



ORIGINAL ARTICLE

CHORUS: A New Dataset of State Interest Group Policy Positions in the United States

Galen Hall¹ , Joshua A. Basseches² , Rebecca Bromley-Trujillo³ and Trevor Culhane⁴ 

¹Department of Sociology, University of Michigan, Ann Arbor, MI, USA; ²Tulane University, New Orleans, LA, USA; ³Department of Political Science, Christopher Newport University, Newport News, VA, USA and ⁴Brown University, Providence, RI, USA

Corresponding author: Galen Hall; Email: galenh@umich.edu

(Received 11 February 2023; revised 27 June 2023; accepted 06 July 2023)

Abstract

Research on the activities and influence of interest groups in state legislatures faces a data problem: we are missing a comprehensive, systematic dataset of interest groups' policy preferences on state legislation. We address this gap by introducing the Dataset on Policy Choice and Organizational Representation in the United States (CHORUS). This dataset compiles over 13 million policy positions stated by tens of thousands of interest groups and individuals on bills in 17 state legislatures over the past 25 years. We describe the process used to construct CHORUS and present a new network science technique for analyzing policy position data from interest groups: the layered stochastic block model, which groups similar interest groups and bills together, respectively, based on patterns in the policy positions. Through two demonstrative applications, we show the utility of these data, combined with our novel analytical approach, for understanding interest group configurations in different state legislatures and policy areas.

Keywords: interest groups; lobbying; comparative legislatures; policy process; environmental policy; network analysis

Introduction

Interest groups are highly active in state politics (Gray and Lowery 1996; Holyoke and Cummins 2020; Lowery, Gray, and Cluverius 2015), and evidence suggests that most interest groups view “policy as prize,” and that their *raison d'être* is to shape public policy in accordance with their preferences (Hacker and Pierson 2014). Qualitative studies have illuminated important contests between organized interests at the state level using case studies from particular policy domains such as energy policy (e.g., Basseches 2023; Basseches et al. 2022; Stokes 2020). But research on the activities and preferences of interest groups in state politics and policymaking faces a data problem

(Anzia 2019). We lack comprehensive data on interest groups' policy positions across most state governments, and consequently, our understanding of coalitional politics at the state level remains piecemeal.

This paper introduces an original and publicly available dataset: the Dataset on Policy Choice and Organizational Representation in the United States (CHORUS). This dataset allows for a systematic comparison of interest groups' policy preferences in different political contexts across 17 states. Within each state, we report all available *policy positions* taken by organized interest groups on individual bills, as reported in lobbying and testimony disclosures over the past four to 25 years, depending on the state. We use these position data to derive inductively determined *policy coalitions* (sets of interest groups with similar policy preferences) and *policy issue spaces* (sets of bills with similar policy themes), which we provide as new units of analysis for the study of organized interest groups' legislative activities.

The 17 states included in this dataset encompass a wide range of political and economic contexts, which will allow scholars to test a vast array of research questions that leverage these differences. More specifically, these states exhibit considerable variation in party control of state legislatures and governorships, legislative professionalization (Squire 2017), and industrial composition. While others have used lobbying positions or similar data from a small handful of state legislatures to study interest group influence and polarization (Butler and Miller 2022; Kroeger 2022; Thieme 2019, 2021), our data encompasses considerably more states and allows scholars to use policy preferences to map coalitional lobbying formations in state capitals.

This paper proceeds as follows: We begin by discussing key bodies of literature on national and state lobbying, identifying ways in which the original dataset presented here can contribute to and build on this work. Next, we present our original dataset, describe the data collection process, and provide summary statistics for the full dataset. We then demonstrate the utility of these data in two key ways. First, we map interest group coalitions in a single state to demonstrate the depth of information provided by this dataset. We show that coalitions identified inductively from policy position data often map onto industry sectors, but with important exceptions, which provides information on the structure of policy preferences that might otherwise be lost. Similarly, the clustering algorithm we apply reveals hierarchical levels of bills centered on different substantive issues, which can provide researchers with a new way of organizing legislative datasets.

Second, we map a subset of interest groups lobbying around renewable energy and climate policy in a set of states with different energy production economies. This application serves to demonstrate how our dataset can be used by other scholars, while also providing insight into interest group alignments around state renewable energy legislation, which is widespread and will be critical to addressing climate change (Bromley-Trujillo *et al.* 2016; Stokes 2020). We make basic observations about the coalitional structures and end with several suggestions for further research based on this approach.

Relevant literature

As Baumgartner and Leech (1998) noted more than two decades ago, interest group literature has been frustrated by basic challenges of conceptualization and

measurement, which would be prerequisites for assessing policy influence and answering a host of questions related to interest group dynamics. They call for “large-scale work” (Baumgartner and Leech 1998, 12) to assess interest group activities in Washington. Anzia (2019) extends a similar call for work at the understudied state and local levels, which she argues could revitalize “policy-focused” interest group research. In recent years, the already large interest group literature has exploded, with vibrant subfields and methodological advances that we cannot hope to cover here (but see Hojnacki et al. 2012). Instead, we will focus on a set of issues affecting the field that we believe have not been adequately resolved. These problems are multileveled, ranging from foundational issues with defining and collecting appropriate data on interest group activities and preferences to higher-level issues such as identifying the scientifically appropriate categorizations to apply to that data, all of which affect the prospects for interest group research to be able to predict and understand influence in the policymaking process.

Data challenge: The problem of missing preferences

The most foundational data challenge has to do with defining and obtaining systematic data on interest group policy preferences. Interest group research would benefit from a standardized way of defining the concepts and categories it seeks to model and providing researchers a clear way to measure those quantities. The most basic fact in the field – so basic it is almost never stated – is that *interest groups* have *preferences over policies*. So the field should be able to answer these questions: 1. What is an interest group?; 2. What are interest groups’ preferences on policies; and 3. How does a researcher gather data on them?

While scholars lack consensus around a single definition of an interest group, many agree that interest groups are organizations that aim to influence public policy (e.g., Anzia 2019; Baumgartner et al. 2009; Bawn et al. 2012; Hacker and Pierson 2014). This means they presumably have policy preferences, but until recently, we have lacked a way to conceptualize and systematically measure these preferences in the American states. Without standardized and comprehensive data sources, prominent interest group research generally uses qualitative case studies (e.g., Stokes 2020), or systematic data collection in a particular policy area (e.g., Culhane, Hall, and Roberts 2021). In cases where comprehensive datasets of policy preferences were created by hand, which to the best of our knowledge has been limited to studies in Washington, D.C. (e.g., Baumgartner et al. 2009), extension to additional years and issues presents significant challenges.

While descriptively rich and extremely generative for theory-building, these studies lacked consistency in the ways they define and measure basic concepts such as “policy preferences.” This situation began to change, at least at the federal level, as scholars began to converge around data generated by lobbying disclosure laws, predominantly the federal Lobbying Disclosure Act of 1995, to define a universe of interest groups (Kim 2018; LaPira and Thomas 2020). The LDA provides a standard source of systematic data on interest groups (as well as various traits about them, such as lobbying spending), and some argued that the issues lobbyists disclosed working on could also provide a standard data source for policies (Baumgartner and Leech 2001).

But several problems remained after the LDA: first, the law did not encompass state legislatures, which are a key venue for interest group activity (Anzia 2019), and have been for a long time (McConnell 1966). In numerous publications, Gray and Lowery (1993, 1996) address this gap by collecting and analyzing an impressive body of data on state-level interest group populations, demonstrating their diversity and density. Recent work has also established automated methods to measure state interest group populations (Garlick and Cluverius 2020). These studies have enabled scholars to consider a range of questions regarding the population ecology of state-level interest groups, including interest group mortality (Nownes and Lipinski 2005), the relationship between interest group density and strategies employed by groups (Lowery et al. 2009), and their association with healthcare policy outcomes (Gray, Lowery, and Benz 2013). Nevertheless, these studies do not provide systematic data on interest group preferences for and against specific policies (bills, in the legislative context).

Second, while the LDA (and several state equivalents) mandated disclosure of “issues” lobbied on, the resulting data are much less well-defined than the data on interest groups themselves. An “issue” is an inherently fuzzy concept whose meaning may vary from one lobbyist’s report to the next. We return to this problem shortly – it is in some sense a higher-level conceptual problem rather than a data one – but note here that researchers can resolve it simply by looking one level lower: at interest group activity on individual *bills*, which, unlike issues, are a well-defined and bounded type of data. LDA disclosures increasingly report individual bills lobbied on (Kim and Kunisky 2021).

Third, and most importantly for our case, the LDA does not mandate disclosure of *preferences* on issues, and neither do most state-level lobbying disclosure laws. So, even though most interest groups experience opposition from others on most issues (Gilens and Page 2014; Givel and Glantz 2001; Salisbury et al. 1987), we lack the crucial data telling us which side a given group falls on. In its absence, scholars must resort to handcrafted surveys, interviews, and other qualitative methods to define and elucidate preferences (Baumgartner et al. 2009; Nownes and Freeman 1998). This gap – the near-total lack of standardized data on preferences – is a major hindrance to understanding how interest groups aim to influence legislative policy¹.

Conceptual challenge: Defining issues and interests

Once in hand, raw data are rarely useful by itself. We must then find patterns in the data that we can use to make predictions. Interest group scholars can benefit from defining agreed-upon concepts, latent in the observed data, which can enter into useful models of politics.² For example, the broader field of legislative studies employs ideal points as one such latent concept: they are inferred via standard methods from standard data sources (roll-call votes), have explanatory and predictive power, and

¹Recent studies have begun to leverage policy positions data from state lobbying disclosures and public statements to fill this gap (Butler and Miller 2022; Crosson, Furnas, and Lorenz 2020; Thieme 2019); we aim to extend these efforts.

²The alternative—treating each interest group and each policy as an independent unit of analysis—can yield useful results (e.g., Butler and Miller 2022), but it leaves us incapable of incorporating more complex structures in our models.

are critical to many models of legislative politics (Poole and Rosenthal 1991; Shor, Berry, and McCarty 2010; Shor and McCarty 2011).

Interest group scholars often focus on two types of latent variables corresponding to interest groups and policies.³ First, they apply categorizations to interest groups that they hope will capture key commonalities in the preferences of those groups. Industry categorizations such as SIC codes give one example (e.g., see their use in Box-Steffensmeier and Christenson 2014). Scholars often assume *a priori* that interest groups in the same industry or economic sector will fall on the same side of a given issue (Garlick and Cluverius 2020). Second, as opposed to studying individual policies, scholars often study interest group activity on “issues” which are taken to encompass many similar policies (Baumgartner et al. 2009; Gilens and Page 2014). Scholars have come up with many answers to this problem, from surveys (Nownes and Freeman 1998) to comprehensive coding schemes for Congressional legislation such as the Policy Agendas Project (Eissler and Jones 2019), although no standard scheme or data source exists for issues at the state level.

Crucially, for both interest groups and policies, scholars still rely on *a priori* categories rather than inductively defined ones. Industry and issue classifications are not chosen based on real-world observation of interest group preferences on bills, but instead because they capture our intuitive notions about what various types of interest groups *should* want from policy. Applying *a priori* industry classifications to interest groups yields categories which do not always predict their preferences over important policy domains such as trade policy (Kim 2017).

Recent work has advanced two different approaches to define empirically grounded latent traits for interest groups. Each captures one half of what we see as the ideal measure: a categorization of interest groups and policies, inferred inductively from preferences, which captures the most explanatorily useful divisions between groups/bills. First, At least one study has attempted to derive interest group categories inductively by applying a statistical model of lobbying to LDA disclosures; however, because of the missing preference data, it cannot capture the important divisions between camps of interest groups active on similar issues but toward opposing ends (Kim and Kunisky 2021).

Second, scholars have leveraged emerging data sources on interest group policy positions to jointly define ideal points for interest groups and politicians at both the national and state levels (Crosson, Furnas, and Lorenz 2020; Thieme 2018, 2019), or done the same using campaign contributions as a proxy for preferences (Bonica 2014), and recently combined campaign contributions and a network of co-signed Supreme Court briefs to estimate interest groups’ ideal points in the judicial domain (Abi-Hassan et al. 2023). These ideal points provide important insights into the interplay of interest groups, political parties, and polarization. However, they cannot wield the same predictive power for interest groups as they do for legislators.

To see why, consider that interest group policy positions, unlike legislators’ roll call votes, are characterized by extremely high nonresponse rates – most interest groups do not state positions on most policies. The ideal point model assumes that a legislator will vote for a bill based on that bill’s distance from the legislator in ideal

³Political scientists employ many other concepts and variables when studying interest groups, which we cannot cover here; we take “industries” and “issues” as our focus because they are two common but empirically ill-defined examples.

space (the closer they are, the more likely a *yea* vote). Yet a conservative interest group such as an antiabortion group may only have a position on a small number of bills, even though many more are located near their ideal point in one or two dimensional ideal space. Interest group studies require a measure that combines the best of both approaches: latent quantities grounded in policy preferences and designed for maximum predictive power. Here, we agree with Anzia (2019), who argues that “[I]f the ultimate goal were to study [interest group] influence, it would be more useful to categorize interest groups according to their policy goals – and to measure how actively they pursue those goals and with what resources.”

Prediction challenge: Understanding influence

With no such categorizations in hand, it seems natural that the field has faced serious problems pursuing that goal: understanding in a comprehensive, systematic way, how and why interest groups influence policy. While studies do examine interest group influence (Baumgartner et al. 2009; Gilens and Page 2014), much of this work centers on the US Congress, which limits the development of novel theories of policy making influence (Anzia 2019). It also lacks a common data source and quantifiable conceptual framework, and has – maybe for this reason – often arrived at contradictory conclusions about, for instance, whether better-resourced groups exert more policy influence. For example, Baumgartner et al. (2009) find little relationship between interest group resources and policy influence. By contrast, Gilens and Page (2014) find that interest group influence is unequal (“biased pluralism”), with business interests having greater say than mass-based, public interest groups. Conversely again, others find that state-level policy mostly does follow majority popular opinion, and that where it deviates, interest groups have little to do with it (Erikson, Wright and McIver 1993; Lax and Phillips 2012).

Perhaps in light of this confusion, many scholars have shifted their focus to study aspects of interest groups that do not directly depend on knowledge of their policy preferences (Hacker and Pierson 2014). Anzia’s (2019) critique of the extant literature aptly summarizes the theoretical opportunities and challenges associated with state-level interest group research. Anzia observes that the vast literature that has developed around the question of political representation “has become one largely about how well the positions of political elites align with citizens’ preference ... as though questions about interest group influence are somehow separate from studies of political representation” (Anzia 2019, p. 343).

Given the challenges presented in the literature above, we hope to advance the interest group literature in a few key ways. First, we offer a dataset that includes interest group preferences, which can be leveraged to answer numerous questions about interest group coalition activity and influence. Second, these data also allow scholars to use an inductive method to categorize interest groups based on their alignment of preferences around actual legislation. Finally, these data can be used to test research questions on organized interests that have, thus far, only been considered at the national level.

The CHORUS dataset

We present the Dataset on Policy Choice and Organizational Representation in the US (CHORUS) along with a novel application of an analytical technique from

network science as a means of addressing the problems identified in the literature. CHORUS contains millions of policy positions taken in lobbying and testimony by interest groups (and individuals) on bills across 17 state legislatures. It thereby provides a solution to the data problem: interest groups, policies, and preferences get operationalized as “organizations which lobbied or testified”, “the bills they lobbied on,” and “the positions they stated,” respectively. We then use stochastic block models (SBMs) to sort interest groups and bills into “blocks” whose members share similar patterns of lobbying positions – so that interest groups in the same block express similar preferences on bills, and bills in the same block receive support or opposition from the same interest groups. These block models answer the conceptual problem outlined above with an approach that unifies the best qualities of the two methods described above.

We begin by describing the steps taken to compile the CHORUS dataset and present the variables available to scholars for use in future research. We also link our dataset to other publicly available sources of data such as Legiscan and FollowTheMoney (FTM), which allows for rich exploration of research questions associated with interest group activities and the state policy process. We then describe the analytical method used to derive empirically grounded categorizations of bills and interest groups. We finish with applied examples showing the analytical possibilities these data open up.

Data collection

Policy position disclosure laws and practices vary across the states. The most transparent laws require interest groups to disclose the positions they took on every bill on which they or a contracted lobbying firm lobbied. Differences in norms and the stringency with which overseers enforce these laws mean that the quality and trustworthiness of these data vary by state. Lobbying positions are typically required to be reported on a twice-yearly or monthly basis.

Where states do not require lobbying position disclosures, they may still provide data on policy positions collected from public testimony, committee meeting minutes, or “witness lists,” which we refer to throughout the paper as “testimony positions,” as opposed to “lobbying positions.” These positions are primarily recorded by legislative committee staff, and most capture positions taken during a particular committee hearing. In most states, anyone, including the public, can submit testimony positions.

Testimony positions differ from lobbying positions in terms of source, timing, and format of reporting and potential biases. Testimony positions generally capture positions disclosed in the more public fora of open meetings rather than behind-closed-doors lobbying, where organizations may face increased incentives to strategically obscure their preferences (Broockman 2012). However, legislative staff record these positions instead of lobbyists themselves, meaning there may be lower risk of strategic misrepresentation in the actual recording of the positions. Because testimony positions are most often recorded by legislative committees during a hearing, lobbying positions are more likely to capture positions taken throughout the legislative process. Testimony positions more often come in the form of free or semi-structured text that must be parsed, while lobbying positions from all but one state came from structured databases.

Given the variability in disclosure laws, our data collection process proceeds differently depending on the data that is accessible. Eight states have fairly transparent lobbying disclosure laws, including Colorado, Iowa, Massachusetts, Montana,

Nebraska, New Jersey, Rhode Island, and Wisconsin. As such, we collect positional lobbying records from these states. Each state makes its data available through a different data portal. Preliminary outreach over email suggested that legislative staff are often unable or slow to provide access to the entire database, so in all but one case (Massachusetts), we wrote custom scrapers to gather data from each state's public-facing data portal.

In addition to the lobbying records, we collected records of positions stated in public testimony minutes or witness lists from legislative committees in an additional 11 states (of which two also have lobbying positions): Texas, Colorado, Illinois, Missouri, South Dakota, Kansas, Arizona, Florida, Ohio, Montana, and Maryland. In all cases, our scrapers stored the raw documents (html web pages and pdfs) in a Google Cloud Storage repository before we extracted the structured position data. This allowed us to capture each website's data only once, and then iteratively develop parsers to transform it into structured data, as well as to maintain a frozen record of the original sources of our dataset.

While the lobbying data were all scraped from html tables or structured databases, only five of the states with testimony positions presented them in a structured format. The remaining data had to be extracted from unstructured or semi-structured text, such as committee minutes, plaintext lists of organizations, or summaries prepared by legislative staffers. We wrote Python scripts to extract these data using a combination of regular expressions and machine learning-based entity recognition.

Because web scrapers can fail to load pages or miss important subdirectories on websites, and extracting structured data from unstructured text is prone to errors, we tested the coverage of our resulting dataset against manual collection of a random subset of policy position data for each state. Lobbying data were considered accurate when lobbying positions on a random set of 100 bills or scraped documents for a given state contained no errors. For testimony data, which was far messier, we set a lower bar, and in most cases, we ceased improving our document parsers once our accuracy reached 95% compared with a manually checked sample.⁴

The row-level metadata varies by state, with some states providing, for example, descriptions of policy preferences motivating a given lobbying position, or the amount of time or money spent on lobbying a given bill, while others do not. Some states list organizations' lobbying budgets while others do not.⁵ While lobbying expenditures provide useful context to policy positions, and we collect them where available, the central dataset consists of policy positions and the relevant actors involved. At a minimum, a row in the resulting dataset contains the columns presented in [Table 1](#).

Data cleaning and deduplication

Because the data come from states that each have unique disclosure laws, reporting systems, and database configurations, they require significant cleaning prior to use as a coherent dataset.⁶ Several cleaning steps were straightforward, including: ensuring

⁴States that provided testimony data in a structured format, for example, json, were not spot checked.

⁵See [Supplementary Appendix 5](#). Data Quality and Additional State Positions Data.

⁶Data cleaning and entity deduplication will continue after publication as new data are collected and processed.

Table 1. Fields in the lobbying position dataset

Field	Description
client_name	The interest group stating the policy position.
client_uuid	A unique identifier for each interest group, which may correspond to several different spellings of the interest group's name.
unified_bill_number	The numerical identifier for a piece of legislation, normalized to 10 digits, e.g., "0000001234" for "HB1234".
unified_bill_prefix	The abbreviated categorical identifier for a piece of legislation, typically "H[ouse]," "S[enate]," "A[ssembly]," "D[raft]," although some states use other prefixes; these are normalized to match Legiscan's conventions.
position	The position on the <i>Bill</i> expressed by the <i>Principal</i> , typically <i>support</i> , <i>oppose</i> , or <i>neutral</i> .
position_numeric	The position on the <i>Bill</i> expressed by the <i>Principal</i> , encoded as a 1 for support, 0 for neutral or other, and -1 for oppose. ^a
start_date	The date the position was taken ^b .
end_date	The date the position ended, if applicable; else left blank. Most states only report start_dates.
lobbyist_rep	Where applicable, the name of the lobbyist representing the principal in this activity.
lobbyist_firm	Where applicable, the name of the lobbying firm representing the principal in this activity.
state	The state in which this activity took place.
unified_session	The legislative session in which the lobbying took place.
committee	The committee in which the position was taken, if available.
record_type	"Lobbying" or "testimony" depending on the type of position.

^aNuanced types of support or opposition are simplified in the numeric position, for example, "support with amendments" is recorded as "1".

^bIf not available, the date the position was submitted is used.

that the data collected from each state are formatted with the same column names, formatting the bill identifiers consistently and in a way that allows easy cross-referencing with Legiscan and other datasets of bill-level information, and removing faulty rows if they occur. By matching interest group names in our dataset with entities found using Google Knowledge Graph, FTM's database, and OpenSecrets' database, we created a large set of predicted entity matches both within and between the three datasets. Following the standard practice (Binette and Steorts 2022), we then took each connected component of the resulting match graph as one entity and stored the FTM IDs, OpenSecrets IDs, Google Knowledge Graph IDs, and name variants for each unique identity under a unique identifier. In Massachusetts, for example, this process reduced the total number of unique entities from 4,371 to slightly over 2,670.⁷

⁷The most significant step in data processing and cleaning required standardizing and collecting various references to unique entities under different names at the intrastate level. For example, "Eversource Energy," a major investor-owned utility in Massachusetts and frequent position taker, is variously referred to as "Eversource," "Eversource Energy, Inc.," "Eversource Corporation," and other permutations. Many states, in particular those with testimony positions, yielded very messy entity names; therefore, deduplication was critical for making this dataset legible. The full details of our deduplication process are included in the [Supplementary Appendix](#). Our machine learning deduplication pipeline achieved a 92.6% F1 score for predicting true/false matches on a held-out set of candidate entity name pairs and a 99% area under the curve measure for comparing the true-positive to false-positive rates.

Database linkage

We augment our dataset of policy positions with metadata on both interest groups and bills. For interest groups, we use the data collected by FTM to add each group's lobbying spending and sector, industry, and business classifications. These data were only available for interest groups with a match in the FTM database. However, by adapting the approach in Garlick and Cluverius (2020), we were able to use a Complementary Naive Bayes classifier to generate guesses for the FTM sector and industry of each non-matched interest group using only their name.⁸ Details on this method are available in the [Supplementary Appendix](#). We did not fill in lobbying spending for interest groups without FTM matches. In the following section, we validate the Naive Bayes approach by comparing it with interest group communities identified using lobbying position data.

For bill-level metadata, we link our database to Legiscan's dataset of state-level legislation, and to a smaller dataset compiled by the National Conference on State Legislatures (NCSL). Legiscan collects bill titles, descriptions, histories, roll call votes, and other key bill information, with coverage extending back to 2009 in some states. NCSL collects similar information (albeit not roll call votes) and additionally tags bills with granular policy area topics, which we found useful for selecting subsets of related bills. The full set of NCSL categorizations and the number of bills available in each category are provided in the [Supplementary Appendix](#). We note that in many cases our dataset for a given state extends further back in time than either of these legislative datasets, which means that there are many years in which we lack pertinent bill-level metadata.

Record coverage

The dataset covers 17 states, comprising nearly 13 million recorded positions from individuals and interest groups between 1997 and 2022. Of these states, two have both lobbying and testimony positions, six have only lobbying positions, and nine have only testimony positions. Lobbying records have strong coverage (nearly 200,000 records per year on average) from 2010 onward, and testimony records exceed 100,000 per year from 2013 onward ([Figure 1](#)). Biennial legislative cycles in many states mean that interest groups give more testimony in even years than odd, while lobbying occurs session-round.

Lobbying and testimony data feature markedly different distributions of records per bill and per interest group. Individuals speaking on their own behalf make up the most common testifiers across most states, a dynamic which is absent from lobbying records; similarly, interest groups with only a few (1–100) associated records are vastly overrepresented in testimony data as opposed to lobbying ([Figure 2](#)). The modal number of position records per bill is about the same (~10 per bill) in lobbying and testimony data. Testimony records feature a small number of bills with very high numbers of records; one Illinois bill has nearly 90,000 associated testimony positions across its several versions.

⁸Roughly speaking, Naive Bayes classifiers attempt to associate words found in each group's name with their FTM sector and industry categories; for instance, they might find that the presence of "Petroleum" in the name increases the relative probability of a group falling in the ENERGY and NATURAL RESOURCES sector and the OIL and GAS industry.

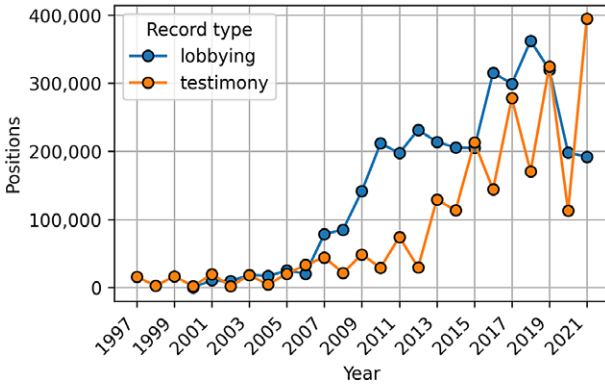


Figure 1. Policy position records available per year by record type. Records of individual testimony are removed (albeit imperfectly), so this chart only reflects the available records of interest group positions. Not all states have available position data in a given year, and we cut off the graph at 2021 as this is the last year in which all states have available data (see Table 2).

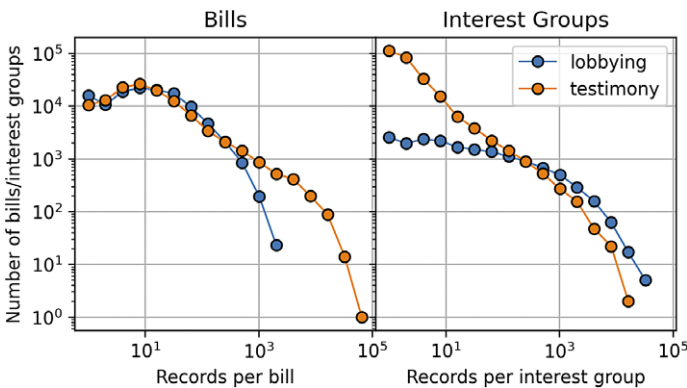


Figure 2. Histograms of records per bill (including individual testimony) and records per interest group (unassociated individuals excluded). Histograms are split according to record type, with data from lobbying records shown in blue and data from testimony in orange. Note logistic x- and y-axis scales.

The distribution of positions – support, oppose, or neutral – also varies depending on record type. “Oppose” positions occur more often than “Support” positions in testimony, while the relative frequencies of each are more balanced in lobbying (see Table 2). This disparity may indicate that some interest groups find publicly supporting certain legislation less politically beneficial than lobbying for it in private. Lobbying data feature far more “Neutral” positions than testimony. This result likely arises because interest groups seeking information-based lobbying, or seeking to negotiate over language/provisions within a bill, find it more useful to do so in private than in public hearings. In addition, groups looking to obscure their true position on a bill are better able to do so by lobbying and recording a neutral position. As neutral positions do not factor into our SBM models, this disparity may not lead to a large difference in the accuracy of the community detection.

Table 2. Summary statistics of positions dataset

State	Record type	Support	Neutral	Oppose	% Neutral	Average positions per bill	Years covered
AZ	Testimony	3007440	124521	2604305	2.2	398.7	2006–2022
CO	Lobbying	214645	919610	464469	57.5	132.7	2003–2023
CO	Testimony	13134	4414	42826	7.3	22.6	2006–2022
FL*	Testimony	13526	4669	35094	8.8	6.2	2004–2022
IA	Lobbying	90545	531455	167782	67.3	35.8	2009–2022
IL	Testimony	1232635	19091	1924315	0.6	166	2013–2022
KS	Testimony	11203	2973	25862	7.4	18.6	2014–2022
MA	Lobbying	111160	210783	153709	44.3	15.9	2010–2021
MD	Testimony	19001	4193	73058	4.4	14.2	2020–2022
MO	Testimony	16110	2927	45812	4.5	6.9	2003–2022
MT	Lobbying	23060	33440	45415	32.8	10	2006–2022
MT	Testimony	15349	4875	32198	9.3	22.3	2017–2021
NE	Lobbying	77103	78092	141489	26.3	23.5	2000–2021
NJ	Lobbying	13687	16359	40315	23.3	4.6	2014–2022
OH	Testimony	10039	4973	27409	11.7	12.1	2015–2022
RI	Lobbying	15198	11210	41313	16.6	16.3	2018–2022
SD	Testimony	19465	1309	44731	2	6	1997–2022
TX	Testimony	140771	90985	415461	14.1	17.9	1997–2021
WI	Lobbying	22212	26006	50494	26.3	6.6	2002–2022

Note: Positions from Senate Bills are not available in Florida. All other states include positions from both chambers, except for Nebraska, which is unicameral.

Identifying policy coalitions

Previous interest group studies have used deductive approaches to assign interest group categories. For instance, Gray *et al.* (2015) hand-coded 26 economic sectors from their state interest group population data. Holyoke (2019) hand-coded data from the National Institute for Money and Politics to categorize interest groups by type and economic sector. While these datasets are comprehensive, they take a considerable amount of resources and time to reproduce and present significant replication and extension challenges. Garlick and Cluverius (2020) use a supervised learning method to classify state interest group populations, which provides a systematic set of data on interest group types, but does not identify coalitions of groups that work around specific pieces of legislation, making their influence more difficult to ascertain.

Inductive approaches, such as Kim and Kunisky (2021), allow scholars to make inferences about interest group populations by examining links between policy actors.⁹ The authors map networks of “legislative communities” composed of interest groups and members of Congress through use of these linkages. This inductive process allows researchers to observe connections between interest groups and politicians directly, capturing opportunities for influence that previous work could not. Our approach to identifying state interest group coalitions also follows an inductive approach. A particularly notable advantage of this approach is that it opens up a black box surrounding group interactions that cannot be inferred by group type or economic sector alone.

⁹Political scientists have also used an inductive approach to estimate ideology based on preferences shared on social media (Bond and Messing 2015) and to consider how coordinated efforts of party organizations in support of candidates affect election outcomes (Desmarais *et al.* 2015).

We identify interest group coalitions by fitting an SBM to the dataset of all lobbying positions for a given state and record type (lobbying or testimony).¹⁰ SBMs are generative statistical models that represent an observed network as a collection of *blocks*, or groups of nodes, each of which has an assigned probability of forming edges internally and with nodes in other blocks. The block assignment (also called the partition) and edge probabilities that give the maximum likelihood of generating the observed network define the node communities. In our case, the observed network is a bipartite network, in which two types of nodes (interest groups and bills) link to each other via three types of edges (neutral, support, or oppose positions – although we opted to ignore neutral positions in estimating the models).¹¹ These blocks of interest groups often, but not always, map onto different industry sectors or organizational types. However, the exceptions to this pattern can provide important insights into the nuances of state politics. Similarly, blocks on bills often – but not always – capture policies within the same issue or topic, according to an *a priori* categorization.

We use the *graph-tool* Python package to fit a degree-corrected categorical hierarchical SBM to our datasets (Peixoto 2015; The *graph-tool* python library 2014). This model has several advantages: the partitions it finds attempt to parsimoniously predict all three types of positions, it finds overlapping communities of bills and interest groups simultaneously,¹² it adjusts for differences in the lobbying rates between interest groups or bills, and its hierarchical nature means it automatically determines several levels of more-to-less granular clusterings. For a review of the SBMs advantages and other uses, see Peixoto (2019). We have uploaded fitted SBMs for each state and record type (testimony and lobbying), as well as code for reestimation of these models. Stochasticity implies that the outputs of each model run will be different, so researchers may choose to run each model several times and pick the result with the best fit, as measured by minimizing the description length (or equivalently, the entropy) of the model.

As an analytical tool for studying policy preferences, SBMs hold several advantages over the more common approach of estimating interest group ideal points. First, SBMs do not presume dimensionality. Whereas ideal point estimation generally sorts interest groups along a single dimension, SBMs infer the optimal number of

¹⁰We combine all years' data to estimate each SBM for two reasons: first, the more data used, the more stable the blocks identified by the algorithm, preventing random stochastic fluctuations from interfering with interpretation of the results; second, we want to identify long-standing and stable camps of interest groups with similar preferences, which requires long-term data on their positions. Capturing only positions from a given year would subject the estimation to arbitrary biases depending on, for example, which bills happened to be introduced that year, the larger macroeconomic context, and so forth.

¹¹In broad terms, this approach to the study of interest group preferences is analogous to the application of topic modeling to large-scale text analysis. Topic modeling (Blei, Ng, and Jordan 2003; Gerlach, Peixoto, and Altmann 2018) allows researchers to inductively find themes in a text corpus that best describe the documents, rather than hand-coding documents using a predetermined list of themes. Similarly, we show that applying our approach to the large-scale policy position data we collected allows us to inductively find communities of interest groups with shared policy interests, which the researcher can then qualitatively interpret to aid understanding of state-level politics.

¹²Kim and Kunisky (2021) argue that their bipartite network model improves on the standard bipartite SBM by including both overlapping node communities and overlapping *link* communities (i.e., links between two nodes can have mixed membership of their own); *graph-tool* addresses the first critique, by finding overlapping node communities, but does not identify link communities.

blocks and levels to represent the policy position graph. Second, SBMs yield more precise predictions. Given the block assignments for an interest group and a bill, the SBM provides a posterior probability of each possible position (support/oppose/none) that the interest group could state on that bill. Finally, and most importantly, SBMs are simply better-fit to the structure of interest group policy position data. These data are characterized by very high nonresponse rates because interest groups focus on only some issues. Ideal point estimation – at least at low dimensionality – gives useful information about *how* an interest group might lobby, but not *when*; SBMs provide both.

Application of CHORUS: mapping policy preferences in Wisconsin

In the demonstrative analysis below, we show how this dataset can be used to address a less-studied descriptive question: what *coalitions*, defined by their policy preferences, characterize the interest group population interacting with a given state legislature? We demonstrate the breadth of this dataset by first providing a broad assessment of the entire interest group population in one state (Wisconsin) and then exploring coalitions across several states on a particular policy area: renewable energy and climate change. This policy area is an ideal demonstration case given the diverse array of interest groups involved in the policy space and research demonstrating the considerable impact of interest groups on policy outcomes (Brulle and Aronczyk 2020; Downie 2017; Karol 2019). Climate change has taken center stage in recent scholarship on American political institutions (Hacker *et al.* 2022), and this field therefore urgently needs research that can pierce the veil of interest group activity in state governments.

We first fit the SBM to all Wisconsin's lobbying records to generate a high-level map of the interest group populations in the state. The resulting hierarchical SBM contains four levels of clusterings. Figure 3 displays the entire output of the SBM, including both the positions interest groups took on bills (represented by green and red edges) and the hierarchically nested clusters into which they fell (represented by the light blue tree structure superimposed on the graph). Evidently, the SBM succeeds in uncovering a large-scale structure in the policy position data; if no such structure existed, the model would not be able to uncover fine divisions at several levels of granularity.

At the finest level, blocks of interest groups, which capture organizations stating similar positions on bills, are remarkably precise and informative. The 136 interest group communities identified include a small cluster with two plant-based meat organizations (Plant Based Foods Association and The Good Food Institute), a cluster featuring all five gun rights advocacy groups in the state, a cluster of all the telecommunications companies in the state, and a cluster of all Native American tribes that lobbied in the state. At higher levels, broader patterns in interest group position-taking emerge. The fourth and least-granular community level features seven policy coalitions, as shown in Figure 4. These include a coalition of progressive policy advocacy groups and unions whose members frequently oppose the positions taken by a community of chambers of commerce and insurance companies.

The SBM generates a hierarchical clustering of bills alongside interest groups. Under the assumption that blocks of interest groups are likely to lobby the same way on similar bills, it is reasonable to infer that the blocks of bills group together similar

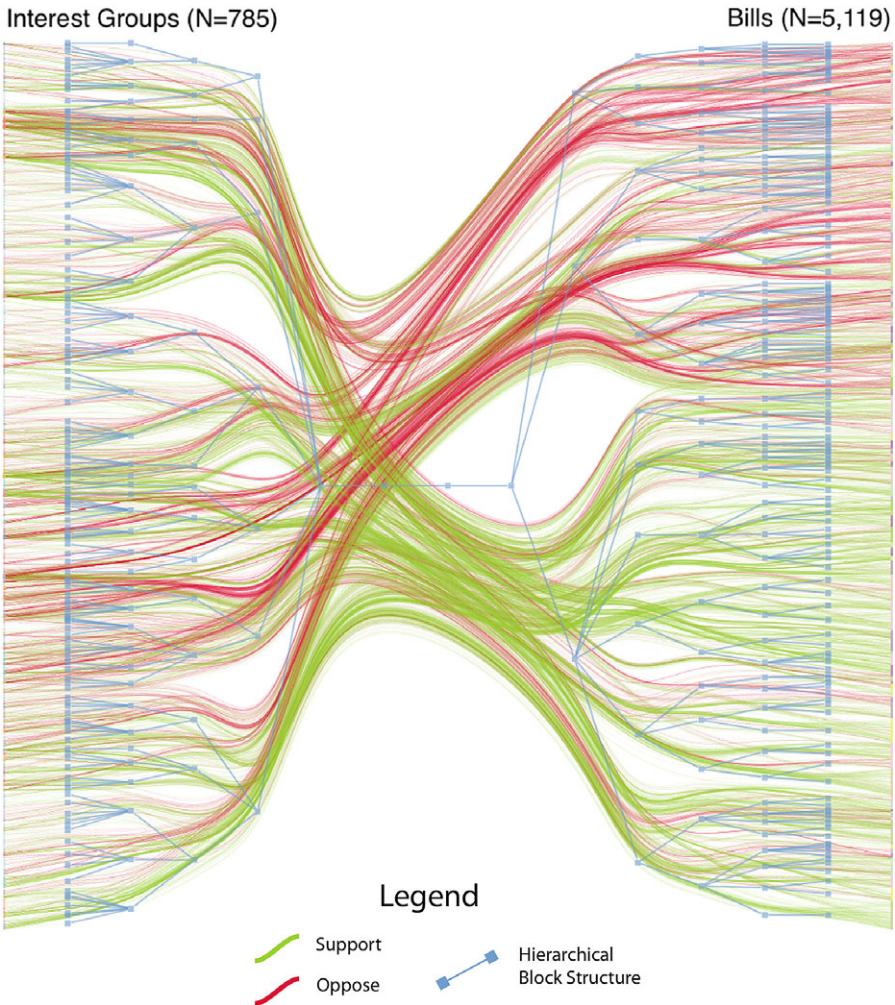


Figure 3. The full network of interest groups and bills in Wisconsin. The bipartite interest group-bill network for Wisconsin, sorted by the hierarchies detected by the SBM. Interest groups ($N = 785$) are on the right and bills ($N = 5,119$) are shown on the left (individual nodes are generally too small to distinguish). Each edge corresponds to a position on a bill, colored green for support and red for oppose. The light blue tree structure shows the divisions of interest groups and bills into hierarchically nested clusters. The full graph includes 34,179 “support” positions and 17,327 “oppose” positions; a subsample of 5,000 edges (positions) are plotted here.

policies with increasing granularity. This allows us to explore the issue space in a systematic and inductive way. To do so, we require a way of quickly characterizing the bills in each block, which we accomplish using a Naive Bayes approach similar to the one used in guessing organizations’ industry classifications (Garlick and Cluverius 2020).

We find characteristic words associated with blocks of bills at each level of clustering by fitting a multinomial Naive Bayes classifier to predict bill blocks using

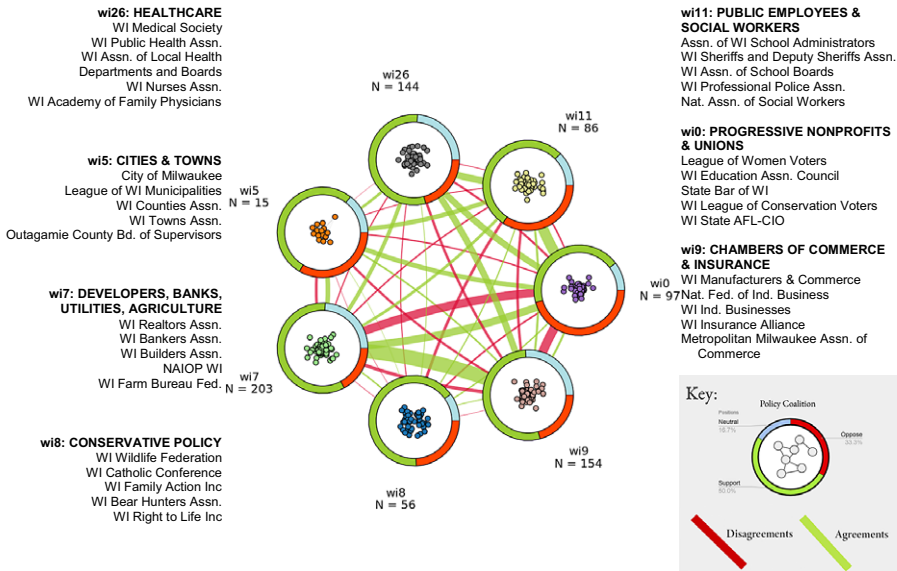


Figure 4. High-level policy coalitions in Wisconsin. The seven circles displayed each represent one of the communities identified at the highest non-trivial level of the SBM (level 4). Each of these policy coalitions has been titled based on a subjective interpretation of the industry categorization and background of its most active members. Edges between groups indicate the number of instances of agreement (green) or disagreement (red) between their members on legislative positions. The pie charts ringing each policy coalition indicate the aggregate distribution of their “support” (green), “oppose” (red), and “neutral” (blue) positions. The N for each coalition indicates the number of coalition members.

word counts from every bill’s title, which in Wisconsin includes a brief one-sentence description of the bill (see [Supplementary Appendix](#) for details). The resulting classifier allows us to find the words most predictive of each bill category at each level of the hierarchy, which provides a means of quickly summarizing the policy issues associated with each bill cluster identified by the SBM. Note that the classifier itself does not perform well as a predictor of a given bill’s block assignment; rather, the parameters of the Naive Bayes model allow us to identify characteristic words for each block solely for the sake of quickly summarizing some of the common themes.

Table 3 shows the fourth-level bill clusters in Wisconsin, the number of bills in each cluster, the percentage that passed both chambers, and the top descriptors discovered using this Naive Bayes approach. Some clusters clearly identify a substantive policy area, such as wi19, which appears to be related to agriculture and food products, or wi21, which appears to be related to abortion and vaccine policy – that is, conservative cultural issues. Others appear to combine several substantive issue areas, which may become disentangled at lower levels of the clustering.

The SBM provides a natural way of defining interest groups on a spectrum from specialists with interests in highly particular issue areas to generalists with broad sets of interests. Intuition suggests that business associations and nonprofits representing a wide array of stakeholders, such as chambers of commerce, should attempt to influence a wide range of policies, while highly specialized organizations such as issue-specific advocacy groups or small firms, engage in targeted lobbying on a small number of issues. Taking the bill clusters assigned by the SBM as indicative of

Table 3. Top-level bill clusters and characteristic descriptors in Wisconsin

Bill cluster	N	% Passed	Top descriptors
14	773	31%	incremental, financial, apprenticeship, tax, franchise
3	700	24%	enforcement, officers, hunting, congressional, raffles
23	623	21%	wind, room, retail, creative, alcohol
1	577	27%	mental, treatment, child, prevention, dental
22	473	13%	absentee, ballots, foodshare, parental, charter
19	451	32%	agricultural, food, dairy, product, labeling
2	358	32%	pharmacy, wellness, compensation, worker, workplace
18	319	8%	sand, frac, mining, homeless, lead
16	263	20%	firearms, transfers, tribal, elder, handguns
4	229	7%	commencement, below, fall, classes, drugs
21	193	10%	abortion, abortions, selective, fetal, vaccination
10	163	11%	complementary, alternative, unpasteurized, professional, smoking

different issue spaces, we can investigate the spread of each interest groups' lobbying efforts across those issue spaces using the concept of entropy. Entropy measures the average amount of information conveyed by an event – in this case, when an interest group lobbies on a bill. The entropy of an interest group k 's lobbying efforts can be defined as follows:

$$H_k = - \sum_i \left(\frac{n_k^i}{N_k} \right) \times \log_2 \left(\frac{n_k^i}{N_k} \right)$$

where i refers to each issue area, n_k^i refers to the number of bills interest group k lobbied on in each issue area i , and N_k refers to the total number of bills interest group k lobbied on. Low-entropy interest groups focus their lobbying on only a few issue areas, while high-entropy groups lobby across many. We calculated the entropy scores for every interest group in Wisconsin using the lowest-level bill clusters as issue areas. Table 4 displays the five highest and five lowest-entropy groups. The highest-entropy groups are nearly all broad-based business associations or municipalities, which we expect to have diverse interests; the five lowest-entropy groups are all issue-specific advocacy organizations or small businesses. These results validate the SBM output, as they indicate that organizations' spread across SBM issue areas matches expected patterns.

Table 4. Highest and lowest-entropy interest groups in Wisconsin, by bills lobbied on

Interest group	Entropy
Wisconsin Manufacturers & Commerce	4.479873233
Fox Cities Chamber of Commerce & Industry	4.165201359
Wisconsin Independent Businesses Inc.	4.156174123
Wisconsin Counties Association	4.154366743
City of Milwaukee	4.087227733
Pres House – UW Madison	0
Wisconsin Council on Community Corrections	0
Ground Water Management & Water Conservation Lobbying Assn.	0
People for the Personal Choice of Raw Milk	0
Can Manufacturers Institute	0

In just the same way, each bill can be assigned an entropy according to the number of interest groups from different policy coalitions which lobbied on it. Each session brings a number of omnibus bills: legislation such as budget proposals or broad initiatives which touch on many different issues and consequently will interest many different organizations. We expect such bills to have high entropy, as interest groups from many different policy coalitions will lobby on them (Brasher, Lowery, and Gray 1999). Conversely, bills which touch on small and defined issues – for instance, minor changes to industry-specific regulations – are expected to have low entropy, as only interest groups in the affected industries will expend effort lobbying them.

The results in Table 5 affirm these expectations. The five highest-entropy bills in Wisconsin over the covered time period include an emergency response to COVID-19, a special session budgeting bill, a wide-ranging emissions reduction bill, another budget bill, and a wide-ranging health care and insurance liability bill. The lowest-entropy bills touch on specific topics such as hearing procedures, miscellaneous vehicular laws, and evidence of sexual offenses.

We can further validate the SBM clustering results by comparing them to the pre-assigned interest group and bill categories taken from FTM and NCSL, respectively. Here we rely on the normalized mutual information (NMI), which measures the amount of information one categorization gives about another, on a zero to one scale. For example, an NMI of one between the FTM classifications and the SBM block assignments indicates that the SBM blocks perfectly capture the industry categories of interest groups; an NMI of zero would indicate that the SBM blocks are completely unrelated to the industry classifications. We calculate the NMI between interest group block assignments and industry categorizations at each hierarchical level of the

Table 5. Highest and lowest-entropy bills in Wisconsin, by lobbying interest groups

Session	Bill	Description	Entropy
2021	AB 1	State government actions to address the COVID–19 pandemic, extending the time limit for emergency rule procedures, providing an exemption from emergency...	4.12
2011–X1	AB 11	State finances, collective bargaining for public employees, compensation and fringe benefits of public employees, the state civil service system...	3.86
2009	SB 450	Goals for reductions in greenhouse gas emissions, for construction of zero net energy buildings and for energy conservation; information, analyses...	3.85
2009	SB 62	State finances and appropriations and making diverse other changes in the statutes.	3.62
2011–X1	AB 1	Limiting noneconomic damages awarded in actions against long–term care providers; actions against manufacturers, distributors, sellers, and promoters...	3.61
2011	SR 22	Prohibiting waiver of public hearing requirement for bills placed on a calendar.	0
2009	SB 58	Permitting third–party testers to administer driving skills tests for certain noncommercial motor vehicle drivers....	0
2011	SB 56	Evidentiary recordings of persons under the age of 18 engaging in sexually explicit conduct and certain sex offenses against children and providing penalties...	0
2021	AB 365	Whip lights on all–terrain and utility terrain vehicles.	0
2011	SB 51	The Wisconsin Small Company Advancement grant program and making an appropriation. (FE)	0

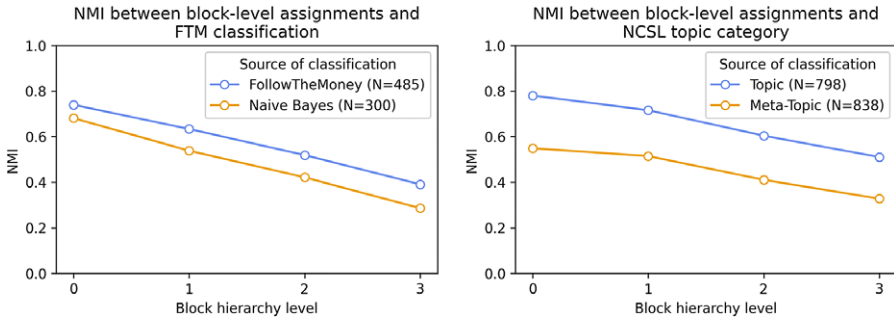


Figure 5. Comparing hand-coded and predicted industry and topic categories with SBM blocks. Left: the normalized mutual information between SBM-assigned interest group clusters and FollowTheMoney industry classifications taken directly from FollowTheMoney's dataset (blue), or inferred via Naive Bayes (orange). Right: the NMI between SBM-assigned bill clusters and topics (blue) or meta-topics (orange) assigned by NCSL. Values for N indicate the number of bills with an assigned category under each given label.

SBM, and distinguish between industries assigned directly by FTM and those we obtained using the Naive Bayes classifier (Figure 5, left).

The results indicate that the most-granular SBM classification closely matches industry divisions among interest groups, with an NMI of 0.79; they also indicate that the Naive Bayes industry classifications, which have a nearly identical NMI, capture almost the same substantive categories of groups, further validating the Naive Bayes approach to interest group classification. We also calculate the NMI between SBM-assigned bill clusters and the bill topics and meta-topics collected from the National Conference of State Legislatures (Figure 5, right). While fewer bills have assigned topics than interest groups do industries, we still observe a similarly high overlap between this external categorization and the SBM blocks, which validates our hypothesis that the SBM picks out substantive issues among bills.

We note that the dataset allows for several more uses not demonstrated here. For instance, because Legiscan provides roll call votes on many of the bills in the dataset, researchers can easily incorporate votes and thus legislators and political parties into the analysis. The bill outcomes Legiscan provides allow researchers to straightforwardly assess interest groups' success in terms of the number of their (dis) favored bills that (fail to) pass. Many of the interest groups in this dataset have also been linked to entity identifiers in the FTM and OpenSecrets databases maintained by the Center for Responsive Politics, allowing researchers to augment these records with lobbying spending information and industry categorizations at both the state and national level. Our incorporation of NCSL bill categorizations also provides a deductive policy taxonomy to complement the inductive approach to bill categorization shown here.

Application 2: Mapping climate and energy policy preferences across four states.

To demonstrate one use of our dataset for making cross-state comparisons, we present a brief analysis of the policy coalitions active in climate and energy politics

in four states: Colorado, Texas, Illinois, and Massachusetts. These states were selected for comparison because they span a range of energy production regimes. Illinois produces the second-most coal and the most nuclear power in the nation; Colorado produces a substantial amount of oil and natural gas; Texas is the largest energy producer of all the states, with a dominant oil extraction industry but also a sizable onshore wind industry; and Massachusetts does not produce fossil fuels within its borders (U.S. Energy Information Administration - EIA - Independent Statistics and Analysis *n.d.*). We isolated all bills in the “energy” meta-topic within NCSL’s database for each state, and using either the second-lowest or lowest (in Texas’s case) SBM block assignments¹³, created networks showing the amount of mutual agreement and disagreement in position-taking on these bills by interest groups in each state (see [Figure 6](#)). We kept only those interest groups which lobbied on five or more of the energy-related bills and excluded policy coalitions with three or fewer members to avoid cluttering the graphs with only loosely-relevant policy coalitions. We then named each policy coalition according to the industries of its members; in every case, one or two industries dominated each policy coalition, making it easy to choose appropriate names. Full lists of all members of each coalition are available in the [Supplementary Appendix](#).

In Colorado, a large number of fossil fuel producers and utility companies receive support from real estate, business associations, and agriculture. This fits expectations for a state whose economy is so dominated (comparatively speaking) by extractive industries like oil and gas. Meanwhile, these interests predominantly oppose environmental advocates, who receive support from municipalities, public health, and progressive, public interest organizations.

In Texas, coalition politics are dominated by a three-way alliance between oil majors, petroleum producers, and gas utilities. Social greens oppose this fossil-fueled alliance, but receive little support from renewable energy companies in doing so. In fact, renewable companies have slight support for oil companies, which arises from bills combining technology-neutral energy incentives. This might explain why fossil fuel interests are so dominant in Texas politics, despite the sizable and rapidly growing wind and solar industry in that state. Interestingly, and very much unlike what we see in other states, electric utilities in Texas are not testifying for/against much legislation as the other major coalitions are. This could be the result of “neutral” positions, but it could also be the result of Texas’ unusual and “hyper-restructured” electric utility sector, in which electric utilities compete with one another in both generation and distribution (and are therefore less monopolistic than in other states).

In Illinois, Social Greens are aligned with the Solar Industry on many bills, perhaps due to that state’s distributed generation policies which empower rooftop solar companies and are supported by environmental groups due to their decarbonization potential. However, this coalition faces strong opposition from unions, fossil fuel companies, and utilities. It makes sense that these interests would be aligned due to such high levels of in-state fossil fuel production, particularly coal. On the other hand, it is interesting that unions seem to oppose different bills than those opposed by the

¹³We use the level of block assignments that yields a readable graph with on the order of 10 coalitions; in most cases, the lowest-level block assignments are far too splintered and numerous to be helpful in this analysis, since they are generated using data from a much larger slice of the dataset than we consider here.

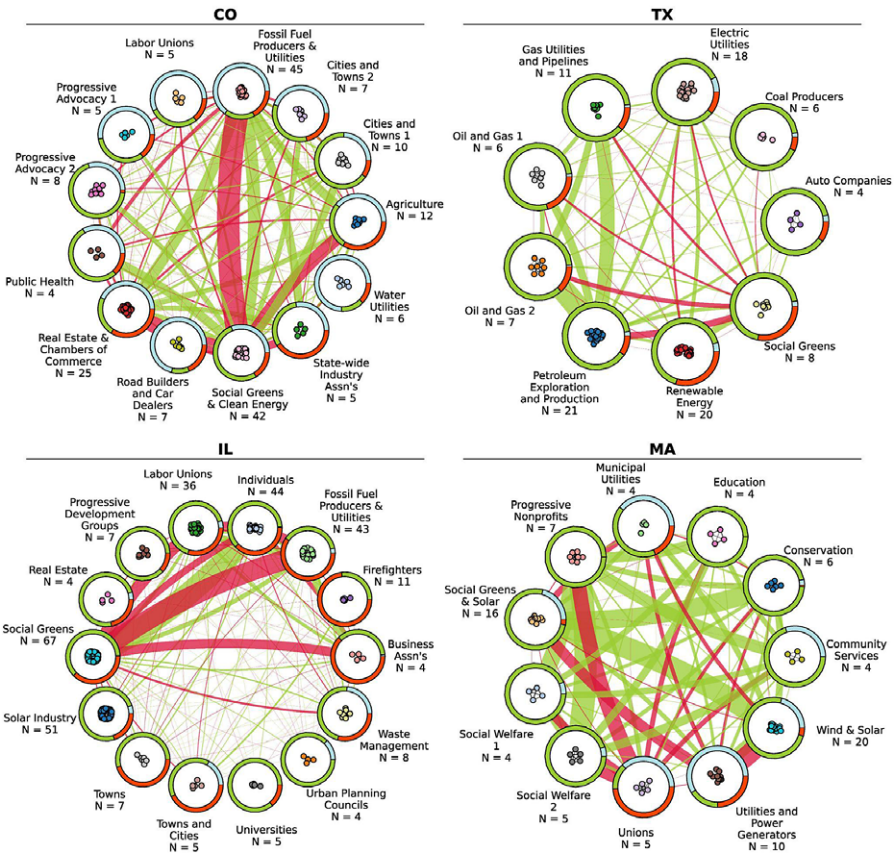


Figure 6. Policy coalition alignment graphs for energy legislation in four states. These charts illustrate trends in testimony and lobbying positions taken on bills in the Energy category identified by NCSL in each state. Each graph shows the policy coalitions inferred by the SBM within the labeled circles. Lines between coalitions indicate the extent of policy preference alignment on energy legislation (green bands) and disagreement (red bands), with line width proportional to the number of times members from each coalition either agreed or disagreed in their stated positions on energy-related bills. The donut charts around each policy coalition indicate the proportion of support, neutral, or oppose positions they stated on all energy-related bills (see Figure 4 for key).

fossil fuels-utilities coalition. This is precisely the type of unexpected finding (perhaps ripe for qualitative research) that our dataset would allow us to observe and that otherwise might be easily overlooked. Meanwhile, as expected, Illinois business associations tend to support the fossil fuels-utilities coalition and oppose Social Greens.

Finally, in Massachusetts, a state that does not produce fossil fuels within its borders, we see no fossil fuel companies leading a coalition, unlike in other states with greater in-state fossil fuel production. We see that the Utilities and Power Generators coalition is sometimes on the same side but other times on the opposing side from the Wind and Solar coalition. This likely depends on the specific language of the bill, and may also be explained by inherent differences between Utilities and Power Generators, given that Massachusetts law generally prohibits investor-owned utilities from

owning generation assets, and Power Generators may be renewables or may be gas (Basseches 2023).

Conclusion

We present the CHORUS dataset and, through the paradigm of stochastic block modeling and community detection, show its usefulness in defining the space of state-level interest group advocacy and legislation relying solely on the policy positions disclosed by interest groups. For researchers' convenience, we have uploaded fitted SBMs for each state in the supplemental material. These applications provide a foundation through which these data can be used, though we hope that this dataset will allow widespread investigation and hypothesis testing of theories in state politics that were previously difficult or impossible to explore.

In particular, our dataset directly responds to Anzia's (2019) call for analyzing variation in state-level politics and policy as a means of advancing the fields of interest groups and democratic representation. Our dataset enables the discovery of "different constellations of interest groups – constellations in which groups ... often have a much larger presence than they do at the national level" (Anzia 2019, p. 344). Our dataset provides researchers with abundant new opportunities to "develop (and test)" new interest group theories, "[using] public policy as an anchor" (Anzia 2019, pp. 345–346). In addition, our dataset can be used to explore interest group-party alignments, promoting theory development and testing in this literature as well. These alignments have been theorized (Bawn *et al.* 2012), but are rarely tested. Policy process literature on advocacy coalitions can further explore coalition activities through use of these data (e.g., Kukkonen, Ylä-Anttila, and Broadbent 2017; Sabatier 1988).

CHORUS may also be fruitful in addressing the "problem of preferences," or the tendency of organized interests to publicly convey preferences that differ from their "true preferences," in order to maintain a strategic advantage (Broockman 2012). The opacity of the disclosure systems and the clunkiness of most states' disclosure websites mean these records rarely come to light, and interest groups therefore face smaller incentives to strategically misrepresent their policy positions in this context. The data we present likely allow for a more accurate examination of preferences revealed by interest groups that lobby state governments over time and across states because of this difference in incentives.

We also anticipate that scholars will find entirely different frameworks within which to analyze these data. For instance, interest group policy positions are natural candidates for ideal point estimation, as Crosson, Furnas, and Lorenz (2020) and Thieme (2019) have shown. They also provide a natural way to define and test theories of political party coalitions, following the approach in Bonica (2014). Scholars could use these data to uncover coalitions within particular policy domains, regardless of whether they observably coordinate. Analysis could show relations of support and opposition between coalitions and assess their relative influence over legislative outcomes as Culhane, Hall, and Roberts (2021) did for energy policy in Massachusetts (e.g., Culhane, Hall, and Roberts 2021).

In addition, these data can be used to test theories of interest group influence and representation. With the explicit knowledge of interest group preferences provided by CHORUS, scholars could test whether these preferences explain the gap between

policy outcomes and citizen's preferences. They could also test the relative influence of political parties versus interest groups in state legislatures. For example, one could test whether minority parties find more success where their interests align with interest groups. Relatedly, studies could identify systematic party-interest group alignments within and across states.

Further work could be done to explain interest group policy preferences and develop novel theories guided by SBMs and other models. These data could also allow testing theories of lobbyist influence, strategy and coordination, in addition to relating lobbying relationships with legislators to campaign finance and other expenditures.

CHORUS builds on previous work that investigates the activities and influence of interest groups in the United States and provides a much needed set of comprehensive data at the state level. We hope that future research will uncover yet more uses for this dataset with the aim of advancing the field of state politics, interest group influence, and in understanding how these ever-present groups shape the actions of state policymakers.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/spq.2024.6>.

Data availability statement. Replication materials are available on SPPQ Dataverse at <https://doi.org/10.15139/S3/RPU1QP> (Hall et al. 2024).

Author contribution. GH conceived of and led the study and carried out all quantitative analysis. TC and GH contributed equally to data collection. JB and RBT contributed theoretical background and the case studies. TC, JB, and RBT wrote the literature review, and authors contributed equally to all remaining sections.

Funding statement. The authors received a grant from the Climate Social Science Network over 2021–2022 to fund data collection efforts.

Competing interest. The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

- Abi-Hassan, Sahar, Janet M. Box-Steffensmeier, Dino P. Christenson, Aaron R. Kaufman, and Brian Libgober. 2023. "The Ideologies of Organized Interests and Amicus Curiae Briefs: Large-Scale, Social Network Imputation of Ideal Points." *Political Analysis* 31 (3): 396–413.
- Anzia, Sarah F. 2019. "Looking for Influence in All the Wrong Places: How Studying Subnational Policy Can Revive Research on Interest Groups." *Journal of Politics* 81 (1): 343–51.
- Basseches, Joshua A. 2023. "Who Pays for Environmental Policy? Business Power and the Design of State-Level Climate Policies." *Politics & Society*. <https://doi.org/10.1177/00323292231195184>.
- Basseches, Joshua A., Rebecca Bromley-Trujillo, Maxwell T. Boykoff, Trevor Culhane, Galen Hall, Noel Healy, David J. Hess, David Hsu, Rachel M. Krause, Harland Prechel, J Timmons Roberts, and Jennie C Stephens. 2022. "Climate Policy Conflict in the U.S. States: A Critical Review and Way Forward." *Climatic Change* 170 (3): 32. <https://doi.org/10.1007/s10584-022-03319-w>.
- Baumgartner, Frank R., Jeffrey M. Berry, Marie Hojnacki, David C. Kimball, and Beth L. Leech. 2009. *Lobbying and Policy Change: Who Wins, Who Loses, and Why*. Chicago, IL: University of Chicago Press.
- Baumgartner, Frank R., and Beth L. Leech. 1998. *Basic Interests: The Importance of Groups in Politics and in Political Science*. Princeton, NJ: Princeton University Press.
- Baumgartner, Frank R., and Beth L. Leech. 2001. "Interest Niches and Policy Bandwagons: Patterns of Interest Group Involvement in National Politics." *Journal of Politics* 63 (4): 1191–213.
- Bawn, Kathleen, Martin Cohen, David Karol, Seth Masket, Hans Noel, and John Zaller. 2012. "A Theory of Political Parties: Groups, Policy Demands and Nominations in American Politics." *Perspectives on Politics* 10 (3): 571–97.

- Binette, Olivier, and Rebecca C. Steorts. 2022. "(Almost) All of Entity Resolution." *Science Advances* 8 (12): eabi8021.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.
- Bond, Robert and Solomon Messing. 2015. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109 (1): 62–78.
- Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58 (2): 367–86.
- Box-Steffensmeier, Janet M., and Dino P. Christenson. 2014. "The Evolution and Formation of Amicus Curiae Networks." *Social Networks* 36: 82–96.
- Brasher, Holly, David Lowery, and Virginia Gray. 1999. "State Lobby Registration Data: The Anomalous Case of Florida (And Minnesota Too!)." *Legislative studies quarterly* 24 (2): 303–14.
- Bromley-Trujillo, Rebecca, J. S. Butler, John Poe, and Whitney Davis. 2016. "The Spreading of Innovation: State Adoptions of Energy and Climate Change Policy." *Review of Policy Research* 33 (5): 544–65.
- Broockman, David E. 2012. "The 'Problem of Preferences': Medicare and Business Support for the Welfare State*." *Studies in American Political Development* 26 (2): 83–106.
- Brulle, Robert, and Melissa Aronczyk. 2020. *Environmental Countermovements: Organised Opposition to Climate Change Action in the United States*. 1st ed. London: Routledge.
- Butler, Daniel M., and David R. Miller. 2022. "Does Lobbying Affect Bill Advancement? Evidence from Three State Legislatures." *Political Research Quarterly* 75 (3): 547–61.
- Crosson, Jesse M., Alexander C. Furnas, and Geoffrey M. Lorenz. 2020. "Polarized Pluralism: Organizational Preferences and Biases in the American Pressure System." *American Political Science Review* 114 (4): 1117–37.
- Culhane, Trevor, Galen Hall, and J. Timmons Roberts. 2021. "Who Delays Climate Action? Interest Groups and Coalitions in State Legislative Struggles in the United States." *Energy Research & Social Science* 79: 102114.
- Desmarais, Bruce A., Raymond J. La Raja, and Michael S. Kowal. 2015. "The Fates of Challengers in US House Elections: The Role of Extended Party Networks in Supporting Candidates and Shaping Electoral Outcomes." *American Journal of Political Science* 59 (1): 194–211.
- Downie, Christian. 2017. "Business Actors, Political Resistance, and Strategies for Policymakers." *Energy Policy* 108: 583–92.
- Eissler, Rebecca, and Bryan D. Jones. 2019. "The US Policy Agendas Project." In *Comparative Policy Agendas: Theory, Tools, Data*, eds. Frank R. Baumgartner, Christian Breunig, and Emiliano Grossman. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780198835332.003.0021> (February 2, 2023).
- Erikson, Robert S., Gerald C. Wright and John P. McIver. 1993. *Statehouse Democracy: Public Opinion and Policy in the American States*. Cambridge: Cambridge University Press.
- Garlick, Alex, and John Cluverius. 2020. "Automated Estimates of State Interest Group Lobbying Populations." *Interest Groups & Advocacy* 9 (3): 396–409.
- Gerlach, Martin, Tiago P. Peixoto, and Eduardo G. Altmann. 2018. "A Network Approach to Topic Models." *Science Advances* 4 (7): eaaq1360.
- Gilens, Martin, and Benjamin I. Page. 2014. "Testing Theories of American Politics: Elites, Interest Groups, and Average Citizens." *Perspectives on Politics* 12 (3): 564–81.
- Givel, Michael S., and Stanton A. Glantz. 2001. "Tobacco Lobby Political Influence on US State Legislatures in the 1990s." *Tobacco Control* 10 (2): 124–34.
- Gray, Virginia, and David Lowery. 1993. "The Diversity of State Interest Group Systems." *Political Research Quarterly* 46 (1): 81–97.
- Gray, Virginia, John Cluverius, Jeffrey J. Harden, Boris Shor, and David Lowery. 2015. "Party Competition, Party Polarization, and the Changing Demand for Lobbying in the American States." *American Politics Research* 43 (2): 175–204.
- Gray, Virginia, and David Lowery. 1996. *The Population Ecology of Interest Representation: Lobbying Communities in the American States*. Ann Arbor, MI: University of Michigan Press. <http://www.press.umich.edu/14367> (December 14, 2022).
- Gray, Virginia, David Lowery, and Jennifer K. Benz. 2013. *Interest Groups and Health Care Reform across the United States*. Washington, DC: Georgetown University Press.

- Hacker, Jacob S., Alexander Hertel-Fernandez, Paul Pierson, and Kathleen Thelen. 2022. "The American Political Economy: Markets, Power, and the Meta Politics of US Economic Governance." *Annual Review of Political Science* 25 (1): 197–217.
- Hacker, Jacob S., and Paul Pierson. 2014. "After the 'Master Theory': Downs, Schattschneider, and the Rebirth of Policy-Focused Analysis." *Perspectives on Politics* 12 (3): 643–62.
- Hall, Galen, Joshua Basseches, Rebecca Bromley-Trujillo, and Trevor Culhane. 2024. "Replication Data for: CHORUS: A New Dataset of State Interest Group Policy Positions in the United States." UNC Dataverse, V1, UNF:6:/8mtS0YGalad3W5ah8gJ9g==[fileUNF]. <https://doi.org/10.15139/S3/RPU1QP>
- Hojnacki, Marie, David C. Kimball, Frank R. Baumgartner, Jeffrey M. Berry, and Beth L. Leech. 2012. "Studying Organizational Advocacy and Influence: Reexamining Interest Group Research." *Annual Review of Political Science* 15 (1): 379–99.
- Holyoke, Thomas T. 2019. "Strategic Lobbying to Support or Oppose Legislation in the US Congress." *Journal of Legislative Studies* 25 (4): 533–52.
- Holyoke, Thomas T., and Jeff Cummins. 2020. "Interest Group and Political Party Influence on Growth in State Spending and Debt." *American Politics Research* 48 (4): 455–66.
- Karol, David. 2019. *Red, Green, and Blue: The Partisan Divide on Environmental Issues.* *Elements in American Politics*. <https://www.cambridge.org/core/elements/red-green-and-blue/D95292C12340F508E0D1BC8945649D1C> (July 23, 2022).
- Kim, In Song. 2017. "Political Cleavages within Industry: Firm-Level Lobbying for Trade Liberalization." *American Political Science Review* 111 (1): 1–20.
- Kim, In Song. 2018. "Lobbyview: Firm-Level Lobbying & Congressional Bills Database." Unpublished manuscript, Cambridge, MA: MIT. <http://web.mit.edu/insong/www/pdf/lobbyview.pdf>. Google Scholar Article Location.
- Kim, In Song, and Dmitriy Kunisky. 2021. "Mapping Political Communities: A Statistical Analysis of Lobbying Networks in Legislative Politics." *Political Analysis* 29 (3): 317–36.
- Kroeger, Mary. 2022. "Groups as Lawmakers: Group Bills in a US State Legislature." *State Politics & Policy Quarterly* 22 (2): 204–25.
- Kukkonen, Anna, Tuomas Ylä-Anttila, and Jeffrey Broadbent. 2017. "Advocacy Coalitions, Beliefs and Climate Change Policy in the United States." *Public Administration* 95 (3): 713–29.
- LaPira, Timothy M., and Herschel F. Thomas. 2020. "The Lobbying Disclosure Act at 25: Challenges and Opportunities for Analysis." *Interest Groups & Advocacy* 9 (3): 257–71.
- Lax, Jeffrey R., and Justin H. Phillips. 2012. "The Democratic Deficit in the States." *American Journal of Political Science* 56 (1): 148–66.
- Lowery, David, Virginia Gray, Jennifer Benz, Mary Deason, Justin Kirkland, and Jennifer Sykes. 2009. "Understanding the Relationship between Health PACs and Health Lobbying in the American States." *Publius: The Journal of Federalism* 39 (1): 70–94.
- Lowery, David, Virginia Gray, and John Cluverius. 2015. "Temporal Change in the Density of State Interest Communities: 1980 to 2007." *State Politics & Policy Quarterly* 15 (2): 263–86.
- McConnell, Grant. 1966. *Private Power and American Democracy*. New York: Random House.
- Nownes, Anthony J., and Patricia Freeman. 1998. "Interest Group Activity in the States." *Journal of Politics* 60 (1): 86–112.
- Nownes, Anthony J., and Daniel Lipinski. 2005. "The Population Ecology of Interest Group Death: Gay and Lesbian Rights Interest Groups in the United States, 1945–98." *British Journal of Political Science* 35 (2): 303–19.
- Peixoto, Tiago P. 2015. "Inferring the Mesoscale Structure of Layered, Edge-Valued, and Time-Varying Networks." *Physical Review E* 92 (4): 042807.
- Peixoto, Tiago P. 2019. "Bayesian Stochastic Blockmodeling." In *Advances in Network Clustering and Blockmodeling*, eds. Patrick Doreian, Vladimir Batagelj and Anuška Ferligoj, 289–332. New York: John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119483298.ch11> (August 31, 2022).
- Poole, Keith T., and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35 (1): 228–78.
- Sabatier, Paul A. 1988. "An Advocacy Coalition Framework of Policy Change and the Role of Policy-Oriented Learning Therein." *Policy Sciences* 21 (2): 129–68.
- Salisbury, Robert H., John P. Heinz, Edward O. Laumann, and Robert L. Nelson. 1987. "Who Works with Whom? Interest Group Alliances and Opposition." *American Political Science Review* 81 (4): 1217–34.

- Shor, Boris, Christopher Berry, and Nolan McCarty. 2010. "A Bridge to Somewhere: Mapping State and Congressional Ideology on a Cross-Institutional Common Space." *Legislative Studies Quarterly* 35 (3): 417–48.
- Shor, Boris, and Nolan McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105 (3): 530–51.
- Squire, Peverill. 2017. "A Squire Index Update." *State Politics & Policy Quarterly* 17 (4): 361–71.
- Stokes, Leah Cardamore. 2020. *Short Circuiting Policy: Interest Groups and the Battle Over Clean Energy and Climate Policy in the American States*. Oxford: Oxford University Press.
- The Graph-Tool Python Library. 2014. https://figshare.com/articles/dataset/graph_tool/1164194/14 (November 29, 2022).
- Thieme, Sebastian. 2018. *Ideology and Extremism of Interest Groups: Evidence from Lobbyist Declarations in Three States*. Rochester, NY: Social Science Research Network. SSRN Scholarly Paper. <https://papers.ssrn.com/abstract=2950719> (February 26, 2021).
- Thieme, Sebastian. 2019. "Moderation or Strategy? Political Giving by Corporations and Trade Groups." *Journal of Politics* 82 (3): 1171–75.
- Thieme, Sebastian. 2021. "A Direct Test of Legislative Gatekeeping." *Legislative Studies Quarterly* 46 (4): 855–88.
- U.S. Energy Information Administration - EIA - Independent Statistics and Analysis. n.d. <https://www.eia.gov/state/> (December 14, 2022).

Author biographies. Galen Hall is a PhD Student in Sociology and Physics at the University of Michigan.

Jossua Basseches is the David and Jane Flowerree Assistant Professor of Environmental Studies and Public Policy at Tulane University.

Rebecca Bromley-Trujillo is an associate professor of Political Science and Research Director of the Wason Center for Civic Leadership at Christopher Newport University.

Trevor Culhane is a researcher in the Climate and Development Lab at Brown University.

Cite this article: Hall, Galen, Joshua A. Basseches, Rebecca Bromley-Trujillo, and Trevor Culhane. 2024. CHORUS: A New Dataset of State Interest Group Policy Positions in the United States. *State Politics & Policy Quarterly* 1–26, doi:10.1017/spq.2024.6