Online Appendix
Mapping Political Communities: A Statistical Analysis of Lobbying Networks in Legislative Politics*[†]

In Song Kim[‡] Dmitriy Kunisky[§]

September 18, 2019

*Funding: This work was supported by the National Science Foundation (SES-1264090, SES-1725235 to I.K.).

[†]Acknowledgements: We thank Pablo Barberá, J. Lawrence Broz, Devin Caughey, Bruce Desmarais, Marina Duque, Jeff Harden, Lisa McKay, Nolan McCarty, Cristopher Moore, Michael Peress, Yunkyu Sohn, Erik Voeten, and Hye Young You for helpful comments. We also thank seminar participants at NYU-Abu Dhabi, Yale University, Duke University, Princeton University, University of California, San Diego, and University of Maryland, and conference participants at the American Political Science Association, Asian Political Methodology and Political Methodology annual meetings.

[‡]Associate Professor, Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA, 02139. Email: insong@mit.EDU, URL: http://web.mit.edu/insong/www/

[§]Ph.D. Student, Department of Mathematics, Courant Institute of Mathematical Sciences, New York University, New York, NY, 10012. Email: kunisky@cims.nyu.edu, URL: http://www.kunisky.com/

A Dataset Construction

A.1 Identifying Bills and Missing Congress Numbers

Identifying congressional bills in lobbying reports is difficult because bills are numbered starting at 1 in each new session of Congress, and often do not appear directly annotated with session numbers in lobbing reports. Using the report filing year to guess the session number 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:

> 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 $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ 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 $\boldsymbol{w}$ representing that text. We then choose the most likely Congress number by maximizing the cosine similarity between the $\boldsymbol{v}_i$ and $\boldsymbol{w}$, choosing bill i^* with index given by

$$i^* = \underset{1 \leq i \leq n}{\operatorname{argmax}} \frac{\boldsymbol{v}_i^\top \boldsymbol{w}}{\|\boldsymbol{v}_i\| \|\boldsymbol{w}\|}. \tag{1}$$

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.2 Filtering

To form the dataset describing the 113th Congress that we analyze in the main text, we *filter* the lobbying database information down to a set of especially active politicians and interest groups. Besides reducing the size of our data, the primary purpose of filtering is to increase the minimum degree of the actors included for analysis. This is important for our models to produce interpretable results: intuitively, it is difficult to accurately model an actor's role in the lobbying marketplace if the actor does not lobby or sponsor very frequently. Computationally, this issue is reflected in poor convergence properties of relevant model parameters for actors with fewer interactions.

It may seem appealing 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. However, this filtering strategy is not necessarily aligned with the goal of finding a large submatrix of the interaction matrix $\mathbf{A}$ (or an induced subgraph of the bipartite lobbying network graph) with only actors 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 $\mathbf{A}$ 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.

The Network Dataset 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. As expected, these actors interact more frequently than typical actors in the entire lobbying network. Both lobbying and 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 (see Figure F.1 for more precise statistics.) In fact, most politician–interest group pairs do not interact at all, 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 most interactions is Senator John Rockefeller (D-WV) and the National Cable Telecommunications Association, which interacted on 26 bills. In the filtered dataset, on the other hand, only 90.23% of pairs of actors do not interact on any bills.

B biLCM Additional Information

B.1 EM Algorithm

We present a more detailed derivation of the EM update equations for the biLCM. After discarding constant terms, the log-likelihood of this model is given by

$$\log \mathbf{P}(\mathbf{A} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} \log \left(\sum_{z=1}^{k} \kappa_z \alpha_{i,z} \beta_{j,z} \right) - \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{z=1}^{k} \kappa_z \alpha_{i,z} \beta_{j,z} \tag{2}$$

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:

$$\begin{aligned} \mathcal{L}(\mathbf{A}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{q}) &\overset{\text{df}}{=} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{z=1}^{k} \left(A_{i,j} q_{i,j}(z) \log \left(\frac{\kappa_z \alpha_{i,z} \beta_{j,z}}{q_{i,j}(z)} \right) - \kappa_z \alpha_{i,z} \beta_{j,z} \right) \\ &\leq \log \mathbf{P}(\mathbf{A} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}). \end{aligned} \tag{3}$$

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

$$q_{i,j}(z) = \frac{\kappa_z \alpha_{i,z} \beta_{j,z}}{\sum_{z=1}^{k} \kappa_z \alpha_{i,z} \beta_{j,z}}. \tag{4}$$

It is also straightforward to differentiate with respect to κ_z, obtaining the update

$$\kappa_z = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} q_{i,j}(z)}{\sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{i,z} \beta_{j,z}}. \tag{5}$$

For the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ 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

$$\alpha_{i,z} = \frac{\sum_{j=1}^n A_{i,j} q_{i,j}(z)}{\lambda_z + \sum_{j=1}^n \kappa_z \beta_{j,z}}, \tag{6}$$

$$\beta_{j,z} = \frac{\sum_{i=1}^m A_{i,j} q_{i,j}(z)}{\mu_z + \sum_{i=1}^m \kappa_z \alpha_{i,z}}, \tag{7}$$

but now we see that to obtain the desired normalization we must in fact take λ_z and μ_z such that they leave us with the simpler

$$\alpha_{i,z} = \frac{\sum_{j=1}^n A_{i,j} q_{i,j}(z)}{\sum_{i=1}^m \sum_{j=1}^n A_{i,j} q_{i,j}(z)}, \tag{8}$$

$$\beta_{j,z} = \frac{\sum_{i=1}^m A_{i,j} q_{i,j}(z)}{\sum_{i=1}^m \sum_{j=1}^n A_{i,j} q_{i,j}(z)}. \tag{9}$$

In particular, these update equations 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. This simplifies the EM calculation substantially compared to the nested coordinate ascents that are usually involved in mixed-membership stochastic block models.

B.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 $1, \dots, m$ and n politicians indexed $1, \dots, n$, 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_j^P interest groups. Note that the total number of edges E is given by the sums of either of these quantities:

$$E = \sum_{i=1}^m d_i^I = \sum_{j=1}^n d_j^P. \tag{10}$$

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 $e_{i,1}^I, \dots, e_{i,d_i^I}^I$. Likewise, from each politician j, produce stubs $e_{j,1}^P, \dots, e_{j,d_j^P}^P$.

2. Generate a uniformly random matching of the two sets $\{e_{i,k}^I\}$ and $\{e_{j,\ell}^P\}$ (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, ℓ in the previous step (so that the total number of edges is the total number of pairs k, ℓ 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 113th Congress dataset here. In particular, in Figure F.2, 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; see Appendix D.1 for further details).

B.3 Temporal Analysis

We present some preliminary results on information the biLCM can provide about how legislation community memberships vary over time. While in the main text we focused on analyzing the 113th Congress only, in Figure F.3 we plot legislation community entropies of several prominent legislators and interest groups for each Congress between the 110th and 114th (spanning from 2007 to 2016). Specifically, we apply the filtering procedure described in Appendix A.2 to the interaction data of each Congress, infer the biLCM with $k = 8$ on the resulting dataset, and compute entropies for several actors that are sufficiently active in all five sessions.

For legislators, we find that trends in legislation community entropy reflect gains or losses of legislative power. For instance, Senator Harry Reid (D-NV) dropped in entropy in the 114th Congress when he lost the position of Senate Majority Leader. Similarly, Senator Barbara Boxer

(D-CA) dropped in entropy in the 113th and 114th Congresses when she lost her positions as chair of the Senate Environment Committee and Senate Ethics Committee and retired at the end of the 114th Congress. Conversely, Senator John Cornyn (R-TX) gained in entropy starting from the 112th Congress when he was elected Senate Minority Whip, and continued to gain entropy in the 113th Congress when he was elected Senate Majority Whip.

For interest groups, in particular firms, trends in entropy appear instead to reflect business success. For instance, the entropy of Motorola steadily dropped over this period, as it was sold to Google in 2012 and again to Lenovo in 2014. The entropy of Arch Coal likewise dropped as it sold its Canyon Fuel Company LLC subsidiary and filed for bankruptcy protection, but rose again as it emerged from bankruptcy protection during the 114th Congress. The entropy of Merck dropped as well as it sold a component of its business to Bayer and underwent a merger in the period leading up to the 112th Congress (the narrowing of its business is starkly visible in its entropy dropping to zero, indicating membership in only one legislation community), but increased again subsequently.

C LSM Additional Information

C.1 Parameter Identification

Several transformations of the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}_i, \boldsymbol{\psi}_j$ 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 $\boldsymbol{S} \in \mathbb{R}^{d\times d}$, each $\boldsymbol{\theta}_i$ may be multiplied by $\boldsymbol{S}$ and each $\boldsymbol{\psi}_j$ by $\boldsymbol{S}^{-1}$; *rotation*, where for any orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ each $\boldsymbol{\theta}_i$ and $\boldsymbol{\psi}_j$ may be multiplied by $\boldsymbol{Q}$; *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 $\mathbf{v} \in \mathbb{R}^d$ may be added to each $\boldsymbol{\theta}_i$ (resp. $\boldsymbol{\psi}_j$) and $\mathbf{v}^\top \boldsymbol{\psi}_j$ (resp. $\mathbf{v}^\top \boldsymbol{\theta}_i$) 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 (6). 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 $\boldsymbol{\theta}_i$ and $\boldsymbol{\psi}_j$ a normal prior with mean zero. The scaling invariance is partly resolved by further setting the covariance matrices of the $\boldsymbol{\theta}_i$ and $\boldsymbol{\psi}_j$ to be equal; the only scaling transformations still admissible are those where $\boldsymbol{S}$ 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 $\boldsymbol{S}$ is a permutation matrix, a diagonal matrix with diagonal entries ± 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 and eight in the two-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 $\widehat{R}$ statistics smaller than 1.01 for all parameters (as reported by `Stan` diagnostic logging) for both $d = 1$ and $d = 2$, and Pearson's ρ 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.

C.2 Dimensionality Selection

We present some heuristics on choosing a dimensionality for the LSM 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 $\mathbf{A}$, 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 $\mathbf{A}$, 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 $\mathbf{A}$, and thus also a heuristic estimate of the dimensionality to take in the LSM. In Figure F.4, we show that a dimensionality of two is a reasonable choice if this heuristic is admitted.

C.3 Computation

The `Stan` code in Listing F.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. 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 our models parameters as used in all visualizations in the main text.

C.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 $\mathbf{X} \in \mathbf{R}^{d\times N}$. We let $\boldsymbol{\mu} \in \mathbf{R}^d$ be the vector of means of the rows of $\mathbf{X}$, and let $\mathbf{X}_0$ be the centered matrix obtained by subtracting μ_i from each entry of the ith row of $\mathbf{X}$. Then, $\mathbf{W} = \frac{1}{N-1}\mathbf{X}_0\mathbf{X}_0^\top \in \mathbf{R}^{d\times d}$ is the empirical covariance matrix of the position samples, which we diagonalize to obtain orthonormal eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_d$ and associated non-negative eigenvalues $\lambda_1, \dots, \lambda_d$. 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: λ_i is the amount of variance of the sampled positions in the direction $\mathbf{v}_i$, and the shape of the set of sampled positions is approximately the ellipsoid with axes in the directions of the $\mathbf{v}_i$ 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 LSM with $d = 2$ in Figure F.6.

Popularity Factor Variances For popularity factors α_i and β_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 coefficient of variation (standard deviation divided by the mean) of each of these quantities. As for the position variances, we visualize all of the coefficients of variation at once, giving in Figure F.7 their histograms for our main dataset. We observe that the distributions of the coefficients of variation are strongly concentrated on the low values; manual inspection reveals that the agents with the highest coefficients 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 coefficient of variation).

Trace Plots In Figure F.9, 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.

C.5 Additional Analysis

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 by nesting the former in the latter. Table F.2 presents the F-statistics and p-values summarizing the significance of including various sets of observable covariates. We find that committee memberships are the only significantly informative covariates of politicians' positions in both the one- and two-dimensional models. 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 affects only one dimension of the two-dimensional LSM significantly..[1]

[1] We thus find evidence that one dimension of the two-dimensional LSM 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

D biSBM Additional Information

D.1 Degree Correction

The biSBM without degree correction is defined by the following adjusted version of Equation (7), obtained by eliminating the popularity factors α_i and β_j:

$$\mathbf{P}(\mathbf{A} \mid \mathbf{B}, \boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{m} \prod_{j=1}^{n} \text{Poisson}(A_{i,j} \mid B_{x_i,y_j}). \tag{11}$$

This simplified model 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 observe in the lobbying network. In our case, the model groups together politicians who sponsor larger numbers of lobbied bills and interest groups who lobby many bills, grouping actors by their overall level of activity rather than the specific legislation they tend to be involved with.

To correct for this, in the main text we follow Larremore, Clauset, and Jacobs (2014) and introduce popularity factors α_i and β_j, and adjust the model to

$$\mathbf{P}(\mathbf{A} \mid \mathbf{B}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{m} \prod_{j=1}^{n} \text{Poisson}(A_{i,j} \mid \alpha_i \beta_j B_{x_i,y_j}). \tag{12}$$

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

$$\hat{\alpha}_i = \frac{\deg(i)}{\sum_{x_{i'}=x_i} \deg(i')}, \;\; \hat{\beta}_j = \frac{\deg(j)}{\sum_{y_{j'}=y_j} \deg(j')}, \tag{13}$$

the fractions of the edges leaving a node's community 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 communities than the biSBM. We discuss this improvement further in Appendix E.

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, 2019). We leave the challenging task of adapting such methods to our setting for future research.

D.2 Assortativity

In Figure F.13, 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*.

E Model Comparison

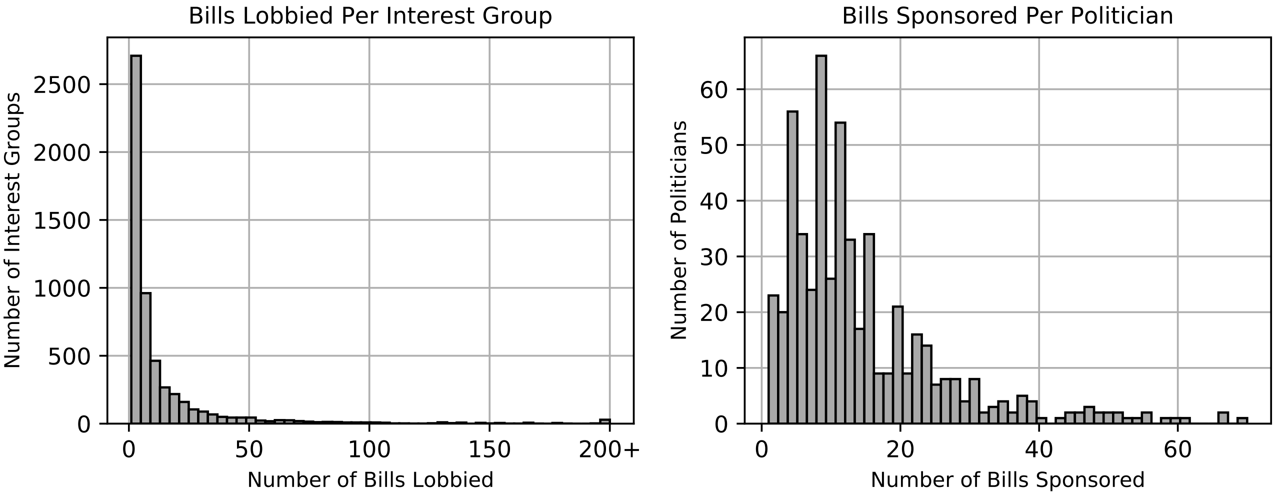
We give a few points of comparison among our latent space and community models, confirming that the models all consistently reflect the same underlying properties of the lobbying network.

Figure F.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 LSM.

Next, recall that in the LSM 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 LSM respect the groupings 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 Figures F.11 and F.12 for the biSBM and dc-biSBM, respectively, observing that the latter is better aligned with the LSM 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 F.14, 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.

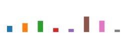
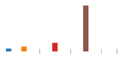
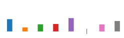
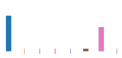
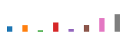
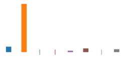
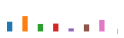
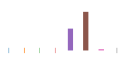
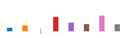
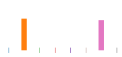
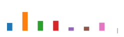
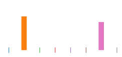
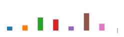
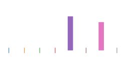
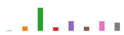
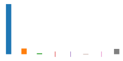
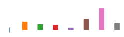
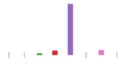
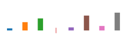
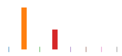
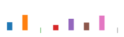
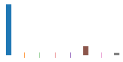
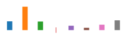
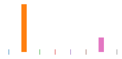
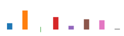
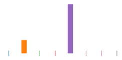
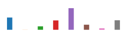
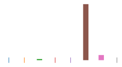
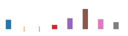
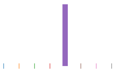
F Supplementary Figures



Interest Group	# Lobbied
Chamber of Commerce	908
Iraq/Afghanistan Veterans of America	719
Healthcare Leadership Council	385
National Cable and Telecommunications Association	350
Xcel Energy	329
Midamerican Energy Holdings	327
American Federation of State, County, and Municipal Employees	309
Duke Energy	293
Trust for America's Health	293
Specialty Equipment Market Association (SEMA)	291

Politician	# Sponsored
Robert Menéndez (D-NJ)	70
David Vitter (R-LA)	67
Mark Begich (D-AK)	66
Amy Klobuchar (D-MN)	61
Bernard Sanders (I-VT)	59
Ron Wyden (D-OR)	58
Dianne Feinstein (D-CA)	55
Sherrod Brown (D-OH)	55
Kirsten Gillibrand (D-NY)	54
Edward Markey (D-MA)	53

Figure F.1: **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.

Highest Entropy		**Lowest Entropy**	
Senator	**Distribution**	**Senator**	**Distribution**
Harry Reid (D-NV)		Cory Gardner (R-CO)	
Jerry Moran (R-KS)		Maria Cantwell (D-WA)	
Patty Murray (D-WA)		Jeff Merkley (D-OR)	
Kay Hagan (D-NC)		Shelley Capito (R-WV)	
Bill Cassidy (R-LA)		Tom Cotton (R-AR)	
Chris Van Hollen (D-MD)		Tim Scott (R-SC)	
Amy Klobuchar (D-MN)		John Isakson (R-GA)	
Charles Schumer (D-NY)		Richard Burr (R-NC)	
Debbie Stabenow (D-MI)		Bob Corker (R-TN)	
Dianne Feinstein (D-CA)		James Risch (R-ID)	
Robert Portman (R-OH)		Martin Heinrich (D-NM)	
Kirsten Gillibrand (D-NY)		Michael Crapo (R-ID)	
Robert Casey (D-PA)		Angus King (I-ME)	
Sherrod Brown (D-OH)		John Hoeven (R-ND)	
Mark Warner (D-VA)		Richard Shelby (R-AL)	

🟦 Healthcare	🟧 Veterans' Affairs	🟩 Technology & Telecom.	🟥 Civil Society
🟪 Retail & Transport.	🟫 Finance & Insurance	■ Energy	■ Universities & Research

Table F.1: **Legislation Community Distribution Examples: Politicians.** We show the legislation community memberships of senators having the most (left panel) and least (right panel) memberships, as quantified by the entropy of the legislation community distribution. As in the table of Figure 3, the distributions are represented as histograms, with eight bars corresponding to the eight legislation communities in the biLCM.

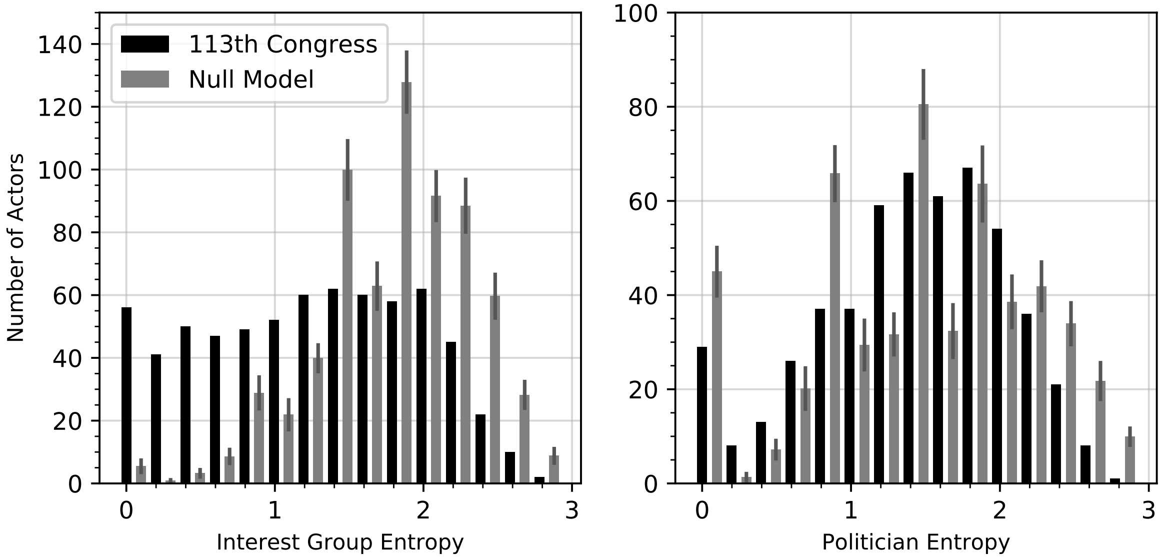


Figure F.2: **Link Community Entropy Distribution vs. Null Model.** We plot the distribution of entropies of link community distributions obtained by the biLCM run on the 113th Congress dataset, compared to an average (per histogram bin) of the same distribution for several draws of the null configuration 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.

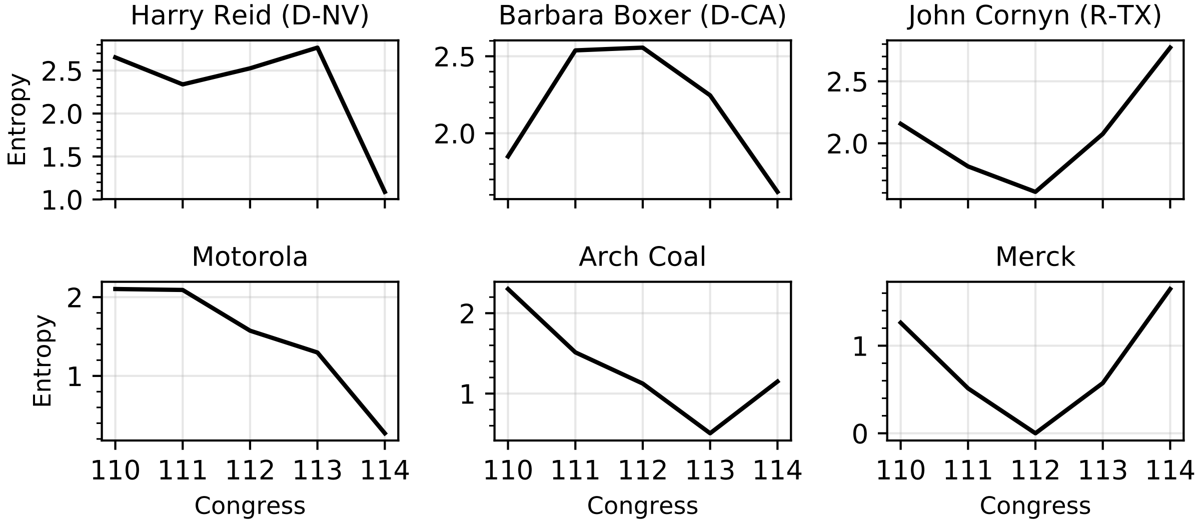


Figure F.3: **Link Community Entropies Over Time.** We plot the legislation community entropy of several legislators and interest groups in biLCM models inferred from data for Congresses between the 110th and 114th. (All datasets are formed with the same filtering procedure as used for the 113th Congress dataset discussed in the main text; see Appendix A.2.)

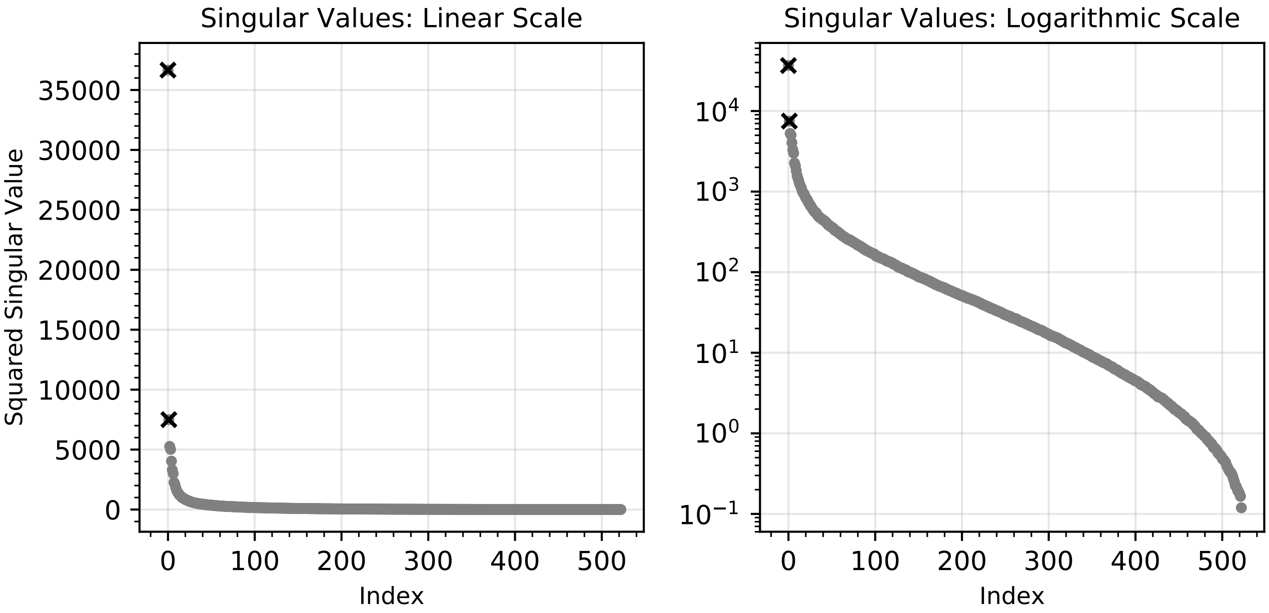


Figure F.4: **Bipartite Adjacency Matrix Singular Value Decay.** We plot the singular values of the interaction matrix **A**, highlighting that the singular values decay rapidly and the first few appear to capture much of the total weight of this matrix. The first two singular values are highlighted, corresponding to our analysis primarily of a two-dimensional latent space model in the main text.

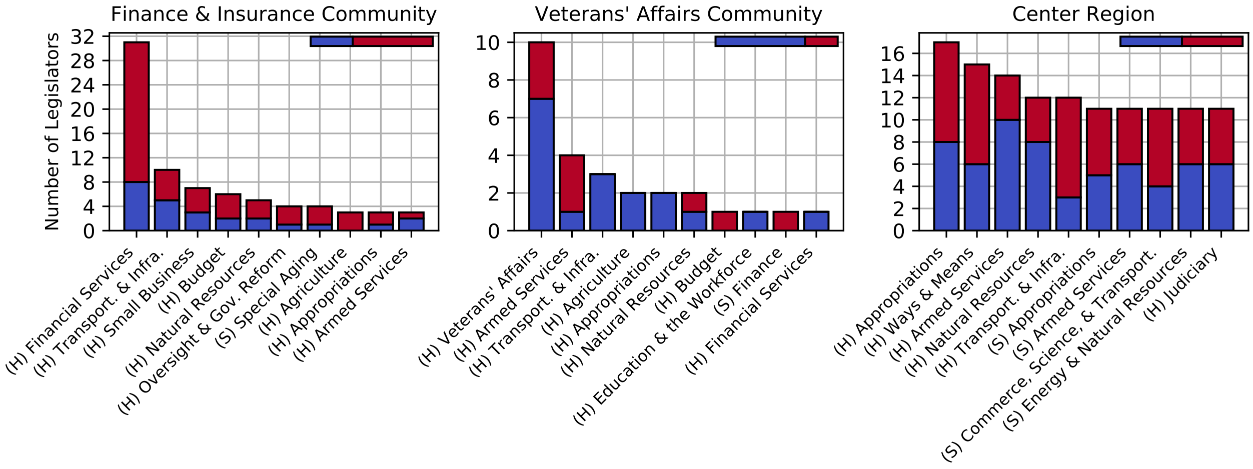


Figure F.5: **Committee Membership of Politicians by Cluster.** We present histograms of the top 10 committee memberships for politicians in the larger clusters highlighted in Figure 5. Bars are divided by political party (red for Republican and blue for Democrat), and the horizontal bars at the top show the overall party distribution of each cluster. Senate committee and House committee labels are prefixed with (S) and (H), respectively.

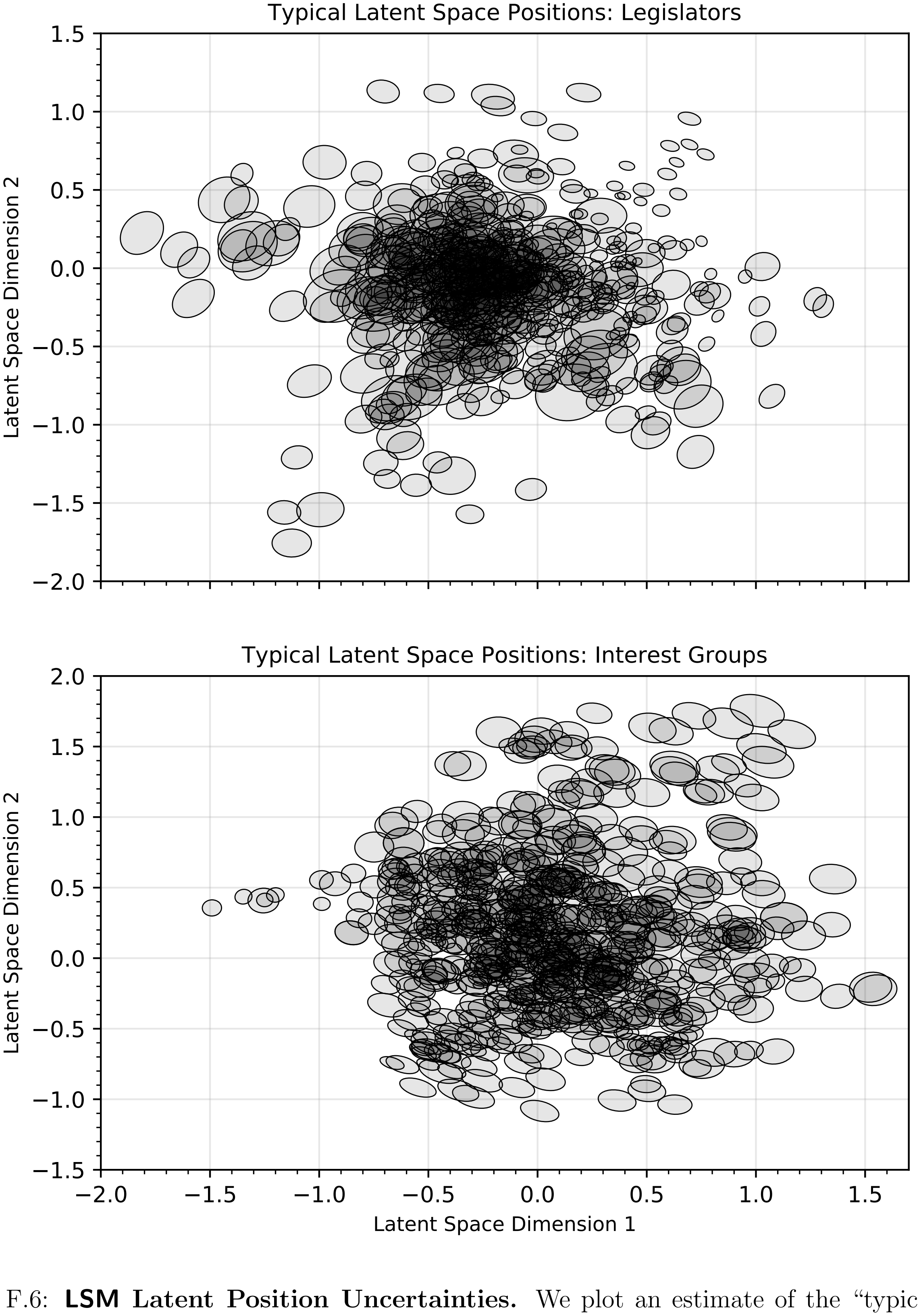
Legislators

Independent Variables	**Restricted**	**Full**	**Restricted**	**Full**	**Restricted**	**Full**
Chamber	✓	✓	✓	✓	✓	✓
Gender	✓	✓	✓	✓	✓	✓
Party	✓	✓	✓	✓	✓	✓
Committees		✓	✓	✓	✓	✓
State	✓	✓		✓	✓	✓
DW-NOMINATE	✓	✓	✓	✓		✓
Dependent Variable	**F-Statistic (p-Value)**		**F-Statistic (p-Value)**		**F-Statistic (p-Value)**	
1D LSM	2.96 (2.80e−9)		1.35 (0.06)		0.85 (0.43)	
2D LSM, Dimension 1	3.76 (1.23e−13)		1.19 (0.18)		0.24 (0.78)	
2D LSM, Dimension 2	3.13 (3.21e−10)		1.10 (0.30)		0.41 (0.67)	

Interest Groups

Independent Variables	**Restricted**	**Full**	**Restricted**	**Full**
Industry		✓	✓	✓
State	✓	✓		✓
Dependent Variable	**F-Statistic (p-Value)**		**F-Statistic (p-Value)**	
1D LSM	13.72 ($<$ 2.2e−16)		0.86 (0.71)	
2D LSM, Dimension 1	9.92 ($<$ 2.2e−16)		0.81 (0.78)	
2D LSM, Dimension 2	13.25 ($<$ 2.2e−16)		1.65 (0.01)	

Table F.2: **Regression Analysis of LSM Covariates.** We present statistics of linear models of covariates against the one- and two-dimensional LSM latent dimensions for legislators and interest groups. We compare full linear models with all available covariates against restricted models, omitting either committee covariates, geographic information, or ideology for legislators, and either industry classification or geographic information for interest groups. In the bottom three rows of each table, for each of the three latent dimensions, we give the F statistics and p-values of the comparison F-test. For legislators, the committee membership covariates are the only covariates with significant explanatory power for the LSM latent space organization. For interest groups, the industry covariates are significantly explanatory of the LSM latent space organization, and, with lesser confidence, geographic covariates are also explanatory of one dimension in the two-dimensional LSM.



Typical Latent Space Positions: Legislators

Typical Latent Space Positions: Interest Groups

Figure F.6: **LSM Latent Position Uncertainties.** We plot an estimate of the "typical set" of each actor's latent position in the two-dimensional LSM, 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 C.4.

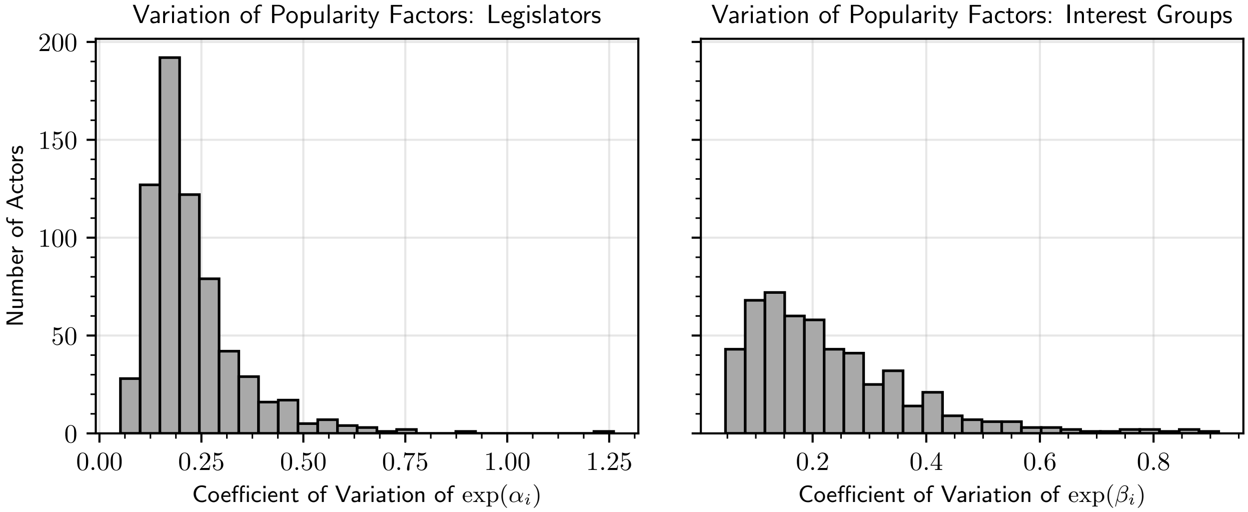


Figure F.7: **LSM Popularity Factor Uncertainties.** We plot distributions of the coefficients of variation (standard deviation as a fraction of the mean) of the sampled popularity factors for all interest groups and all politicians.

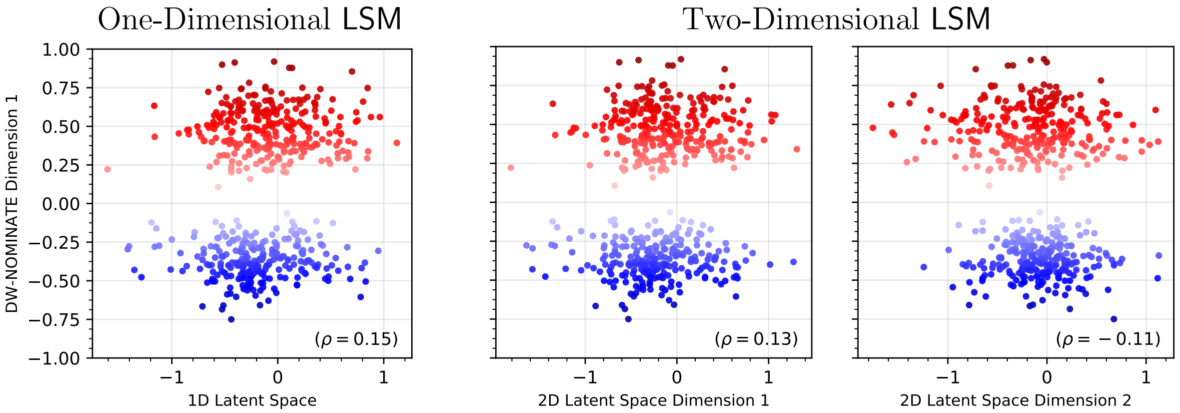


Figure F.8: **DW-NOMINATE Ideology vs. LSM.** We illustrate the lack of correlation between politicians' latent space positions obtained from one- and two-dimensional LSM estimates and the DW-NOMINATE ideology dimension. Points are colored on a linear scale from blue to red according to the DW-NOMINATE ideology dimension.

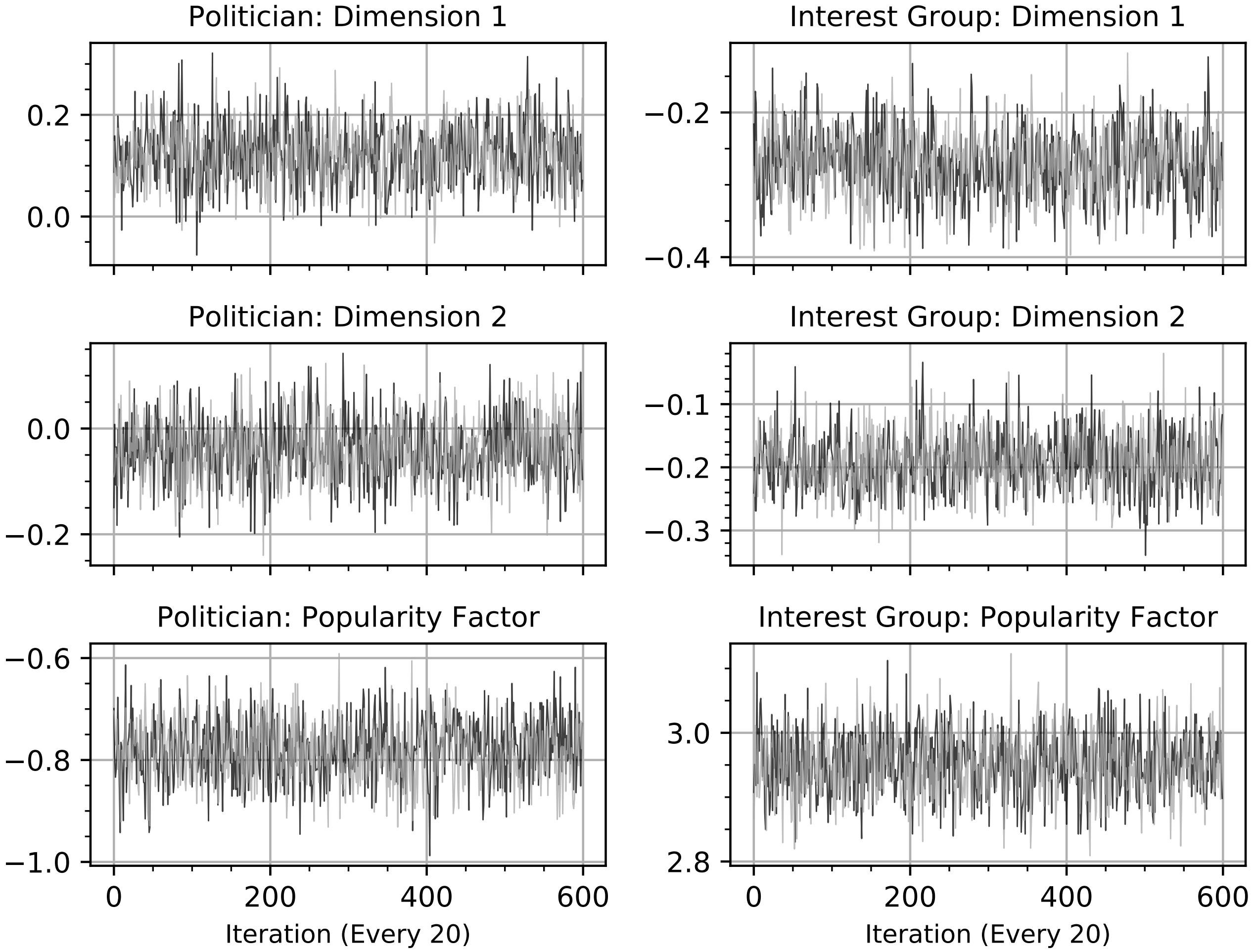
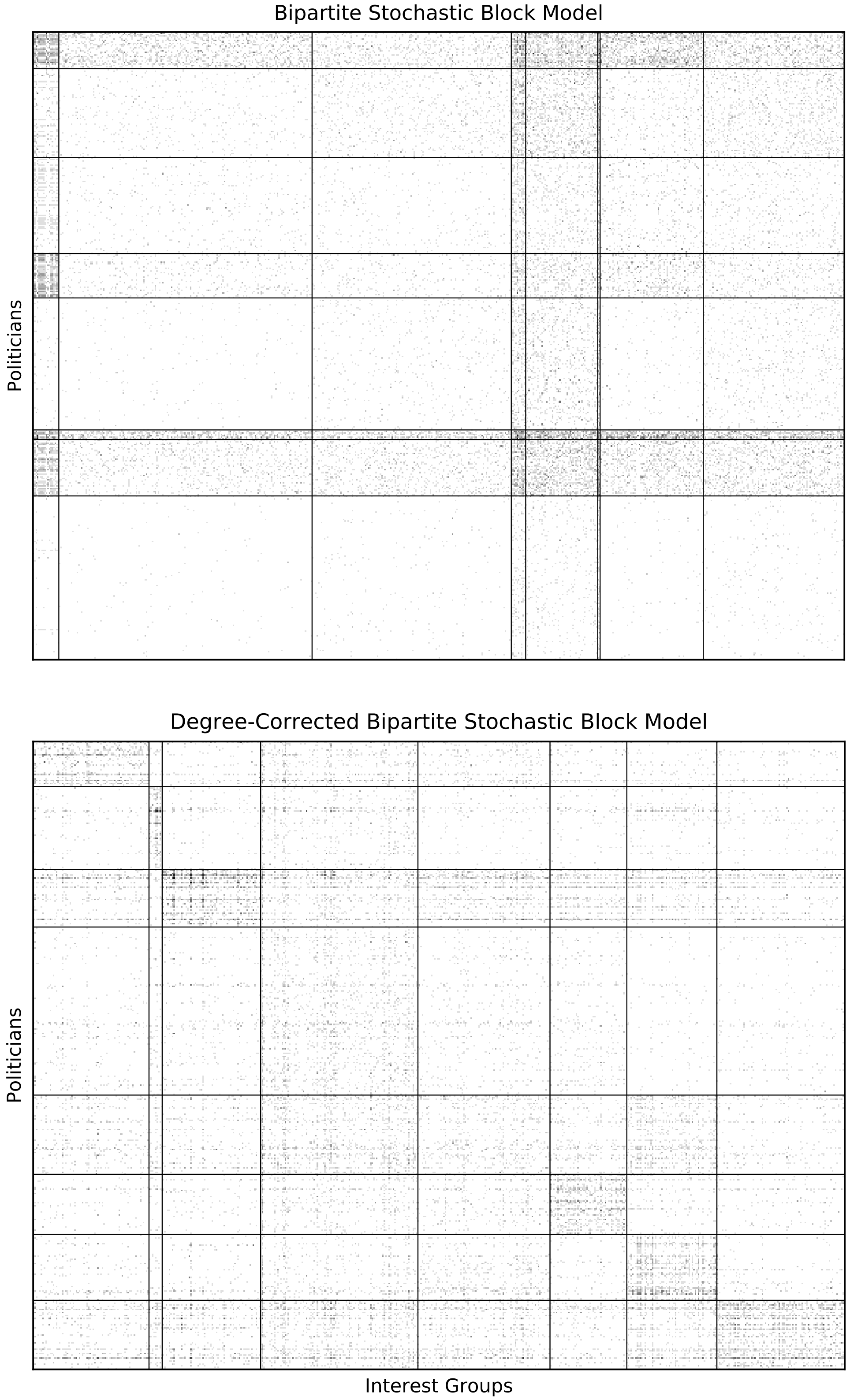


Figure F.9: **LSM Trace Plots.** We show trace plots for the LSM model parameters (two latent space dimensions and one popularity factor) drawn from two (out of four) MCMC chains following the warmup period, for one interest group, the Chamber of Commerce, and one politician, Senator Barbara Boxer (D-CA).

```
// Latent space network model.
//
// Implements the model
//   A_{ij} ~ 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


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 F.1: We provide Stan code defining a sampler for the posterior distribution of the LSM. 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.



Bipartite Stochastic Block Model

Degree-Corrected Bipartite Stochastic Block Model

Figure F.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 communities that each model infers. Both models identify community structure, and, as expected, the dc-biSBM infers more balanced community sizes.

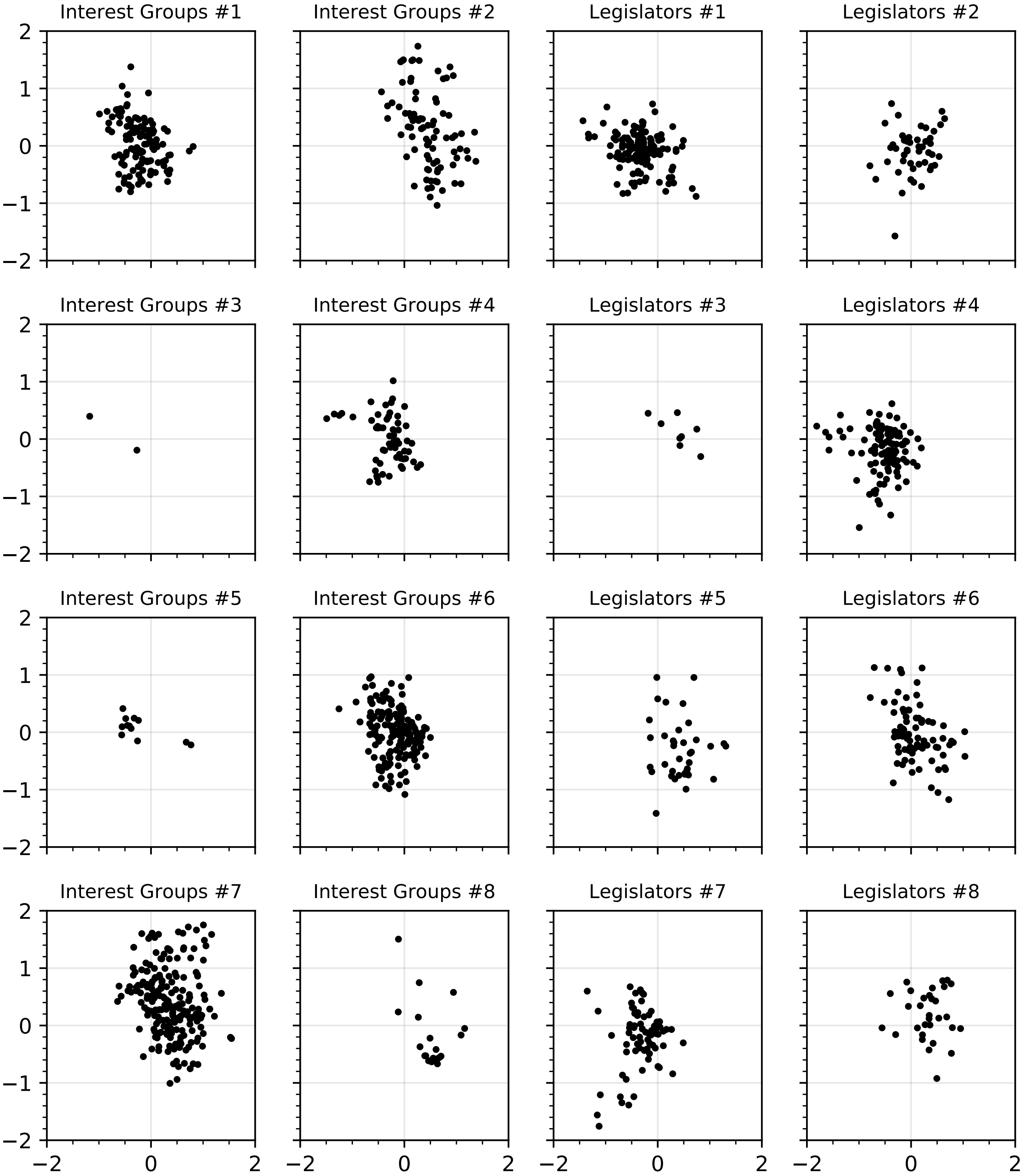


Figure F.11: **LSM vs. biSBM.** We plot the latent positions of legislators and interest groups from the LSM, split by their memberships in communities from the biSBM. For the sake of visual clarity, we no longer vary the sizes of the plotted latent space positions based on popularity factors. Localization of groups in the latent space suggests that the LSM is somewhat consistent with the biSBM, but much of the fine-grained interest group clustering we observe in Figure 5 is not reflected in the biSBM results.

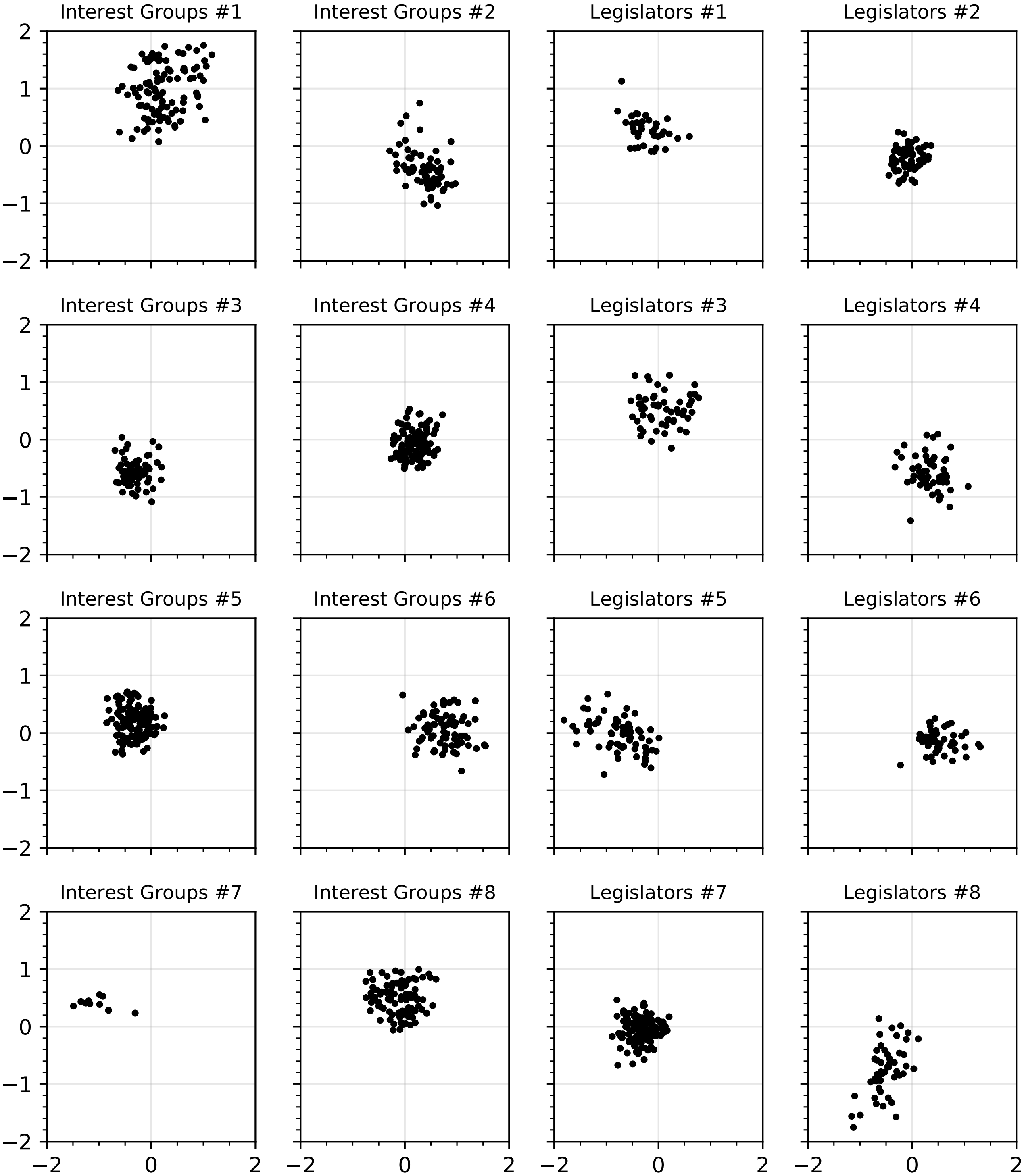


Figure F.12: **LSM vs. dc-biSBM.** We plot the latent positions of legislators and interest groups from the LSM, split by their memberships in communities from the dc-biSBM, with both run on the main dataset. Almost all communities found by dc-biSBM are closely localized in the latent space, showing that the dc-biSBM captures much of the same information as the LSM.

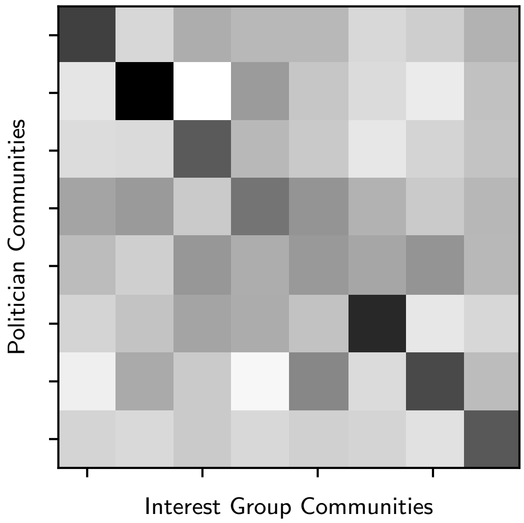


Figure F.13: **Assortativity in dc-biSBM.** We plot the mean interactions between each pair of communities in the dc-biSBM. If the communities are suitably permuted, we observe that the diagonal entries of the resulting matrix are typically the largest, indicating that the model finds an assortative community structure.

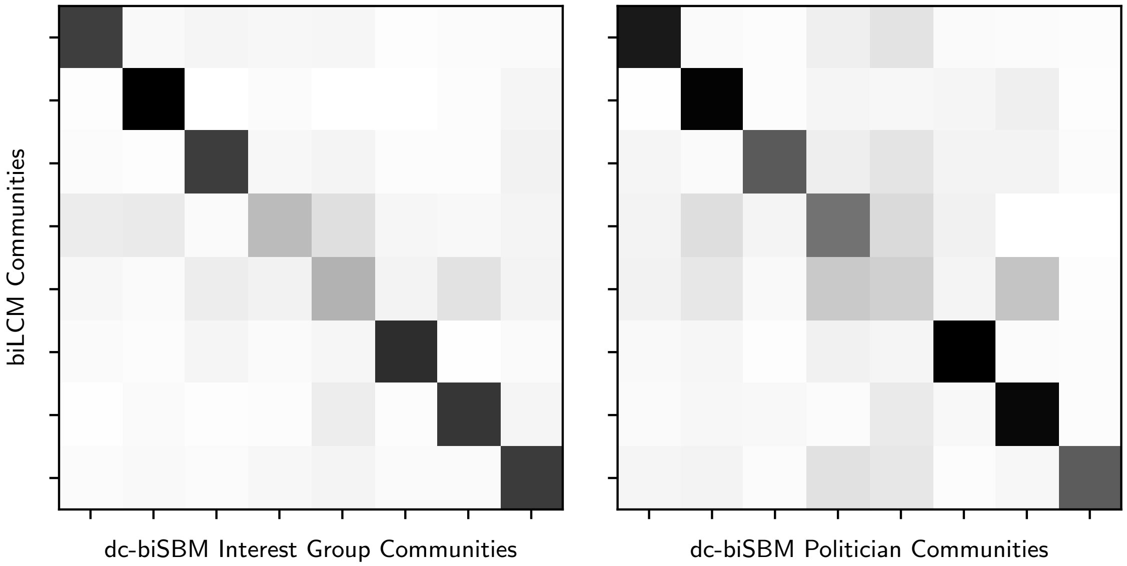


Figure F.14: **dc-biSBM vs. biLCM.** We plot communities of the dc-biSBM against those of the biLCM, measuring pairwise alignment by the mean probability that an actor in the given dc-biSBM community has of interacting in the given biLCM community. We observe that, when community labels are suitably permuted, it is possible to match communities from either model into well-aligned pairs.

References

Bafumi, Joseph, Andrew Gelman, David K Park, and Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13: 171–187.

Bailey, Michael, and David W Brady. 1998. "Heterogeneity and Representation: The Senate and Free Trade." *American Journal of Political Science* 42 (2): 524–544.

Jackman, Simon. 2001. "Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking." *Political Analysis* 9 (3): 227–241.

Karrer, Brian, and M.E.J. Newman. 2011. "Stochastic Blockmodels and Community Structure in Networks." *Physical Review E* 83.

Kluger, Yuval, Ronen Basri, Joseph T. Chang, and Mark Gerstein. 2003. "Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions." *Genome Research* 13 (4): 703–716.

Kubinec, Robert. 2018. "Generalized Ideal Point Models for Time-Varying and Missing-Data Inference.". Working Paper.

Larremore, Daniel B., Aaron Clauset, and Abigail Z. Jacobs. 2014. "Efficiently Inferring Community Structure in Bipartite Networks." *Physical Review E* 90.

Minhas, Shahryar, Peter D. Hoff, and Michael D. Ward. 2019. "Inferential Approaches for Network Analysis: AMEN for Latent Factor Models." *Political Analysis* 27 (2): 208–222.