Replicating
Text as Policy: Measuring Policy Similarity through Bill Text Reuse

Fridolin Linder, Bruce Desmarais, Matthew Burgess, and Eugenia Giraudy (2018)

Goal of the Study

- Identification of comparable policy actions (e.g., adoption, consideration) across U.S. states

- Approach policy adoption in terms of a continuum of similarity rather than a dichotomy of equivalence

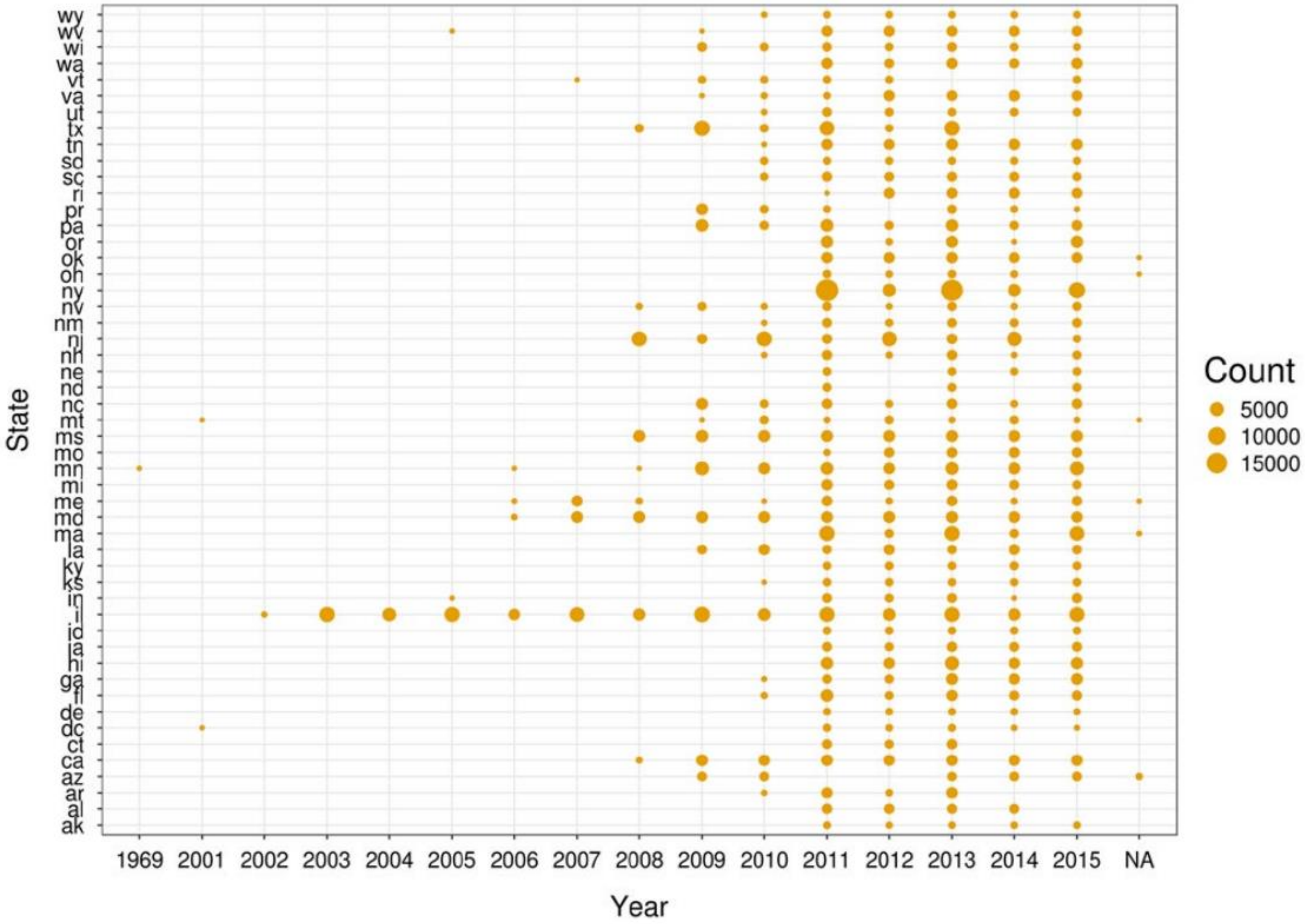
- Latent variable measured through text reuse: policy similarity

Study Data Pipeline

- Acquire raw data 500,000 State Bills + metadata from 2008-2015 (Burgess et al. (2016) and the Sunlight Foundation)

Study Data Pipeline

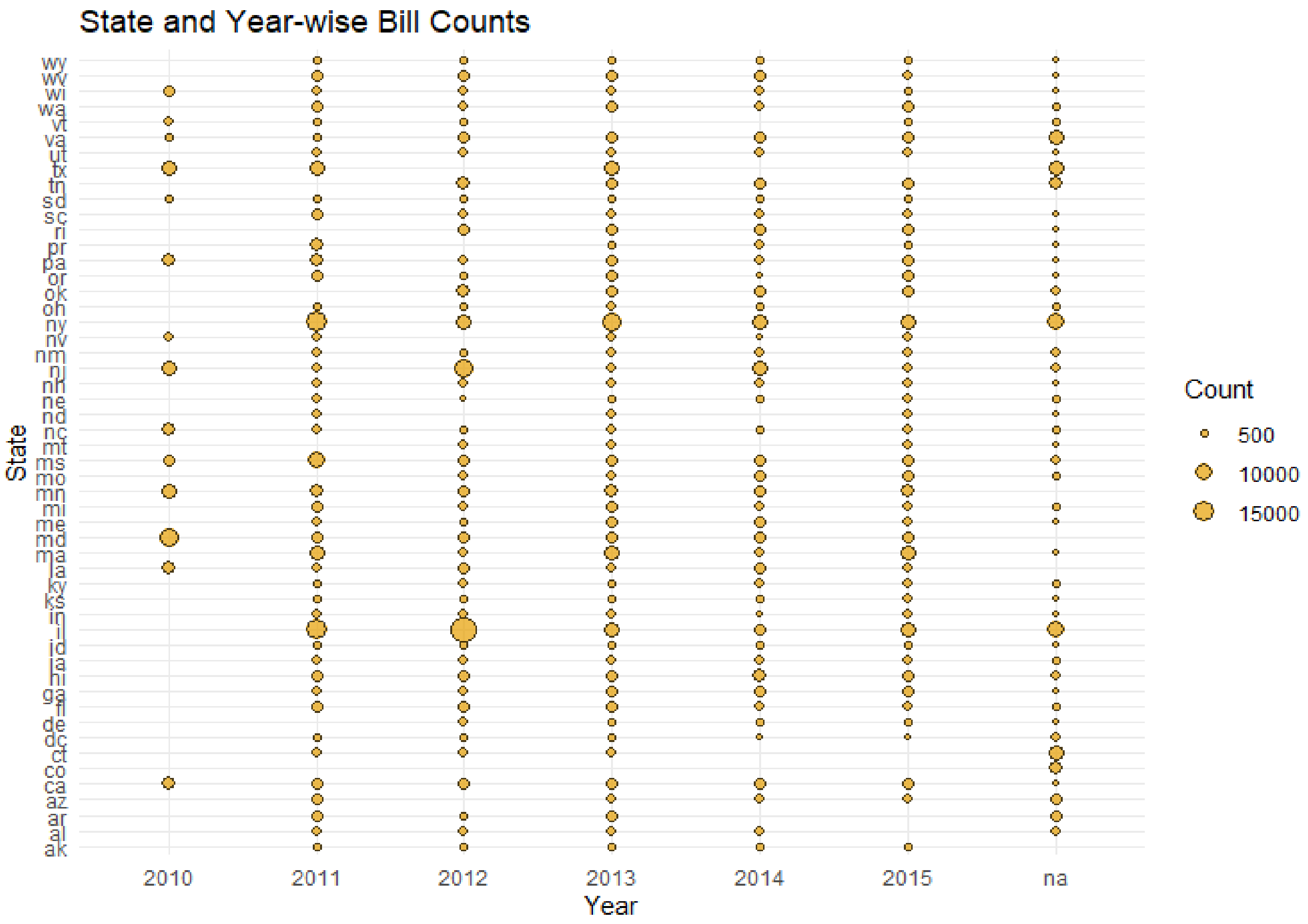
- Acquire raw data 500,000 State Bills + metadata from 2008-2015 (Burgess et al. (2016) and the Sunlight Foundation)



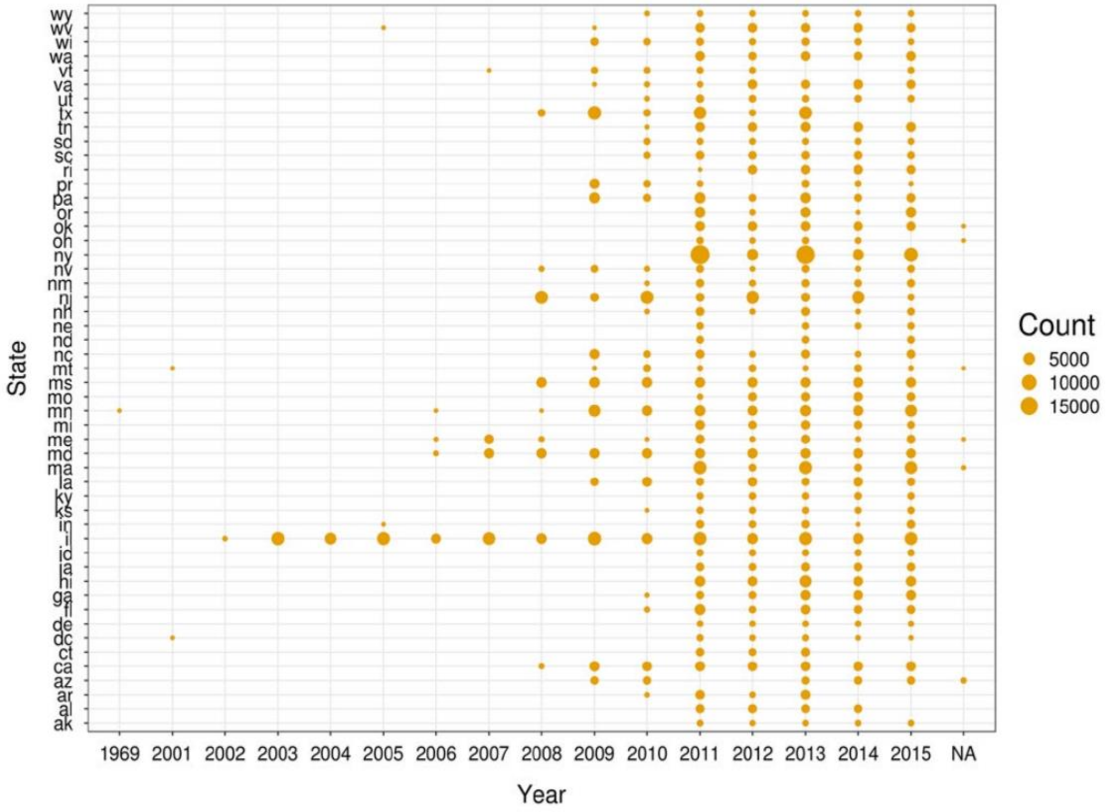
Study Data Pipeline

Acquire raw data 500,000 State Bills + metadata from 2008-2015 (Burgess et al. (2016) and the Sunlight Foundation)

❑ **Replication Step: Utilize Raw bill data made available by authors, n = 571554**



Replication Visualization



Authors' Visualization

Study Data Pipeline

1. Acquire raw data 500,000 State Bills + metadata from 2008-2015 (Burgess et al. (2016) and the Sunlight Foundation)

2. Query the 500 most similar bills each bill using the ElasticSearch algorithm

Study Data Pipeline

1. Acquire raw data 500,000 State Bills + metadata from 2008-2015 (Burgess et al. (2016) and the Sunlight Foundation)

2. Query the 500 most similar bills each bill using the ElasticSearch algorithm

    ❑ Replication Step: Not viable, due to cost of implementation: cloud deployment + API purchase

Study Data Pipeline

1. Acquire raw data 500,000 State Bills + metadata from 2008-2015 (Burgess et al. (2016) and the Sunlight Foundation)

2. Query the 500 most similar bills each bill using the ElasticSearch algorithm
   - ❑ Replication Step: Not viable, due to cost of implementation: cloud deployment + API purchase

3. For each bill pair in use the Smith-Waterman algorithm to find the longest aligned text sequence

4. For each bill pair in use the boilerplate algorithm to reweight its alignment score

Study Data Pipeline

1. Acquire raw data 500,000 State Bills + metadata from 2008-2015 (Burgess et al. (2016) and the Sunlight Foundation)

2. Query the 500 most similar bills each bill using the ElasticSearch algorithm
    - ❑ Replication Step: Not viable, due to cost of implementation: cloud deployment + API purchase

3. For each bill pair in use the Smith-Waterman algorithm to find the longest aligned text sequence

4. For each bill pair in use the boilerplate algorithm to reweight its alignment score

    - ❑ Due to resource constraints, replication of the process leading us from the raw file to the alignments was not possible.

Information Availability for Replication

	Available to Authors	Available for Replication
Raw Bill Data	Yes	Yes, but somewhat different version
Alignments Data	Yes	Yes
NCSL Data	Yes	No
State Diffusion Tie Data	Yes	Yes
Ideology Data	Yes	Yes
Names of Sponsors in Raw Data	Yes	No but technically Yes

Desmarais et al. (2015)

Shor and McCarty (2011)

Open States API, Legiscan API, manual matching

Robustness Checks

1) Comparing Alignment Scores with NCSL Data

❑ Assessing how well alignment scores predict whether bills are in the same NCSL table

❑ Replication Step: Not viable for replication. NCSL Data unavailable.

❑ Scraping of NCSL procedure not replicable, due to significant changes in Website structure.

❑Scraping also requires access to the Bing search API through Microsoft Azure – cost constrained

Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) **Diffusion Networks and Text Reuse**

Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) **Diffusion Networks and Text Reuse**

REPLICABLE !

Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) **Diffusion Networks and Text Reuse**

- Evaluate whether text reuse corresponds to the transfer of policy: presence of a diffusion network tie between two states as a predictor of text reuse

- A tie from state i to state j in the diffusion network indicates that state j has frequently emulated state i’s policies in the preceding 35 years.

- Expectation: Diffusion tie between two states should be predictor of text reuse .

Robustness Checks

1) Comparing Alignment Scores with NCSL Data

2) Diffusion Networks and Text Reuse

- Evaluate whether text reuse corresponds to the transfer of policy: presence of a diffusion network tie between two states as a predictor of text reuse
- A tie from state i to state j in the diffusion network indicates that state j has frequently emulated state i's policies in the preceding 35 years.
- **Expectation: Diffusion tie between two states should be predictor of text reuse** .

Data:

- Alignment score between two bills. Each bill has associated state.
- **Dependent Variable**: Aggregate alignment scores across state pairs, over entire alignments dataset (dropping DC and Puerto Rico) . N = 2352
- **Independent Variables**:

1) State Diffusion Pairs (2008), acquired from Desmarais et al. (2015) data repo. Binary : 1: Tie exists, 0 no tie
2) Total number of bills for each state for all available years, acquired from raw bills data

Robustness Checks

1) Comparing Alignment Scores with NCSL Data

2) Diffusion Networks and Text Reuse

- **Expectation:** Diffusion tie between two states should be predictor of text reuse .
- **Data:**

❑ **Dependent Variable**: Aggregate alignment scores across state pairs, over entire alignments dataset (dropping DC and Puerto Rico) . N = 2352

❑ **Independent Variables**:

1) State Diffusion Pairs (2008), acquired from Desmarais et al. (2015) data repo . Binary : 1: Tie exists, 0 no tie
2) Total number of bills for each state for all available years, acquired from raw bills data

- **Methodology**

➢ Represent aggregate alignment scores between state pairs as 49 X 49 undirected matrix.

➢ Represent state diffusion ties between state pairs as 49 X 49 undirected matrix.

➢ Represent number of bills by each state as 49 X 49 matrix, by taking the outer product of the state wise counts with itself

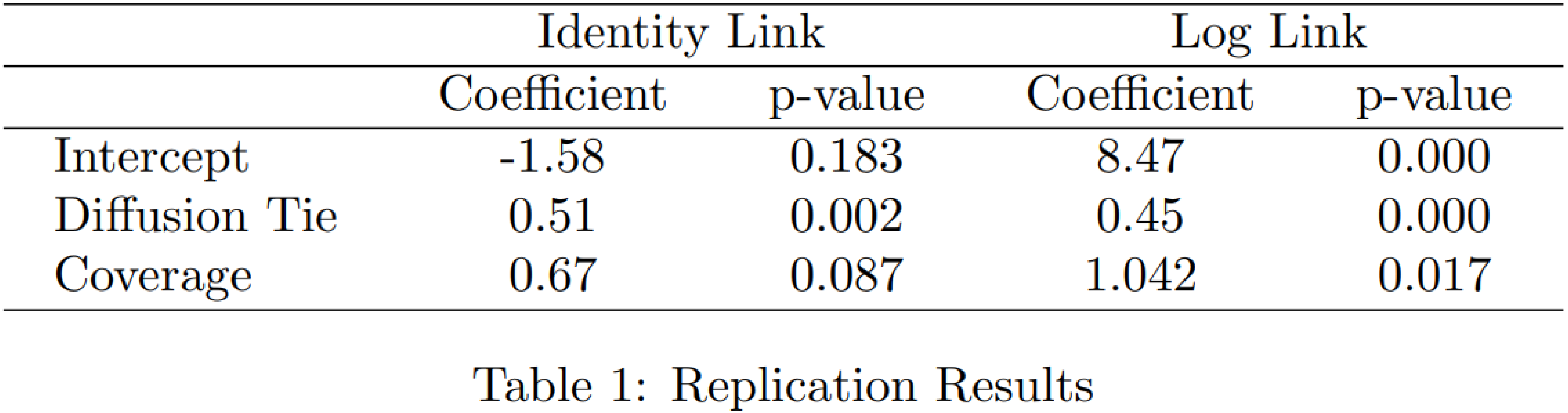
➢ Coefficients calculated with OLS regression and normalized with standard deviation of cross-state alignment scores. P-values found based on 1,000 QAP permutations (5000 in study), since pairs of states cannot be considered independent of each other in this case.

Robustness Checks

1) Comparing Alignment Scores with NCSL Data

2) Diffusion Networks and Text Reuse

Results



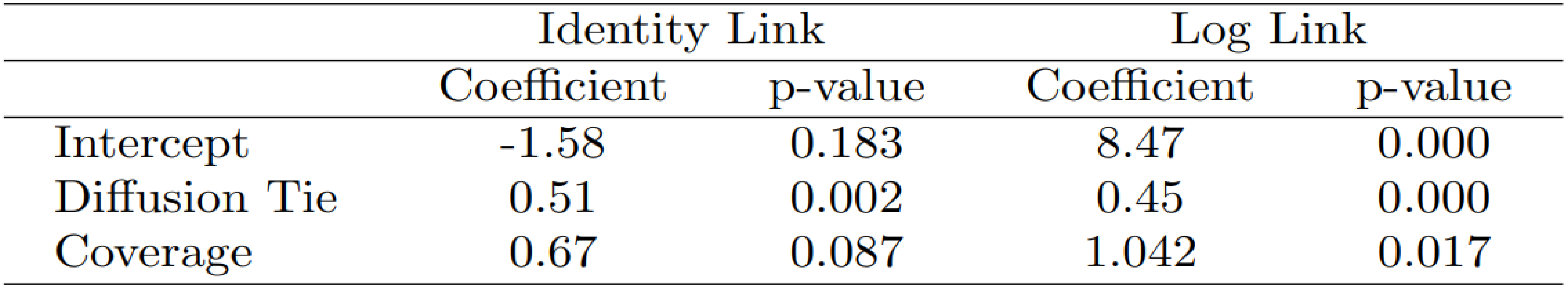
	Identity Link		Log Link	
	Coefficient	p-value	Coefficient	p-value
Intercept	-1.58	0.183	8.47	0.000
Diffusion Tie	0.51	0.002	0.45	0.000
Coverage	0.67	0.087	1.042	0.017

Table 1: Replication Results

Robustness Checks

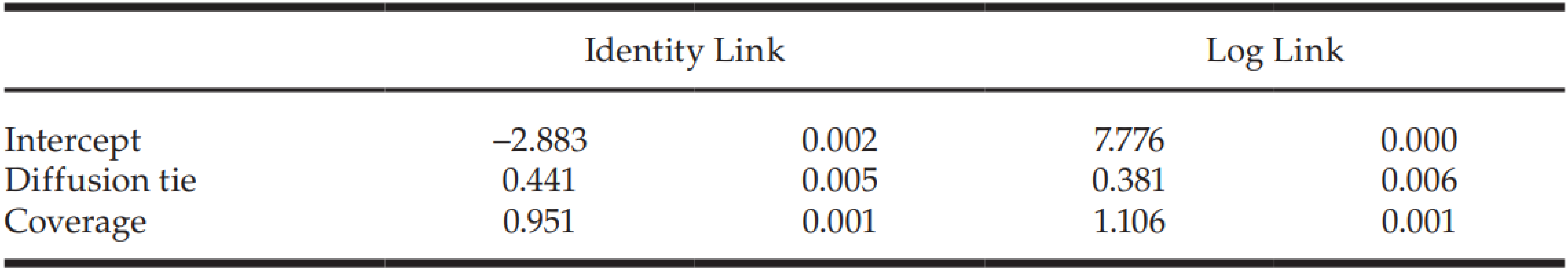
1) Comparing Alignment Scores with NCSL Data

2) Diffusion Networks and Text Reuse



	Identity Link		Log Link	
	Coefficient	p-value	Coefficient	p-value
Intercept	-1.58	0.183	8.47	0.000
Diffusion Tie	0.51	0.002	0.45	0.000
Coverage	0.67	0.087	1.042	0.017

Table 1: Replication Results



	Identity Link		Log Link	
Intercept	–2.883	0.002	7.776	0.000
Diffusion tie	0.441	0.005	0.381	0.006
Coverage	0.951	0.001	1.106	0.001

Table 2: Study Results

Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) Diffusion Networks and Text Reuse
3) **Ideological Distance between bill sponsors and bill text reuse**

- Find a bill’s sponsors and their ideology score as found by Shor and McCarty (2011)
- For each alignment pair, find squared difference between each bill’s average sponsor ideology.
- **Expectation: Bill pairs with low ideology distances should have higher alignment scores**

Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) Diffusion Networks and Text Reuse

3) **Ideological Distance between bill sponsors and bill text reuse**

- Find a bill’s sponsors and their ideology score as found by Shor and McCarty (2011)
- For each alignment pair, find squared difference between each bill’s average sponsor ideology.
- **Expectation: Bill pairs with low ideology distances should have higher alignment scores**
- **Data:**
  - **Dependent Variable**: Alignment Score between pair of bills - available for replication
  - **Independent Variable**: Ideological Distance between pairs of bills

Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) Diffusion Networks and Text Reuse
3) **Ideological Distance between bill sponsors and bill text reuse**

- Find a bill's sponsors and their ideology score as found by Shor and McCarty (2011)
- For each alignment pair, find squared difference between each bill's average sponsor ideology.
- **Expectation: Bill pairs with low ideology distances should have higher alignment scores**
- **Data:**
  - **Dependent Variable**: Alignment Score between pair of bills - available for replication
  - **Independent Variable**: Ideological Distance between pairs of bills
    - ❑Raw bills data: Names of sponsors available, however: highly variable formatting, no unique identifier for a sponsor
    - ❑Shor and McCarty Ideological scores: Available on archived repo. Name formatted consistently

Matching bills with ideological distance, in current state, is not viable!

Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) Diffusion Networks and Text Reuse

3) **Ideological Distance between bill sponsors and bill text reuse**

- **Expectation: Bill pairs with low ideology distances should have higher alignment scores**
- **Data:**
  - **Dependent Variable**: Alignment Score between pair of bills - available for replication
  - **Independent Variable**: Ideological Distance between pairs of bills
    - ❑ Raw bills data: Names of sponsors available, however: highly variable formatting, no unique identifier for a sponsor
    - ❑ Shor and McCarty Ideological scores: Available on archived repo. Name formatted consistently
    - ❑ Authors utilize Open States API to query each bill. Only keep those bills which return sponsor names through query. n = 88 million (40% of total number of alignments).
    - ❑ Query results not satisfactory. Replication instead uses Legiscan API.

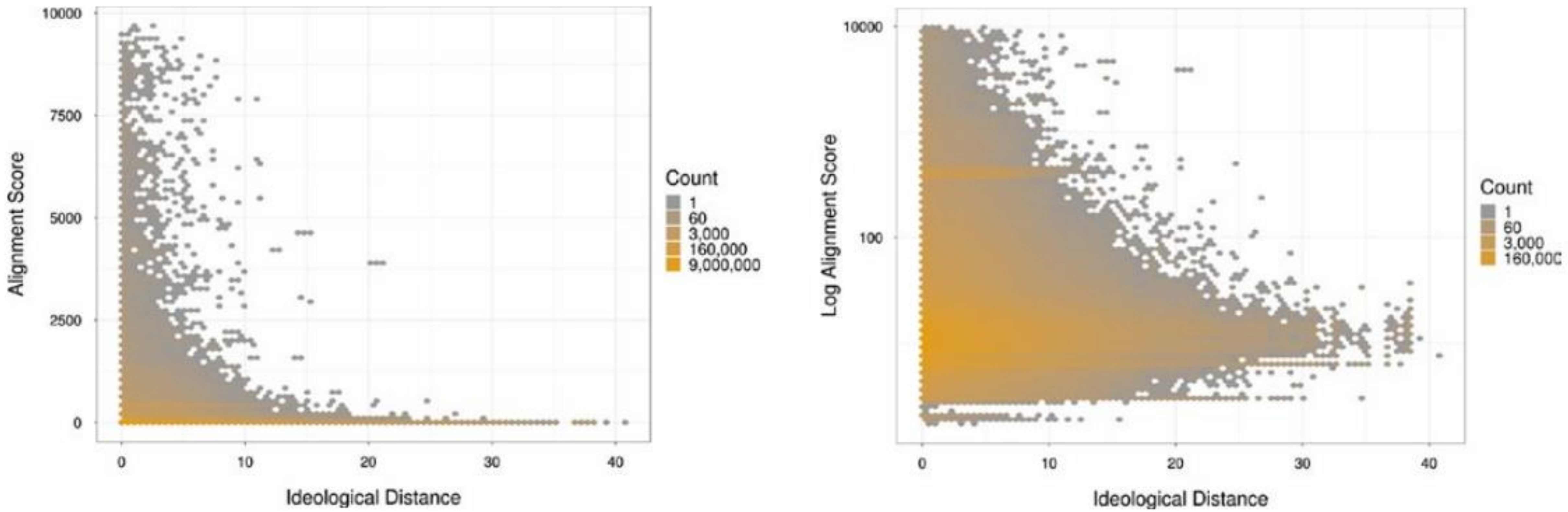
Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) Diffusion Networks and Text Reuse
3) **Ideological Distance between bill sponsors and bill text reuse**

- **Expectation: Bill pairs with low ideology distances should have higher alignment scores**
- **Data:**
  - **Dependent Variable**: Alignment Score between pair of bills - available for replication
  - **Independent Variable**: Ideological Distance between pairs of bills
    - ❑ Raw bills data: Names of sponsors available, however: highly variable formatting, no unique identifier for a sponsor
    - ❑ Shor and McCarty Ideological scores: Available on archived repo. Name formatted consistently
    - ❑ Authors utilize open states API to query each bill. Query results not satisfactory
    - ❑ Replication instead uses Legiscan API + Manual Matching.
    - ❑ **Instead of sampling raw bills; sample alignments (10) for state, year pair. Then use bill numbers from these sample alignments to find corresponding raw bills. Ensuring, sampled data has both left and right alignment bills.**
    - ❑ **Query these bills in Legiscan. Find sponsor names. QC results to have year and raw bill name partly match legiscan results' name. Finally manual matching based on logic to increase matches**

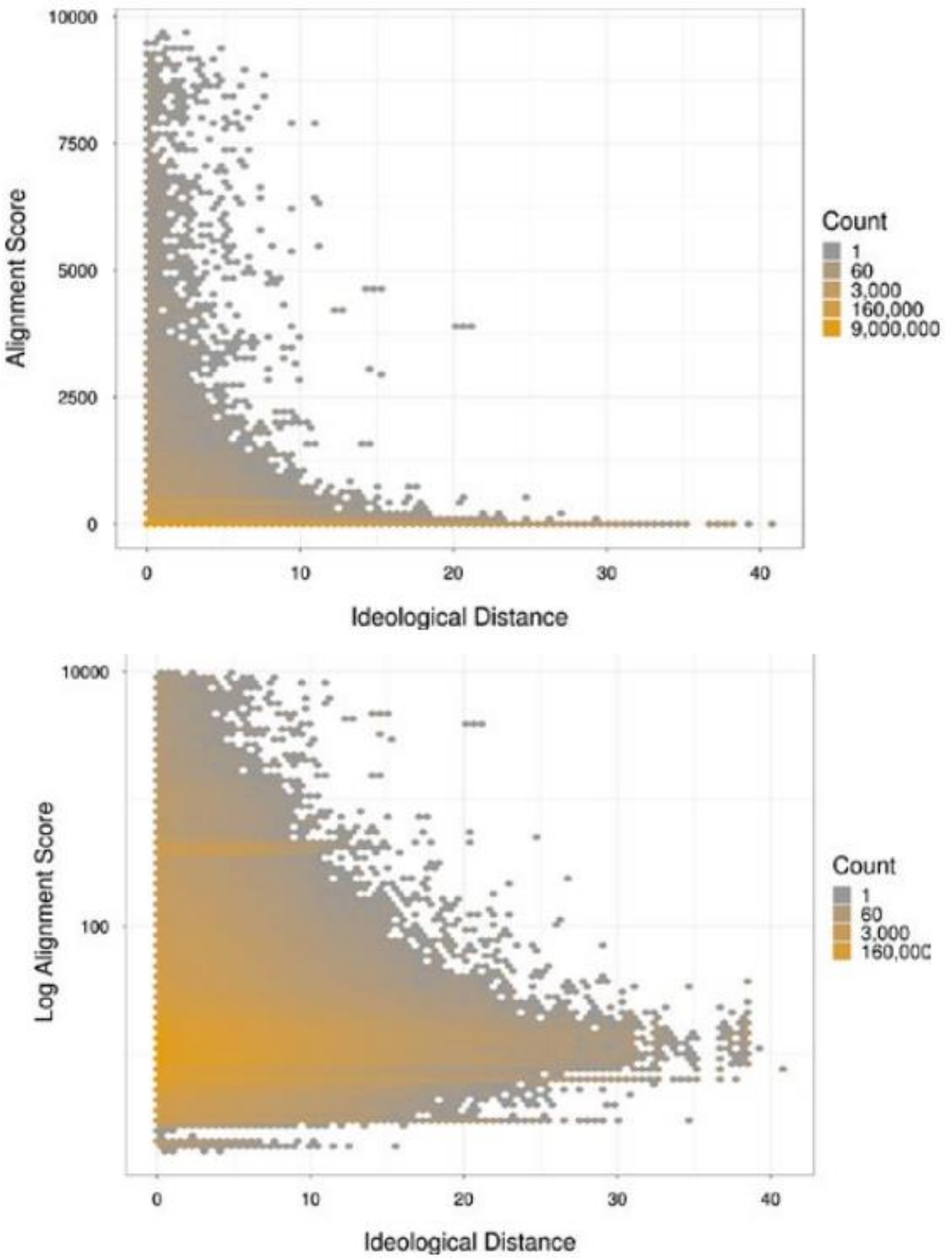
Instead of sampling raw bills; sample alignments (10) for state, year pair.

Possible issue with sampling ?

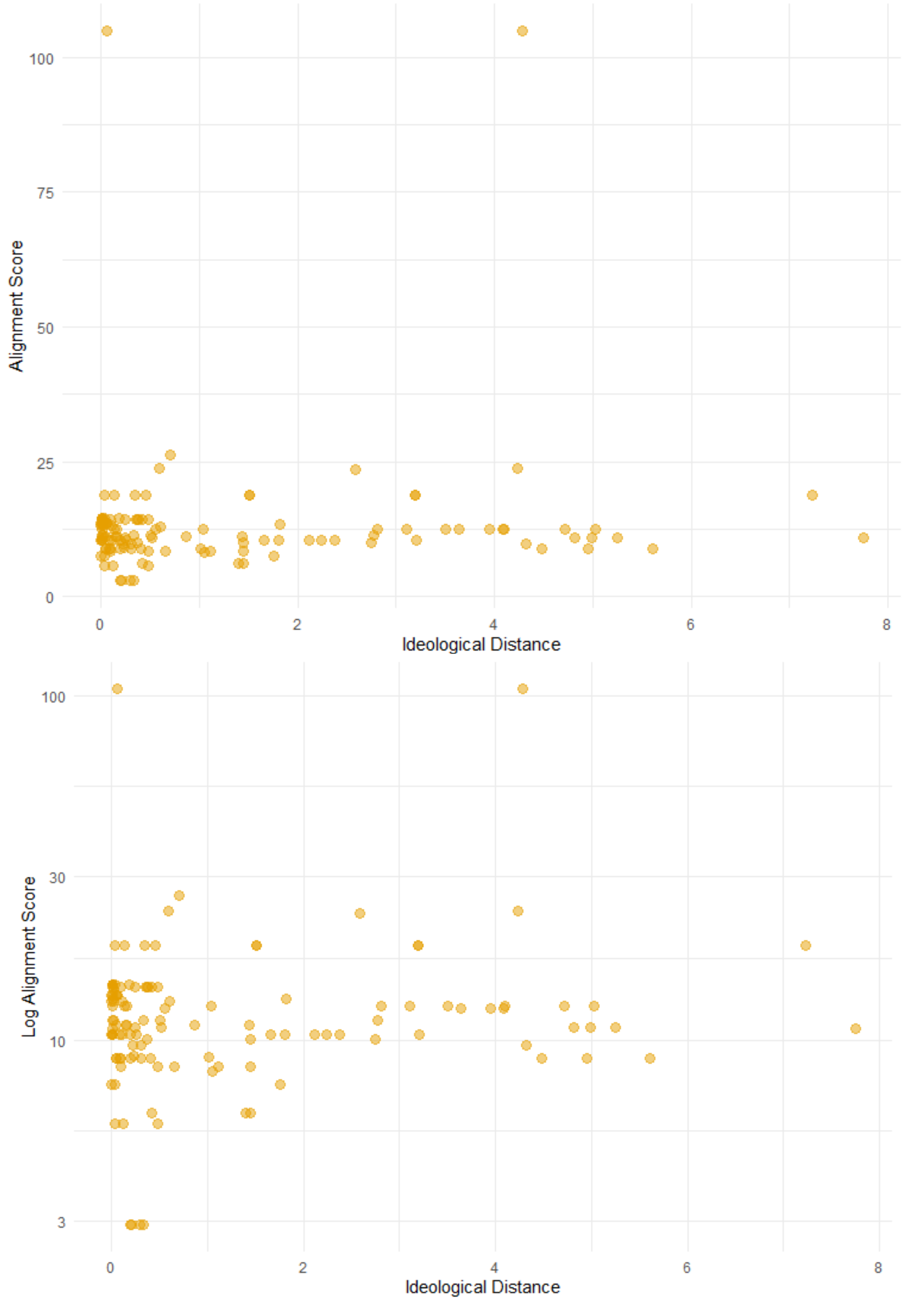


(a) Hexbin plot with alignments on natural scale. **(b)** Hexbin plot with alignments on log_{10} scale.

Study :Alignment Score-Ideology Distribution



Replication: Alignment Score-Ideology Distribution



Study: Alignment Score CDF

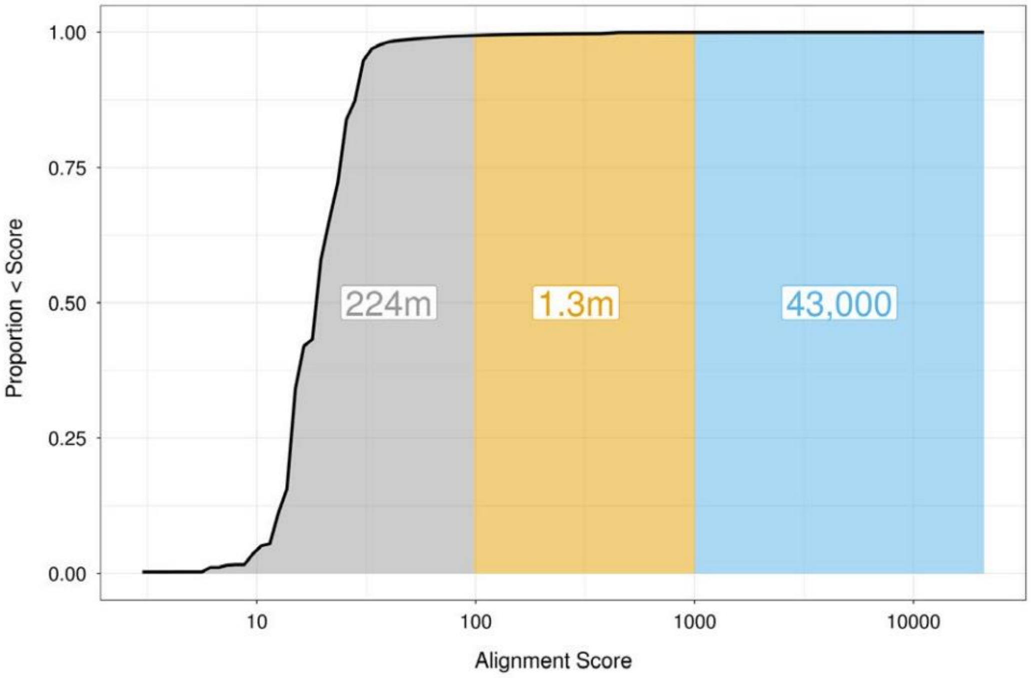
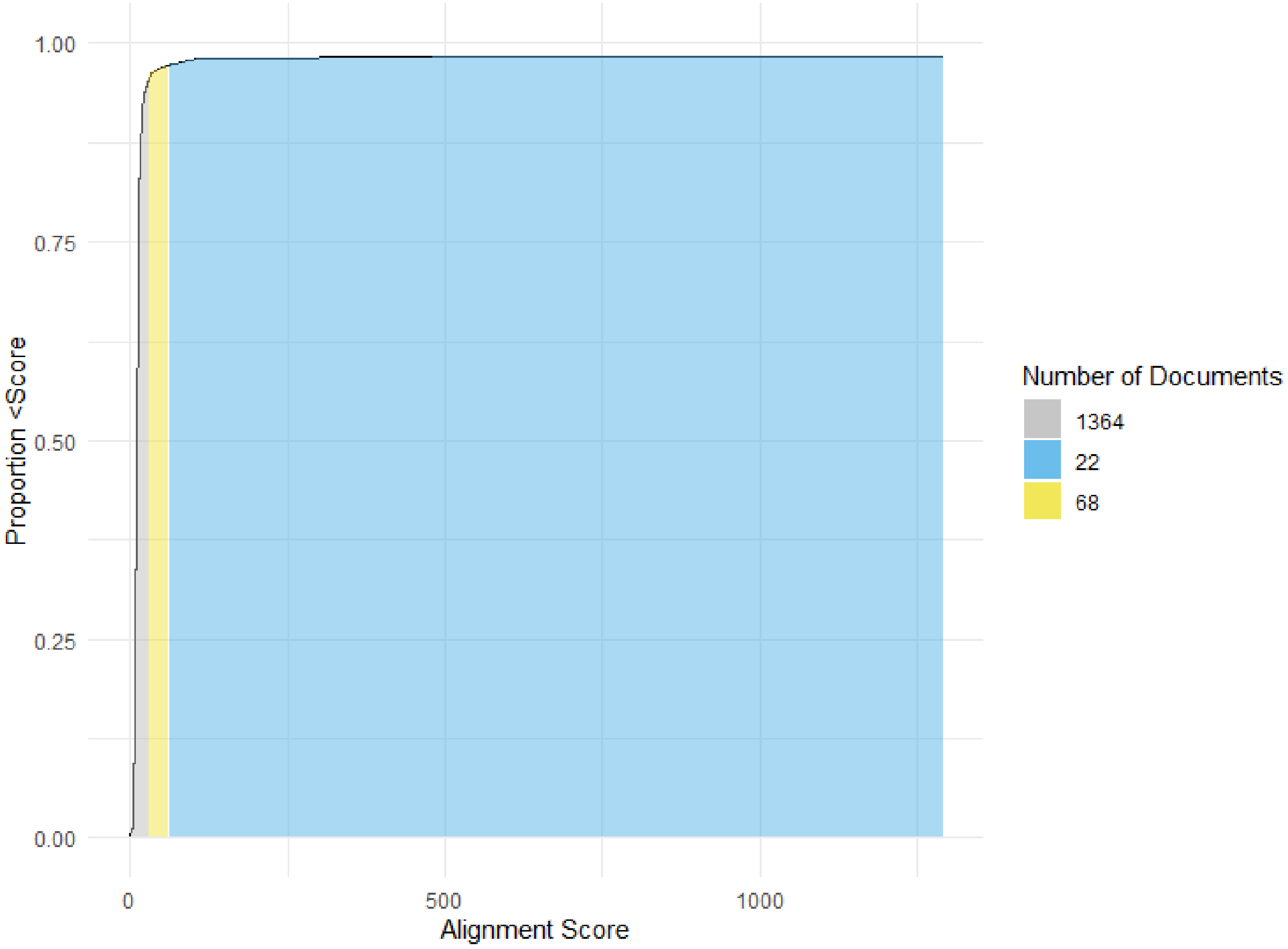


Figure 2. Cumulative Frequency Distribution of the Dyad-Level Alignment Scores.

Replication: Alignment Score CDF



Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) Diffusion Networks and Text Reuse
3) **Ideological Distance between bill sponsors and bill text reuse**

- **Expectation: Bill pairs with low ideology distances should have higher alignment scores**
- **Data:**
  - **Dependent Variable**: Alignment Score between pair of bills - available for replication
  - **Independent Variable**: Ideological Distance between pairs of bills
    - ❑ Raw bills data: Names of sponsors available, however: highly variable formatting, no unique identifier for a sponsor
    - ❑ Shor and McCarty Ideological scores: Available on archived repo. Name formatted consistently
    - ❑ Authors utilize Open States API to query each bill. Only keep those bills which return sponsor names through query.

      Study n = 88 million (40% of total number of alignments).
    - ❑ Query results not satisfactory. Replication instead uses Legiscan API + Manual Matching.
    - ❑ Replication N = 1480 alignments. 90th Quantile: 17.63
    - ❑ Keeping only those alignments with matched sponsors: N = 123

Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) Diffusion Networks and Text Reuse

3) **Ideological Distance between bill sponsors and bill text reuse**

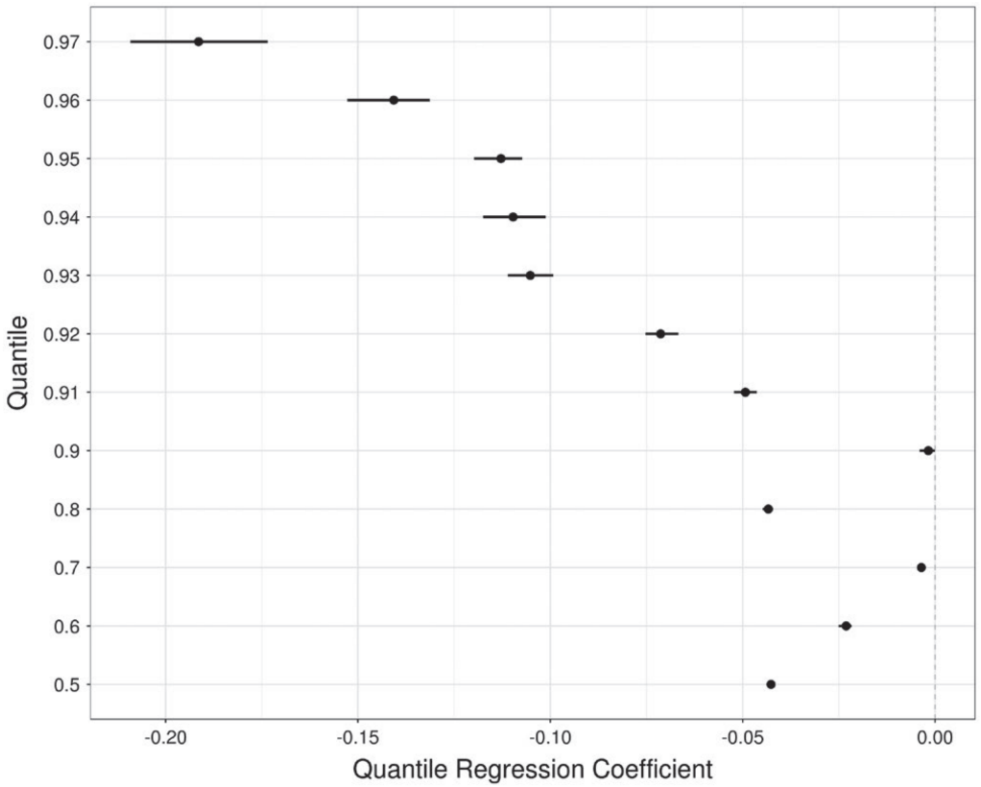
- **Expectation: Bill pairs with low ideology distances should have higher alignment scores**
- **Data:**
  - **Dependent Variable**: Alignment Score between pair of bills - available for replication
  - **Independent Variable**: Ideological Distance between pairs of bills
- **Methods**
  - ❑ **Quantile Regression.**
  - ❑ **Bootstrap to find confidence intervals**
  - ❑ **Not well suited for our sample. Bootstrapping will not create population representative estimates**

Robustness Checks

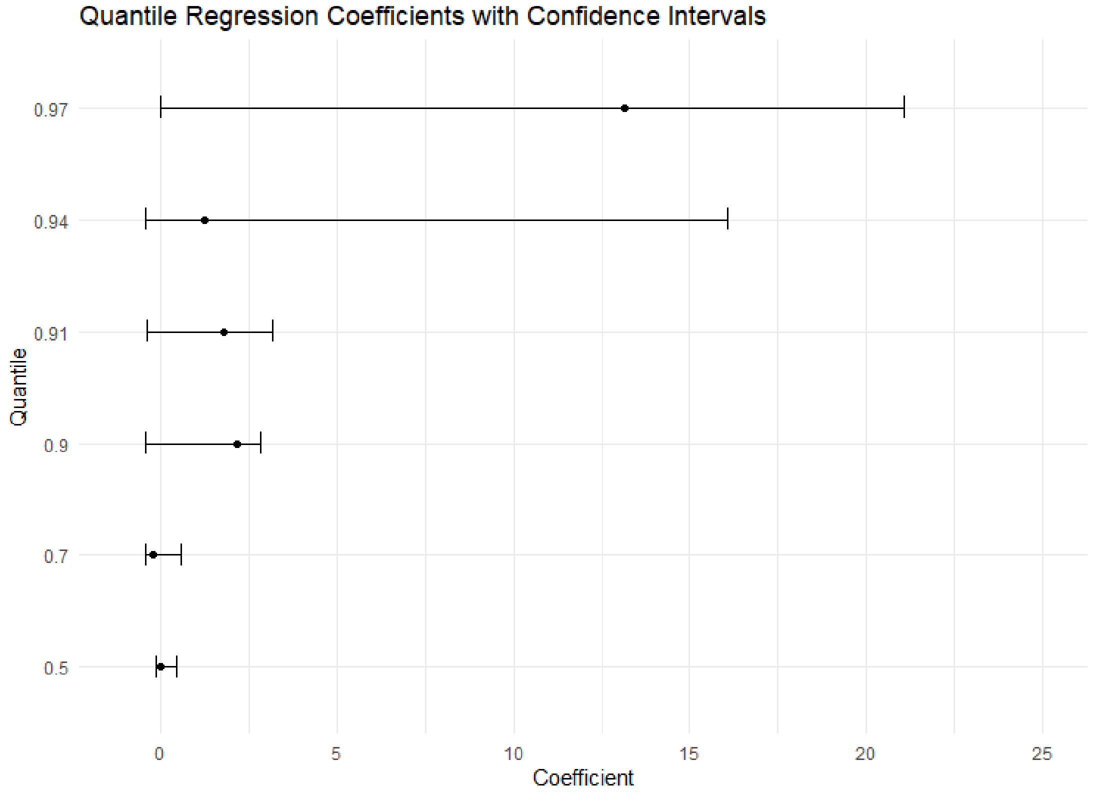
1) Comparing Alignment Scores with NCSL Data
2) Diffusion Networks and Text Reuse
3) **Ideological Distance between bill sponsors and bill text reuse**

- **Results**

Study Results



Replication Results



Robustness Checks

1) Comparing Alignment Scores with NCSL Data
2) Diffusion Networks and Text Reuse
3) Ideological Distance between bill sponsors and bill text reuse

Asymmetric Politics and Policy Similarity Across the U.S. States

Hypothesis: policies proposed by Republican state legislators are more similar to policies proposed by Republican legislators in other states than policies proposed by Democratic state legislators

Data:

1) **Alignment scores for bill pairs sponsored by Republicans and Democrats**

Replication Constraints

1) **To sample appropriately, ex ante information on bill sponsor ideology needed.**
2) **Even if possible through repeated sampling and testing, probability of matching bill with ideological distance, through sponsor name, very low. Other issues: Multiple Bill sponsors**

Autopsy

1. Data Availability
    - NCSL Data Missing
    - Raw Data somewhat different version
    - Supplementary Datasets such as Ideology scores require relatively more tedious searching in archived repos.
2. Resource Constraints:
    - Computationally intensive: Alignments Data consists of 570 million rows.
    - Difficult to truly sample from population. Reading files chunk-wise and sampling, till demanded sample n fulfilled
    - Multiple methods required libraries and deployments that have significant costs
3. Information Availability :
    - No Unique Identifier available for bill, i.e. Bill number not unique
    - Variable information on bill sponsor names and identifiers.
    - No clear indication by authors on what type of date information used. Date Created/Date Introduced/Date Signed
    - Outdated Methods: Significant changes in Open States API, Bing search API, NCSL website structure and information classification since 2018.
    - Two different variables for diffusion ties with unknown naming convention. Metadata doesn’t provide information.

Autopsy

High difficulty: Matching bills with Names.

- Multiple entries for same legislator name

Steps Taken:

1) Shor and McCarty provide ideological scores over different legislator sessions. Legislators flipping parties, become more or less conservative/democratic over time.
2) To match bill sponsors with names and states, need to make sure that if full name is available either in bill or through LegiScan, then unique match for name and state exists in ideology dataset
3) Find all legislators with duplicate names to check if two completely different legislators in the state have the same name, unique n = 150.

Manual check: Observe legislature membership periods over time, house and party. Use google and GPT 4.0 to research about each member.

Rules used to determine if duplicate names referred to same person

- If multiple names have membership in the same house, in the same party, in the same year
- If multiple names have membership in different house, in the same party, in semi-consecutive years
- If multiple names have membership in same house, but different parties, in semi-consecutive years
- If multiple names have membership in different house, different parties, in semi-consecutive years

No cases found where the above weren’t referring to the same person! Keep most recent information on legislator

Autopsy

High difficulty: Matching bills with Names.

- Multiple entries for same legislator name
- Matching Bill sponsor names with ideology dataset:

Open States returns only a single bill result when searched by bill number. However bill number is not unique like bill id. Querying using the entire bill title string lead to variable results.

Decided to use LegiScan

Autopsy

High difficulty: Matching bills with Names.

- Multiple entries for same legislator name
- Matching Bill sponsor names with ideology dataset:
- Open States: Not Viable.
- Decided to use LegiScan

Steps taken:

1) Use LegiScan
2) Get Banned by LegiScan
3) Use another email with LegiScan
4) Identifying best performing query method for LegiScan.
5) Utilize bill title search. Multiple results yielded
6) Select best result based on exact string match between title search and results (90% threshold).

Autopsy

High difficulty: Matching bills with Names.

Matching Bill sponsor names with ideology dataset:

Open States: Not Viable.

Decided to use LegiScan

Steps taken:

1) Use legiscan
2) Get Banned by LegiScan
3) Use another email with LegiScan
4) Identifying best performing query method for LegiScan.
5) Utilize bill title search. Multiple results yielded
6) Select best result based on exact string match between title search and results (90% threshold).
7) Realize sampling raw bills and finding sponsor names doesn't help. Have to sample first from Alignments and extract those bills from raw bill dataset, to find sponsor names!
8) Exhaust available API calls.
9) Use another email for new key. Found reliable matches for 30% of sample.

Autopsy

High difficulty: Matching bills with Names.

- Multiple entries for same legislator name
- Matching Bill sponsor names with ideology dataset:
- Open States: Not Viable.
- Decided to use LegiScan. 30% of Sample returned with full names.
- Expanding Matched sample through manual investigation:
    - ❑ Subset to bills in sample **with at least first and last names** and no LegiScan match.

        Eg: John Doe v/s J Doe, Bob Murr v/s Robert Murr

        - ❑ Match with ideology dataset on last name and state.
        - ❑ J. Doe matches with X doe, Y Doe, John Doe
        - ❑ Keep only row where J Doe Matches with John Doe, Bob Murr matches with Robert Murr. Drop Ambiguous cases.

    - ❑ Subset to bills in sample with **only last name** and no LegiScan match.
        - ❑ Match with ideology dataset on last name and state.
        - ❑ Keep only bills with unique match, i.e. unique last name amongst legislators in state.

    - ❑ Append LegiScan Matches with manually matched rows.
    - **Original Sampled n** = 2585 bills
    - **LegiScan matches** = 735 bills
    - **Total Bills with matched sponsors** = 1150

    - **Original Sampled Alignments** = 1480
    - **Matched Alignments** = 123

Autopsy

High difficulty: Matching bills with Names.

- Multiple entries for same legislator name
- Matching Bill sponsor names with ideology dataset:
- Open States: Not Viable.
- Decided to use LegiScan. 30% of Sample returned with full names.
- Expanding Matched sample through manual investigation:
  - ❑ Subset to bills in sample **with at least first and last names** and no LegiScan match.

    Eg: John Doe v/s J Doe, Bob Murr v/s Robert Murr
    - ❑ Match with ideology dataset on last name and state.
    - ❑ J. Doe matches with X doe, Y Doe, John Doe
    - ❑ Keep only row where J Doe Matches with John Doe, Bob Murr matches with Robert Murr. Drop Ambiguous cases.
  - ❑ Subset to bills in sample with **only last name** and no LegiScan match.
    - ❑ Match with ideology dataset on last name and state.
    - ❑ Keep only bills with unique match, i.e. unique last name amongst legislators in state.
  - ❑ Append LegiScan Matches with manually matched rows.
  - **Original Sampled n** = 2585 bills
  - **LegiScan matches** = 735 bills
  - **Total Bills with matched sponsors** = 1150

  - **Original Sampled Alignments** = 1480
  - **Matched Alignments** = 123

Extension

- Utilize methods like BERT based topic modelling to find what policy areas are more likely to have high/low alignment. Variation across party dyads
- Use LLMs to find similarity between texts. Might be better at finding thematic similarities that may not be represented in direct text alignments.
- Utilize methodology in different parliamentary systems. For eg. Indian parliament, where most laws - not directly related to how state fund is spent - are decided at federal level through voting by state legislators. Therefore, state level law making procedures are subject to dynamic of whether Federal ruling party is the same as a state's ruling party.

Is there greater alignment in policies amongst states, when states' ruling party is the same as federal ruling party ?