that we observed previously in the lobbying network. In our case, the model is likely to cluster politicians who sponsor larger numbers of lobbied bills and interest groups who lobby many bills together, essentially grouping actors by their overall popularity rather than the specific legislation they tend to be involved with. To correct for this, we follow Larremore, Clauset, and Jacobs (2014) and introduce additional parameters $\alpha_i$ and $\beta_j$ per interest group and politician respectively, and
adjust the model given by (5) to

\[ A_{i,j} \sim \text{Poisson}(\alpha_i \beta_j B_{x_i, y_j}), \]  

(14)

for \( \alpha_i, \beta_j > 0 \). We must now identify these new parameters, since among the sets of parameters \( \alpha, \beta, \) and \( B \), a positive constant can be multiplied to any one and divided from any other without changing the model. This is easily resolved by constraining \( \sum_{x_i=a} \alpha_i = \sum_{y_j=b} \beta_j = 1 \) for all \( a \in [k] \) and \( b \in [\ell] \). Letting \( \text{deg}(\bullet) \) be the degree of an agent, the maximum likelihood values may be calculated as

\[
\hat{\alpha}_i = \frac{\text{deg}(i)}{\sum_{x_i=x_i'} \text{deg}(i')}, \quad \hat{\beta}_j = \frac{\text{deg}(j)}{\sum_{y_j=y_j'} \text{deg}(j')},
\]

(15)

the fractions of the edges leaving a node’s cluster that are leaving the node itself. Due to that interpretation of the new parameters, the resulting model is named the Degree-Corrected Bipartite Stochastic Block Model (dc-biSBM). As has been shown in Karrer and Newman (2011) and Larremore, Clauset, and Jacobs (2014), degree correction is an important modeling mechanism in networks with outlying high-degree nodes, in which setting the dc-biSBM typically obtains more meaningful clusterings than the biSBM. This distinction in our setting is discussed further in Appendix A.5.

A.3.2 Assortativity

In Figure A.8, we plot the mean interaction value between actors of each pair of communities discovered by the dc-biSBM. We observe that, after applying a suitable permutation to the community labeling, that it is possible to match politician and interest group communities into strongly interacting pairs, such that most other pairs are weakly interacting. This suggests that the dc-biSBM output may be interpreted as placing politicians and interest groups into joint strongly interacting communities, a property known as assortativity.

A.4 biLCM Additional Information

A.4.1 EM Algorithm

In this appendix, we present a more detailed derivation of the EM update equations for the link community model. After discarding constant terms, the log-likelihood of this model is given by

\[
\log P(A | \alpha, \beta, \kappa) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} \log \left( \sum_{z=1}^{k} \kappa_z \alpha_{i,z} \beta_{j,z} \right) - \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{z=1}^{k} \kappa_z \alpha_{i,z} \beta_{j,z}
\]

(16)

We then apply the standard technique of introducing parameters \( q_{i,j}(z) \) that form a probability distribution over \( z \) for fixed \( i \) and \( j \) and applying Jensen’s inequality to obtain the objective function
of our optimization task, a lower bound on the log-likelihood:

$$\mathcal{L}(A, \alpha, \beta, \kappa, q) \overset{df}{=} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{z=1}^{k} \left( A_{i,j} q_{i,j}(z) \log \left( \frac{\kappa_{z} \alpha_{i,z} \beta_{j,z}}{q_{i,j}(z)} \right) - \kappa_{z} \alpha_{i,z} \beta_{j,z} \right)$$  \hspace{1cm} (17)

$$\leq \log P(A | \alpha, \beta, \kappa).$$

We then seek to maximize \( \mathcal{L} \) via coordinate ascent. Maximizing with respect to the \( q_{i,j}(z) \) with all other parameters fixed simply sets these parameters to the values that make Jensen’s inequality sharp, which are

$$q_{i,j}(z) = \frac{\kappa_{z} \alpha_{i,z} \beta_{j,z}}{\sum_{z=1}^{k} \kappa_{z} \alpha_{i,z} \beta_{j,z}}. \hspace{1cm} (18)$$

It is also straightforward to differentiate with respect to \( \kappa_{z} \), obtaining the update

$$\kappa_{z} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} q_{i,j}(z)}{\sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{i,z} \beta_{j,z}}. \hspace{1cm} (19)$$

For the \( \alpha \) and \( \beta \) parameters, we must constrain our optimization to respect the normalization that for all \( z \), \( \sum_{i=1}^{m} \alpha_{i,z} = \sum_{j=1}^{n} \beta_{j,z} = 1 \). If add to \( \mathcal{L} \) Lagrange multiplier terms \( \lambda_{z}(1 - \sum_{i=1}^{m} \alpha_{i,z}) + \mu_{z}(1 - \sum_{j=1}^{n} \beta_{j,z}) \), then we obtain the updates

$$\alpha_{i,z} = \frac{\sum_{j=1}^{n} A_{i,j} q_{i,j}(z)}{\lambda_{z} + \sum_{j=1}^{n} \kappa_{z} \beta_{j,z}}, \hspace{1cm} (20)$$

$$\beta_{j,z} = \frac{\sum_{i=1}^{m} A_{i,j} q_{i,j}(z)}{\mu_{z} + \sum_{i=1}^{m} \kappa_{z} \alpha_{i,z}}, \hspace{1cm} (21)$$

but now we see that to obtain the desired normalization we must in fact take \( \lambda_{z} \) and \( \mu_{z} \) such that they cancel out these denominators and leave us with the simpler

$$\alpha_{i,z} = \frac{\sum_{j=1}^{n} A_{i,j} q_{i,j}(z)}{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} q_{i,j}(z)}, \hspace{1cm} (22)$$

$$\beta_{j,z} = \frac{\sum_{i=1}^{m} A_{i,j} q_{i,j}(z)}{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} q_{i,j}(z)}, \hspace{1cm} (23)$$

which in particular do not involve any coupling between \( \alpha_{i,z} \) and \( \beta_{j,z} \), meaning that a single iteration of updates suffices for the M step of our EM algorithm, simplifying the calculation substantially compared to the nested coordinate ascents that are usually involved in mixed-membership stochastic block models.

**A.4.2 Null Model Analysis**

The *configuration model*, a standard tool in the theory of random graphs, gives a natural way to build a null model for our analyses. In this model, the degree distribution of the graph is retained,
but all connections are “rewired” randomly. More precisely, in our bipartite setting, if there are \(m\) interest groups indexed 1, \ldots, \(m\) and \(n\) politicians indexed 1, \ldots, \(n\), suppose that interest group \(i\) is connected to a total of \(d^I_i\) politicians, and politician \(j\) is connected to a total of \(d^P_j\) interest groups. Note that the total number of edges \(E\) is given by the sums of either of these quantities:

\[
E = \sum_{i=1}^{m} d^I_i = \sum_{j=1}^{n} d^P_j. \tag{24}
\]

Now, we perform the following random process to generate a draw from the configuration model:

1. From each interest group \(i\), produce \(d^I_i\) “stubs” or partial edges, which we may label \(e^I_{i,1}, \ldots, e^I_{i,d^I_i}\). Likewise, from each politician \(j\), produce stubs \(e^P_{j,1}, \ldots, e^P_{j,d^P_j}\).

2. Generate a uniformly random matching of the two sets \(\{e^I_{i,k}\}\) and \(\{e^P_{j,\ell}\}\) (note that by the preceding remark, the two have equal size).

3. Generate a bipartite graph by putting an edge between interest group \(i\) and politician \(j\) whenever \(e^I_{i,k}\) and \(e^P_{j,\ell}\) were matched for some \(k, \ell\) in the previous step (so that the total number of edges is the total number of pairs \(k, \ell\) for which these two stubs were matched).

It is simple to verify that the configuration model satisfies the following desirable properties. First, every draw of the configuration model has the same degree distribution as the original graph; that is, the collections of numbers \(\{d^I_i\}\) and \(\{d^P_j\}\) remain the same. And second, the configuration model draws a uniform sample from the bipartite graphs having those degree distributions. In this sense, the configuration model is the canonical null model that retains the degree distribution of a graph, but “forgets” all other details.

To understand how exceptional our findings with the biLCM are, we analyze the link community distribution entropy under the configuration model for our main dataset here. In particular, in Figure [A.9] we compare the entropy distributions obtained from the biLCM run on this dataset with averaged distributions obtained from many draws of the configuration model applied to the same dataset (as we would hope, those distributions concentrate fairly well around a single representative null entropy distribution). We see that the entropies of the link community distributions for interest groups in the actual dataset are much lower, meaning the link community distributions are much more concentrated, than those for the null model. This gives further evidence that the communities we find are statistically significant, and not, for instance, merely artifacts of the graph’s degree distribution (not a farfetched possibility, as less sophisticated community detection algorithms often suffer from detecting spurious communities consisting only of high-degree nodes).
A.5 Model Comparison

We give a few points of comparison among our latent space and community models, confirming that the models all consistently reflect the same underlying properties of the lobbying network.

Figure A.10 shows the interaction matrix $A$, with its rows and columns ordered according to the groupings given by the biSBM (top panel) and the dc-biSBM (bottom panel). Both models show that there exist clear “checkerboard” community patterns in the lobbying network data, and, as mentioned before, without degree correction the biSBM suffers from the flaw of grouping high-degree nodes together. (Indeed, one of the interest group communities found without degree correction consists of exactly the two interest groups with highest degree, the Chamber of Commerce and Iraq and Afghanistan Veterans of America.) The dc-biSBM, in contrast, gives much more balanced community sizes and less concentration of the highest-degree agents. Degree correction also yields community memberships which, especially for interest groups, correspond more closely to the clusters we found ex post in the LSNM.

Next, in the LSNM we assign latent positions to each agent, while in the community models we divide the agents into discrete communities. This prompts the question of whether the latent positions of the LSNM respect the clusterings suggested by the community models, i.e. of whether the communities identified the block models are localized in the latent space. We address this question in Figure A.11 and Figure A.12 for the biSBM and dc-biSBM models, respectively, observing that the latter both gives less weight to high-degree nodes, and is more aligned with the LSNM geometric representation.

Lastly, in the main text we claimed that the communities of the dc-biSBM and the biLCM were well-aligned, and could be viewed as having the same substantive interpretations. We illustrate this in Figure A.13, plotting the mean probability with which an actor from a given dc-biSBM community interacts in a given biLCM community and observing that, after a suitable permutation, for each dc-biSBM community there is indeed one best-aligned biLCM community.
Figure A.3: Distribution of Political Actions. We present the distributions of politician sponsorship and interest group lobbying counts over the 113th Congress. Note that these statistics are aggregated before any of the filtering performed to form the dataset we work with later. Bills sponsored only count those bills lobbied by at least one interest group. The activity counts and names of the most active politicians and interest groups are presented in the tables below.
Figure A.4: Bipartite Adjacency Matrix Singular Value Decay. We plot the singular values of the interaction matrix $A$, highlighting that the singular values decay rapidly and the first few appear to capture much of the total weight of this matrix. The first two singular values are highlighted, corresponding to our analysis primarily of a two-dimensional latent space model in the main text.

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Restricted</th>
<th>Full</th>
<th>Restricted</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry State</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Dependent Variable</td>
<td>$F$-Statistic ($p$-Value)</td>
<td>$F$-Statistic ($p$-Value)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1D LSNM</td>
<td>13.72 ($&lt;2.2e-16$)</td>
<td>0.86 (0.71)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D LSNM, Dimension 1</td>
<td>9.92 ($&lt;2.2e-16$)</td>
<td>0.81 (0.78)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D LSNM, Dimension 2</td>
<td>13.25 ($&lt;2.2e-16$)</td>
<td>1.65 (0.01)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.1: Regression Analysis of LSNM Covariates (Interest Groups). Statistics of linear models of covariates against the one- and two-dimensional LSNM latent dimensions for interest groups. We compare full linear models with all available covariates against restricted models, omitting either industry classification covariates or geographic information. In the bottom three rows, for each of the three latent dimensions, we give the $F$ statistics and $p$-values of the comparison $F$-test. The test illustrates that the industry covariates are significantly explanatory of the LSNM latent space organization, and, with lesser confidence, geographic covariates are also explanatory of one dimension in the two-dimensional LSNM.
Figure A.5: **LSNM Latent Position Uncertainties.** We plot an estimate of the “typical set” of each actor’s latent position in the two-dimensional LSNM, in the form of an ellipse containing most of the latent position samples drawn by MCMC sampling. The computations giving the parameters of these ellipses are detailed in Section A.2.4.
Figure A.6: **LSNM Popularity Factor Uncertainties.** Distributions of the coefficients of variation (standard deviation as a fraction of the mean) of the sampled popularity factors for all interest groups and all politicians.

Figure A.7: **LSNM Trace Plots.** We show trace plots for the model parameters (two latent space dimensions and one popularity factor) drawn from two (out of four) MCMC chains following the warmup period, for one interest group, the Chamber of Commerce, and one politician, Barbara Boxer (D-CA).
Latent space network model.

Implements the model

\[ A_{ij} \sim \text{Poisson}(\mu_{ij}) \]

where

\[ \mu_{ij} = \exp(a_i + b_j - ||x_i - y_j||_2^2) = \exp(aa_i + ba_j + 2<x_i, y_j>) \]

The names of the variables in the model are:

- \(aa_i\) : row_factor_adj
- \(ba_j\) : col_factor_adj
- \(x_i\) : row_embedding
- \(y_j\) : col_embedding

```plaintext
data {
  int<lower=1> D; // dimension of latent space
  int<lower=2> N_row; // number of rows
  int<lower=2> N_col; // number of columns
  int N_fixed_row; // number of pinned rows
  int N_fixed_col; // number of pinned columns
  int<lower=0> edges[N_row, N_col]; // connection strength data
  int fixed_row_index[N_fixed_row]; // indices of fixed row actors
  matrix[N_fixed_row, D] fixed_row_embedding; // positions of fixed row actors
  int fixed_col_index[N_fixed_col]; // indices of fixed column actors
  matrix[D, N_fixed_col] fixed_col_embedding; // positions of fixed column actors
}

transformed data {
  int flat_ix;
  int flat_edges[N_row * N_col];
  flat_ix = 1;
  for (j in 1:N_col) {
    for (i in 1:(N_row)) {
      flat_edges[flat_ix] = edges[i][j];
      flat_ix = flat_ix + 1;
    }
  }
}

parameters {
  vector<lower=0.01>[D] cov_embedding_diag;
  matrix[N_row, D] row_embedding;
  matrix[D, N_col] col_embedding;
  real mu_col_factor_adj;
  real<lower=0.01> var_row_factor_adj;
  real<lower=0.01> var_col_factor_adj;
  vector[N_row] row_factor_adj;
  row_vector[N_col] col_factor_adj;
}
```
model {
    vector[N_row * N_col] means;
    int fixed_row_flag;
    int fixed_col_flag;

    for (i in 1:N_row) {
        fixed_row_flag = 0;
        for (j in 1:N_fixed_row) {
            if (i == fixed_row_index[j]) {
                row_embedding[i,] ~ normal(fixed_row_embedding[j,], 1e-4);
                fixed_row_flag = 1;
            }
        }
        if (fixed_row_flag == 0) {
            row_embedding[i,] ~ normal(0.0, cov_embedding_diag);
        }
    }

    for (j in 1:N_col) {
        fixed_col_flag = 0;
        for (k in 1:N_fixed_col) {
            if (j == fixed_col_index[k]) {
                col_embedding[,j] ~ normal(fixed_col_embedding[,k], 1e-4);
                fixed_col_flag = 1;
            }
        }
        if (fixed_col_flag == 0) {
            col_embedding[,j] ~ normal(0.0, cov_embedding_diag);
        }
    }

    row_factor_adj ~ normal(0.0, var_row_factor_adj);
    col_factor_adj ~ normal(mu_col_factor_adj, var_col_factor_adj);

    means = to_vector(
        rep_matrix(row_factor_adj, N_col) +
        rep_matrix(col_factor_adj, N_row) +
        2.0 * row_embedding * col_embedding);

    flat_edges ~ poisson_log(means);
}

Listing A.1: Stan code defining a sampler for the posterior distribution of the LSNM. Note that for \( d = 1 \) the model is greatly simplified, and it is much more efficient to remove the extra dimension from the types related to the latent space position distribution.
Figure A.8: **Assortativity in dc-biSBM.** We plot the mean interactions between each pair of communities in the dc-biSBM. If the communities are suitably permuted, we observe that the diagonal entries of the resulting matrix are typically the largest, indicating that the model finds an assortative community structure.

Figure A.9: **Link Community Entropy Distribution vs. Null Model.** We plot the distribution of entropies of link community distributions obtained by the biLCM run on the full 113th Congress dataset, compared to an average (per histogram bin) of the same distribution for several draws of the null configuration model generated from the same dataset. We observe that for interest groups the actual link community distributions are much more concentrated than those obtained in the null model, while for politicians the link community distributions are comparable.
Figure A.10: **Effect of Degree Correction.** We show the results of the biSBM and the dc-biSBM by permuting the interaction matrix to group together the clusters that each model infers. Both models identify community structure, and, as expected, the dc-biSBM infers more balanced cluster sizes.
Figure A.11: **LSNM vs. biSBM.** We plot the latent positions of legislators and interest groups from the LSNM, split by their memberships in clusters from the biSBM. For the sake of visual clarity, we no longer vary the sizes of the plotted latent space positions based on popularity factors. Localization of groups in the latent space suggests that the LSNM is somewhat consistent with the biSBM, but much of the fine-grained interest group clustering we observe in Figure 3 is not reflected in the biSBM results.
Figure A.12: **LSNM vs. dc-biSBM.** We plot the latent positions of legislators and interest groups from the LSNM, split by their memberships in clusters from the dc-biSBM, with both run on the main dataset. Almost all clusters found by dc-biSBM are closely localized in the latent space, showing that the dc-biSBM captures much of the same information as the LSNM.
Figure A.13: **dc-biSBM vs. biLCM.** We plot communities of the dc-biSBM against those of the biLCM, measuring pairwise alignment by the mean probability of an actor in the given dc-biSBM community to interact in the given biLCM community. We observe that, when community labels are suitably permuted, it is possible to match communities from either model into well-aligned pairs.