CHORUS: A new dataset of state interest group policy positions in the United States

Supplemental Methods

1. Interest Group Deduplication

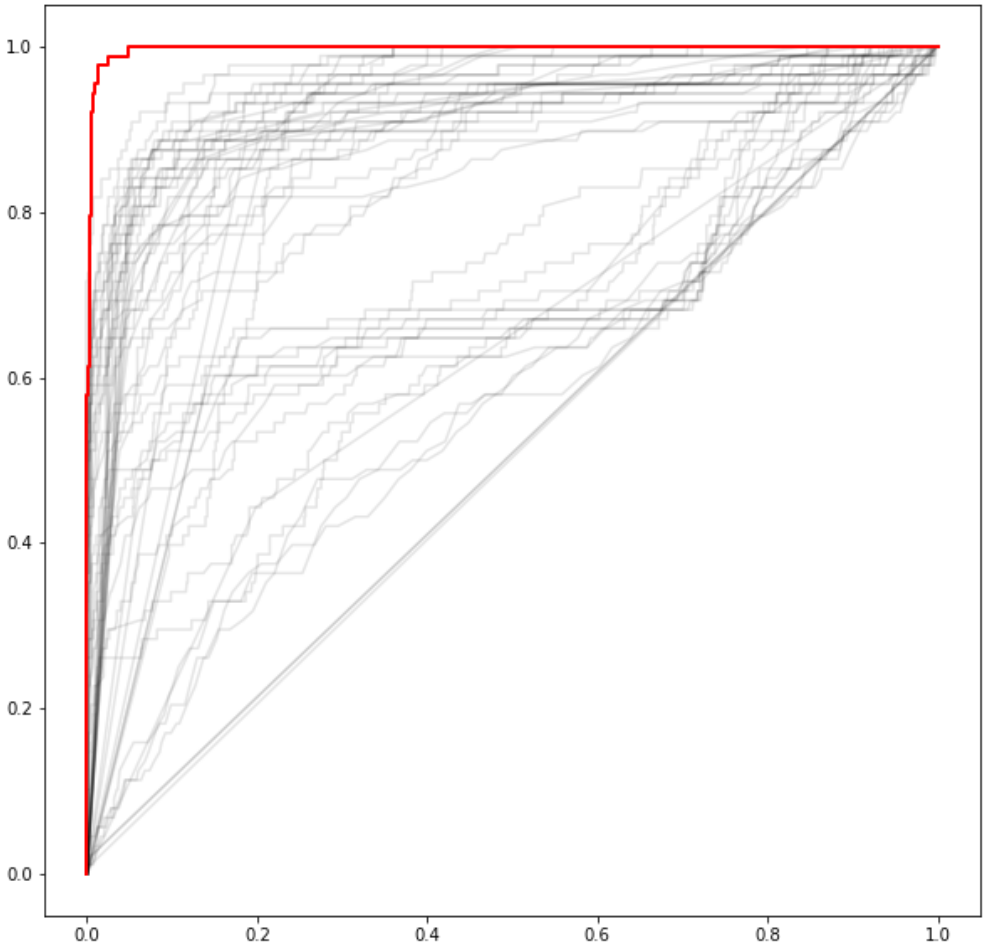
The set of unique interest group names collected from each state's available data is larger than the set of actually-unique interest groups active in the state. This is because interest groups routinely use different spellings, abbreviations, suffixes and prefixes; misspell their names in reporting documents; and change names or register positions under a parent organization. We therefore created a deduplication pipeline to efficiently and accurately merge multiple references to the same organization under one unique identifier. It is extremely difficult to reach 100% accuracy with deduplication, because of the vast number of names that must be compared. Given N unique client names, there are N^2 possible pairs which are confirmed matches — i.e. both refer to the same entity — which for our dataset quickly reaches an intractably high number. To make this task tractable, then, we combined some simplifying assumptions that allowed us to compare smaller numbers of names at a time (a method called *blocking*), an efficient machine learning pipeline to identify potential matches within that constrained set, references to multiple external datasets of entities to improve linkage frequency, and human verification of identified matches (which are typically much smaller in number). Our primary goal was to minimize false positives in which two similar names which refer to different entities get merged, thereby combining all the records associated with both entities. Such combinations can heavily skew the results of analysis. If *Acadia Investments* and *The Acadia Center* get merged in deduplication, suddenly we have an organization which appears to lobby both on financial regulations and environmental policy, which will get grouped very differently by the stochastic block model than either of the component organizations.

We outline each of the steps in our deduplication process below.

1. Blocking. We blocked our dataset by only considering name pairs which had a greater than 20% cosine similarity using Tf-idf featurization, and which did not contain non-matching sequences of numbers. The top 50 matches by cosine similarity are easy to calculate for all entities in our dataset, and pairs with lower than 20% cosine similarity are empirically extremely unlikely to match. Likewise, pairs with different numbers in the name – such as different union locals – are almost always distinct entities.
2. Text featurization. For every candidate pair, we generated many measures of phonetic and semantic (dis)similarity, including hand-made features such as the number of first letters shared by both, or a boolean indicating whether an acronym in one name spells out the first letters of words in the other. We included multiple variants of Tf-idf cosine similarity as well as composite measures from a fuzzy string-matching package (Bachmann 2022). We also include a measure of semantic similarity created using neural network embeddings of sentences created from a BERT model (Reimers and Gurevych 2019). The cosine similarities of these embeddings give a rough general-purpose measure of the semantic similarity of the two different names, which can be useful for identifying matches separated by a synonymous substitution. Nearly all of the semantic

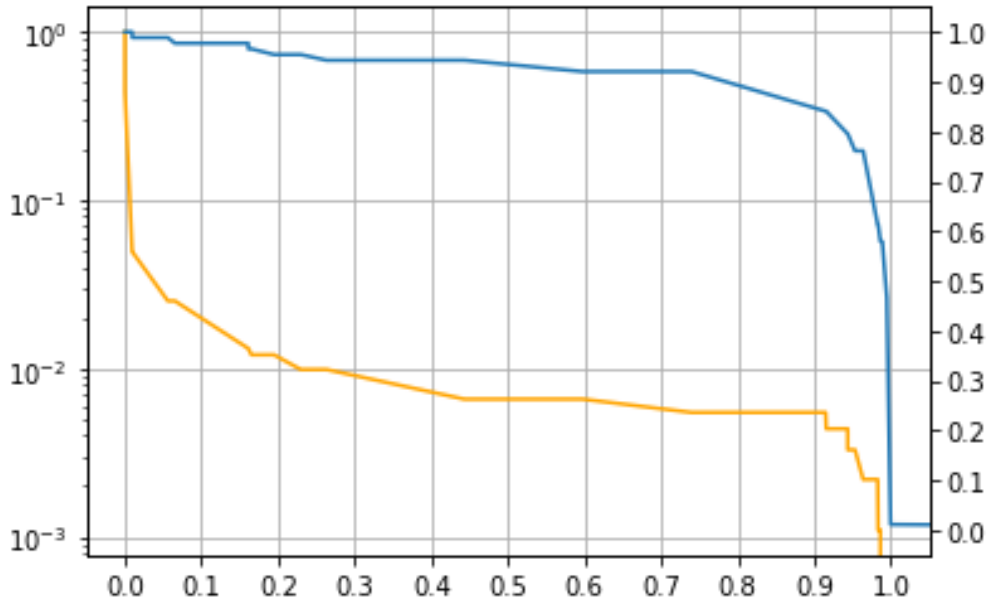
similarity measures were calculated both before and after removing any US state names or abbreviations from the entity names, to account for the common practice of sometimes including "[STATE NAME]" in a trade association or nonprofit's reported name and other times removing it.

3. To automatically identify matches using the featurized pairs of names, we trained a machine learning model on a dataset of ~12,000 labeled pairs, iteratively training on a batch of 500 and then extending the batch by adding the 500 most difficult-to-predict examples from the remaining training set until the out-of-sample F1 score was maximized. Many different ML models can serve for this type of task, but we ultimately settled on the CascadeForest, an augmented, layered variant of the common Random Forest model, because it performed the best (Zhou and Feng 2020).
4. After training, the best model achieved a very high ROC-AUC score of **0.99** on the held-out set of 1,000 randomly-sampled examples, as well as an F1-score of **92.6%** and accuracy of **98.7%**. The figure below shows the ROC-AUC plot of every individual similarity measure used, in gray, as well as our estimator in red. A line closer to a right angle is better; our estimator performs very well for this sort of task.



5. To ensure a very low rate of false positive matches, we retained only all matches with P(match) above 90% for hand checking. In the chart below, the left y-axis (orange line) is the false positive rate; the right y-axis (blue line) is the true positive rate; and the x axis is the probability threshold

of the predictions. Predictions above 90% confidence have a very low false positive rate but also fail to capture many real matches; even above 50% confidence, around 5% of real matches aren't caught.



6. To boost the number of valid entity matches identified, and to provide useful metadata on many of our entities, we performed the same entity-matching procedure by comparing our dataset's entities to those from three external datasets: first, FollowTheMoney's dataset of registered lobbyists in nearly every State, from which we can gather the sector, industry, and business classification of many of our interest groups, as well as other associated data on lobbying spending and political contributions which FTM collects; second, OpenSecrets' database of registered Congressional lobbyists and interest groups, which gave us one means of potentially linking national interest groups across states, and information on their national-level political activities; and finally, Google Knowledge Graph's entity information database, which via their search API was often able to connect entities with very different names which were nonetheless the same because of either a corporate rebranding or a merger.
7. Finally, once every match between our dataset of interest group names and itself, FollowTheMoney, OpenSecrets, and Google Knowledge graph was by-hand confirmed, in each state we created a network linking all names with a verified match and assigned one unique client ID to each connected component of the resulting graph. This step follows standard deduplication procedure (Binette and Steorts 2022).

2. Web Scraping & Parsing

To construct custom web scrapers and parsers in Python for each of the 17 states, we first identified where the positions data was stored on each legislature's website. This involved reviewing the website structure and relevant web pages to locate the information we needed. Once the relevant web pages were identified, we wrote scrapers in the Scrapy library to download the raw web pages or documents. These were then

stored in Google Cloud to have a permanent copy of each website. Next, we wrote parsing scripts to transform the scraped html tables into datasets and to extract positions data from unstructured or semi-structured text sources using regular expressions. This involved defining the necessary regular expressions and using them to extract the desired information from the downloaded web pages or documents. Once the data was extracted, we wrote custom cleaning scripts for each state to ensure that the data was in a consistent format and ready for analysis. This involved checking for missing values, correcting errors, and standardizing the format of the data. Finally, we loaded all of the resulting data together in one file to be used for analysis. This allowed us to easily access and compare the position data from each state.

3. Database Linkage

We linked our dataset to external sources of information on interest groups (FollowTheMoney and OpenSecrets; see Deduplication section), and sources of metadata on legislation. We include two sources of legislative metadata: first, Legiscan's open access dataset of bill histories for all US states, which typically extends back to around 2009; and second, the National Conference on State Legislature's datasets of bills collected under different topic areas. NCSL collects bills within 13 broad topic areas and 478 specific topic areas. Legiscan and NCSL data were merged by creating unique unified identifiers for every bill, and the resulting dataset was again merged with our own dataset of bills collected from lobbying and testimony positions.

4. Network construction and Stochastic Block Models

To construct a signed bipartite network linking interest groups and bills, we first identified all of the interest groups and bills that were included in the positions data. We then created a network with two sets of nodes, one for the interest groups and one for the bills. Next, we added edges to the network that linked each interest group to the bills on which they took a position. The edge weights were set to indicate the interest group's position on the bill, with +1 indicating support, -1 indicating opposition, and 0 indicating neutrality.

A stochastic block model (SBM) is a probabilistic graphical model that partitions a set of nodes into distinct groups, or blocks, such that the probability of an edge between nodes depends only on the blocks to which the nodes belong. SBMs are commonly used in network analysis to identify the underlying structure of a network and to make predictions about the formation and evolution of edges. The model is called "stochastic" because it includes a random component that captures the uncertainty inherent in real-world networks.

The graph-tool package is a powerful open-source tool for analyzing and manipulating graphs in Python. It includes a number of algorithms for estimating the parameters of a stochastic block model (SBM) from observed data, including the block structure, the block assignment of nodes, and the edge probabilities. These algorithms typically use an iterative approach, starting with an initial guess for the model parameters and then refining the estimates through a series of steps .

In cases such as ours, it is useful to incorporate additional information about the edges into the model, such as categorical covariates that describe the type or nature of the relationship between nodes. This can be accomplished by extending the basic SBM to include edge covariates. In this type of model,

the probability of an edge between two nodes still depends on the blocks to which the nodes belong, but it also depends on the value of the edge covariate. This allows the model to capture more nuanced patterns in the network and to make more accurate predictions about the formation and evolution of edges. To estimate the parameters of an SBM with edge covariates, one can use a modified version of the standard algorithms for estimating SBMs, implemented in the graph-tool package. We include a single edge covariate indicating whether each position was “support” or “oppose”. Although “neutral” positions can easily be included in the model as a third possible value for the covariate, they do not typically improve the resolution or intuitive validity of the results, because organizations with starkly opposing interests may often state “neutral” positions on the same piece of legislation.

SBMs were created and estimated for networks constructed from position data of each record type (lobbying or testimony) and from each state independently. Estimation was completed on a Google Colab notebook which is available upon request. A state with two types of records gets two associated SBMs. We do not combine lobbying and testimony data from the same state due to the dramatic differences in the mode of collection, composition, and relevant costs and benefits associated with disclosure in both data types. The BlockModel objects which store the SBM results were saved for use later in the analysis and in future analyses; they can also be iteratively refined even after adding additional records to the dataset. In every case, we estimated the SBM for a 5-core of our bipartite network in which every interest group lobbied on at least five bills and every bill had at least five support/oppose positions stated by interest groups.

5. Data Quality and Additional State Positions Data

5.1. Lobbying Data Quality

State	Additional Information Available	Notes	File Format	Example Record	Relevant State Law	Penalties / Stringency
Colorado	Lobbying spending data available		CSV	https://www.sos.state.co.us/lobby/SearchSubject.do	24-6-301 (Colorado Sunshine Law)	Increasing fines for each day a report is late. “The Secretary of State’s Office may impose fines, suspend, revoke or bar a person or entity from registration, refer the matter to the General Assembly, provide notice to the General Assembly when a substantial violation has occurred, apply to the district court for the issuance of an order in accordance with Section 24-6-309(2), C.R.S., or determine another remedy in accordance with Section 24-6-301, C.R.S.” - see https://www.sos.state.co.us/pubs/lobby/files/guidanceManual.pdf
Iowa	Spending data by lobbying firm for each client is available. Comments on positions are available but rare.		HTML table	https://www.legis.iowa.gov/lobbyist/reports/declarations	Iowa Code 2023, Chapter 68B	Civil penalty of not more than two thousand dollars for each violation. Possible revocation of lobbying license.
Massachusetts	Position on each stage of a bill's progress	Data provided via email from the Massachusetts legislature.	Excel		Session Law - Acts of 2009 Chapter 28	Violation of Massachusetts lobbying rules may be punished by: a fine between $100 and $10,000; and/or imprisonment in state prison for not more than 5 years, or in a jail or house of correction for more than 2.5

						years. G.L. § 3-48.
Montana	Lobbying spending by session; subjects lobbied on		JSON	https://lobbyist-ext.mt.gov/LobbyistRegistration/public/searchRegistry/home	MCA 5-7	"A person who violates any of the provisions of this chapter is subject to civil penalties of not less than $250 and not more than $7,500 according to the discretion of the district court, as court of original jurisdiction. A lobbyist who violates any of the provisions of this chapter must have the lobbyist's license suspended or revoked according to the discretion of the court. Any legislator adjudged in violation of the provisions of this chapter is additionally subject to recall under the Montana Recall Act, Title 2, chapter 16, part 6, and the violation constitutes an additional basis for recall to those mentioned in 2-16-603(3)."
Nebraska	Hearing transcripts available on bill page (example: https://nebraskalegislature.gov/bills/view_bill.php?DocumentID=24605). Expenses available (example: https://nebraskalegislature.gov/lobbyist/view.php?link=view_lobbyist&id=2840)	Lobbyist statements of activity submitted each session, no position-specific dates	HTML table	https://nebraskalegislature.gov/lobbyist/view.php?link=view_form&form=form&RegistrationID=11930	Nebraska Regulations TITLE 4 - CHAPTER 6	For regular reports, late filing fee of $25 for each day, not to exceed $750 per statement. For special reports, late filing fee of $100 for each day for ten days. After the tenth day, an additional late filing fee of one percent of the amount of the receipts and expenditures which were required to be reported per day - not to exceed ten percent of the amount of the receipts and expenditures which were required to be reported. See: Sections 49-1463.02; 49-1483.03(2); and 49-1488.01, Neb. Rev. Stat.
New Jersey	Expenditures by client available. Scanned PDFs of lobbying records pre-2016 are available (PDF data not included in the CHORUS dataset)	Client names are non-unique: the same acronym can refer to multiple distinct organizations, which are undifferentiated in the lobbying records. Additional investigation into a particular record can reveal which organization lobbied for which bill.	Excel	https://www3-elec.mwg.state.nj.us/ELEC_AGAA/EntitySearch.aspx	19:25-20	"any person who is found to have committed a violation of the Act or this subchapter shall be liable for a civil penalty of up to $1,000 for that violation."

Rhode Island	Compensation, expenditures, expenses, contributions, lobbying subjects, executive agencies lobbied	Positions reported on monthly or quarterly basis	PDF table	https://risos-lrd-production-public.s3.amazonaws.com/reports/session_2018/period_12/610_12235_report.pdf	Chapter 139.1 The Rhode Island Lobbying Reform Act	"appropriate relief, which may include an order to pay a civil penalty of up to five thousand dollars ($5,000) per violation, and revocation of the applicable registration for a period of up to three (3) years.", i.e. penalties of up to $5,000 per violation and revocation of the entity's right to lobby in Rhode Island for up to 3 years.
Wisconsin	Comments on some positions available. Effort and lobbying spending information available for each bill. Example: https://lobbying.wi.gov/Who/PrincipalInformation/2019REG/Information/8149?tab=Profile		HTML table	https://lobbying.wi.gov/Who/PrincipalInformation/2019REG/Information/8298?tab=Profile	Wisconsin Statutes Chapter 13 subchapter III	Various fines per late or unreported lobbying interest - see page 50 of https://ethics.wi.gov/Resources/LobbyingBestPracticesandOverview.pdf

Table A-1: *Lobbying Data Quality*

5.2. Testimony Data Quality

State	Additional Information Available	Notes	Data Type	File Format	Example Record
Arizona	Bill and sponsorship information available as json	Individuals and organizations can register positions online without attending a hearing or writing a letter via the Request to Speak system: https://apps.azleg.gov/RequestToSpeak	Online registration system	JSON	https://apps.azleg.gov/api/Bill/?calendarId=10000&includePositions=true&includeSponsors=true&includeActions=false&includeTransmitted=false
Colorado		There is significant variation in the language describing positions - the records lack a standardized format that is easily parseable.	Committee hearing notes	HTML unstructured text	https://www.leg.state.co.us/CLICS2006A/commsumm.nsf/91320994cb8e0b6e8725681d005cb995/70d082620870dc76872571380065aff2?OpenDocument
Florida		Only House meeting appearances are digitized - Senate meeting appearances are scanned handwritten records,	Witness list	JSON	https://www.myfloridahouse.gov/LD/default.aspx

		and are not included in the dataset.			
Illinois		Inconsistent distinction between “Firm Business or Agency” and “Representation” columns in source data - same organization may use fields interchangeably, which negatively impacts data quality.	Witness list	HTML table	https://ilga.gov/legislation/witnessslip.asp?DocNum=1&DocTypeID=HB&LegID=83490&GAID=13&SessionID=88&GA=99&SpecSess=
Kansas	Testimony texts available.	Data in the CHORUS dataset is from a consistently formatted source and usually has written testimony attached to each position. Some additional testimony available in meeting minutes pdfs, as well as sporadic links on committee pages. These are not scraped as their formatting and reporting is inconsistent.	Meeting Testimony	HTML table	http://www.kslegislature.org/li_2014/b2013_14/committees/ctte_h_ins_1/documents/date-choice-2014-03-17/
Maryland	Testimony texts available.		Witness list	HTML table	https://mgaleg.maryland.gov/mgawebsite/Legislation/WitnessSignup/HB0001?ys=2020rs
Missouri		Testimony lists with positions for house and senate starting in 2019 - records do not differentiate between individuals and organizations. House committee reports have witnesses since 2000, but are formatted badly so not included in the dataset. Senate minutes available via FTP.	Committee minutes	PDF (unstructured)	www.house.mo.gov/billtracking/bills151/sumpdf/HB0830C.pdf

Montana	Committee hearing audio and video available		Committee minutes	PDF (unstructured)	https://sg001-harmony.sliq.net/00309/Harmony/en/PowerBrowser/ViewHandoutFile?contentEntityId=40100&handoutId=65070
Ohio	Testimony texts available.		Committee minutes	HTML table	https://www.legislature.ohio.gov/legislation/legislation-committee-documents?id=GA131-HB-214
South Dakota		Records do not differentiate between individuals and organizations.	Committee minutes	HTML unstructured text	https://mylrc.sdlegislature.gov/api/Documents/98056.html#page=620
Texas		Many formats of testimony; PDFs with records spanning multiple pages can be difficult to parse. Witness lists available via FTP.	Witness list	HTML/PDF unstructured text	https://capitol.texas.gov/tlodocs/85R/witlistmtg/html/C420201806280900 1.HTM

Table A-2: *Testimony Data Quality*