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

Ayush Lahiri

Introduction

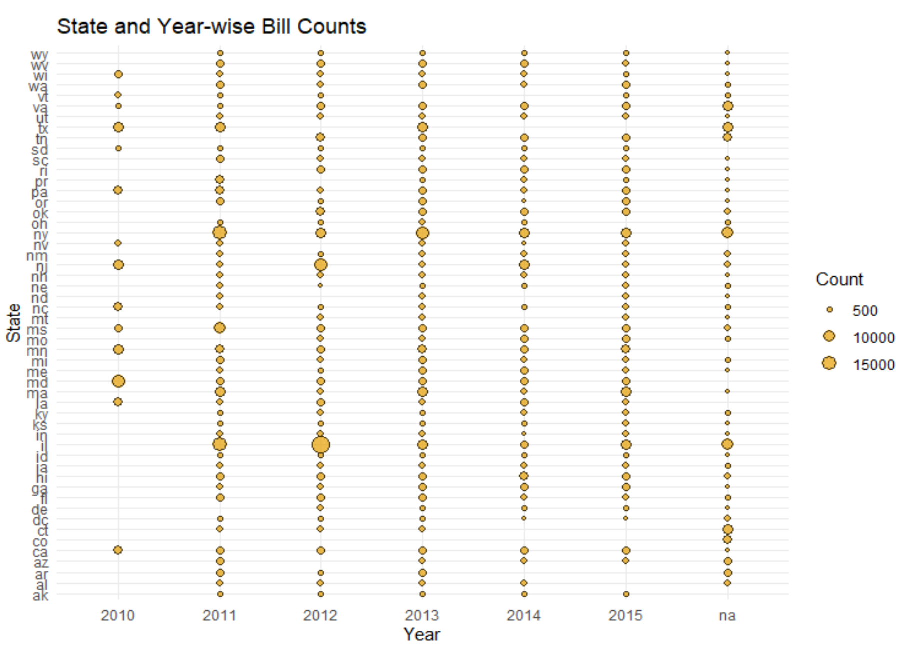
The 2018 study by Fridolin Linder, Bruce Desmarais, Matthew Burgess, and Eugenia Giraudy, investigates policy similarity amongst US states through bill text reuse. They measure, the latent variable of policy similarity effectively through text reuse. The authors show that text in legislation is an aggregate representation of the dimensions underlying the policies proposed therein. These dimensions include the domain of the policy, the ideological position enacted by the policy, the level of specificity in the policy enactment, and several other salient features of policy that are communicated through the text in legislation. By generating continuous numeric, text alignment scores between pairs of bills the authors approach the question of policy similarity as a continuum rather than a dichotomy. Once text alignment scores are achieved, multiple validation tests are undertaken to assess whether text-reuse–based measure of policy similarity correlates with various benchmarks drawn from theory and prior research.

Overview of replication

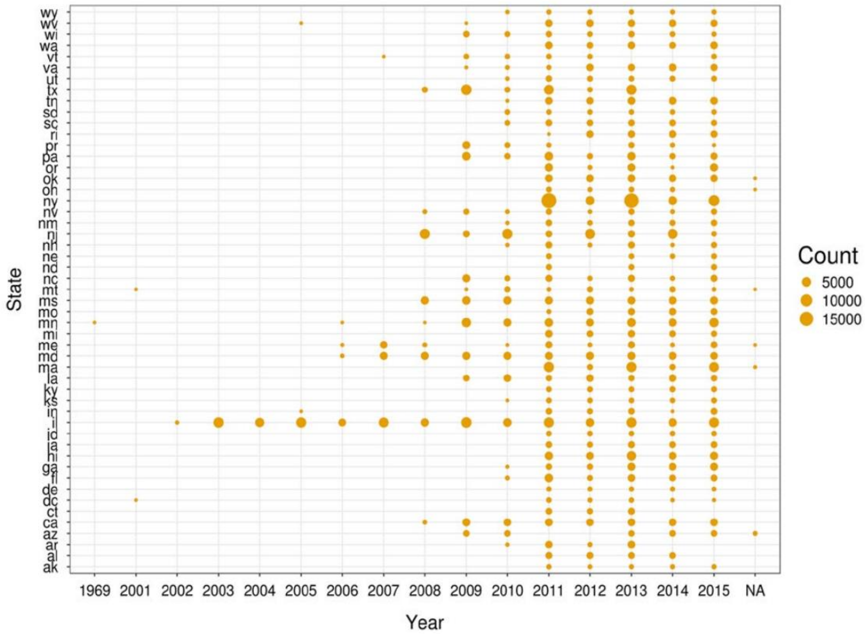
I consider the paper across 3 separate methodological sections. 1) Creating the data pipeline from the raw bills data to finding the final text alignments between pairs of bills and scores, 2) robustness checks: a) predicting co-placement of bills in NCSL tables b) testing significance of relationship between state diffusion networks and text alignments c) testing significance of difference in ideology scores of sponsors of a bill pair and text alignments 3) investigating distribution of alignment scores for republican sponsored bill pair versus democrat sponsored bill pairs. I am only able to replicate robustness check b), completely and to a partial extent robustness check c), amongst all three sections and subsections. The reasons for the same are discussed in the autopsy section. Below I discuss, differences in results amongst the replications I was able to undertake.

Differences in Replication Results:

1) We first, begin by assessing the raw bills data as made available by the authors on Harvard Dataverse and compare our results for the counts bills available by year and state to their results.



Replication Results: Number of Bills by state and year



Study Results: Number of Bills by state and year

Firstly, I find that the data, the authors refer to as their raw bills data and what is available on their publicly available repo have some important differences. They show bill data availability from 1969 but only make data available, 2008 onwards. Unavailability of data up until 2008 can be explained by the authors analysis focused on bills only after 2008 and hence may have chosen to make only that subset available. However, in the available data, I find the earliest bill with non-missing values for full bill text is only from 2010 onwards along with slight differences in state and year wise distributions. This highlights an important discrepancy between the raw data used by the author's and what is made available. While the counts after 2010 do not differ significantly based on visual inspection, it is additionally unclear what the authors interpret as NA values in their plot. I choose to interpret NA values as those bills without any bill text information. This distribution of missing bill text also differs significantly (assuming congruency between mine and the author's interpretation).

2) Given the text alignment scores data made available by the authors and the dataset of ties in state diffusion networks made available by Desmarais et al (2015), I am able to undertake the replication procedure, to assess the robustness of the alignment scores. This is done by replicating the assessment of whether existence of a tie in diffusion networks between states, predicts the average alignment scores between states. Following the study's procedure, I regress the state dyad matrices for whether a diffusion exists or not, the policy coverage between states (outer product of the number of bills introduced in the state across all available years, with itself) on the alignment scores between state pairs and again on the logarithmic transformation of the alignment scores. Coefficients are calculated with OLS regression and normalized with standard deviation of cross state alignment scores. P-values are based on 1,000 QAP permutations, instead of 5000 permutations as undertaken by the authors due to computational resource constraints. Comparison of Results are shown below:

	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

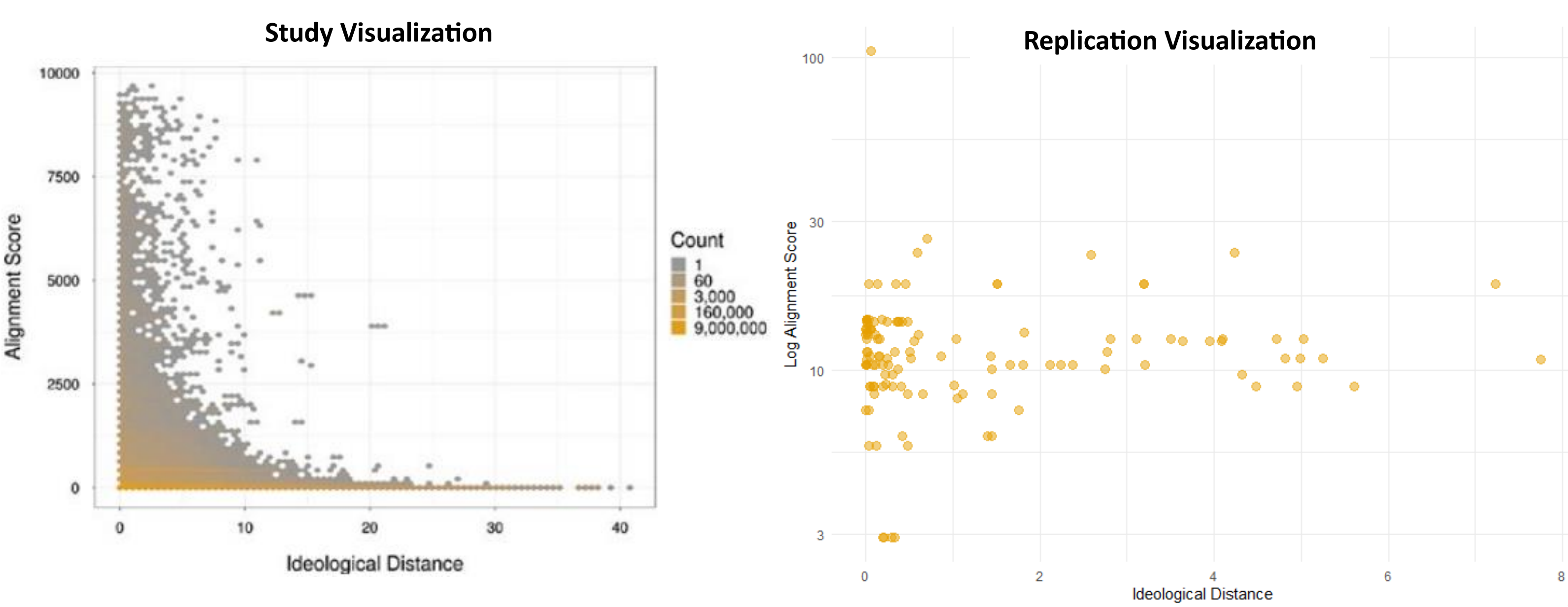
	Identity Link		Log Link	
Intercept	–2.883	0.002	7.776	0.000
Diffusion tie	0.441	0.005	0.381	0.006
Coverage	0.951	0.001	1.106	0.001

Table 2: Study Results

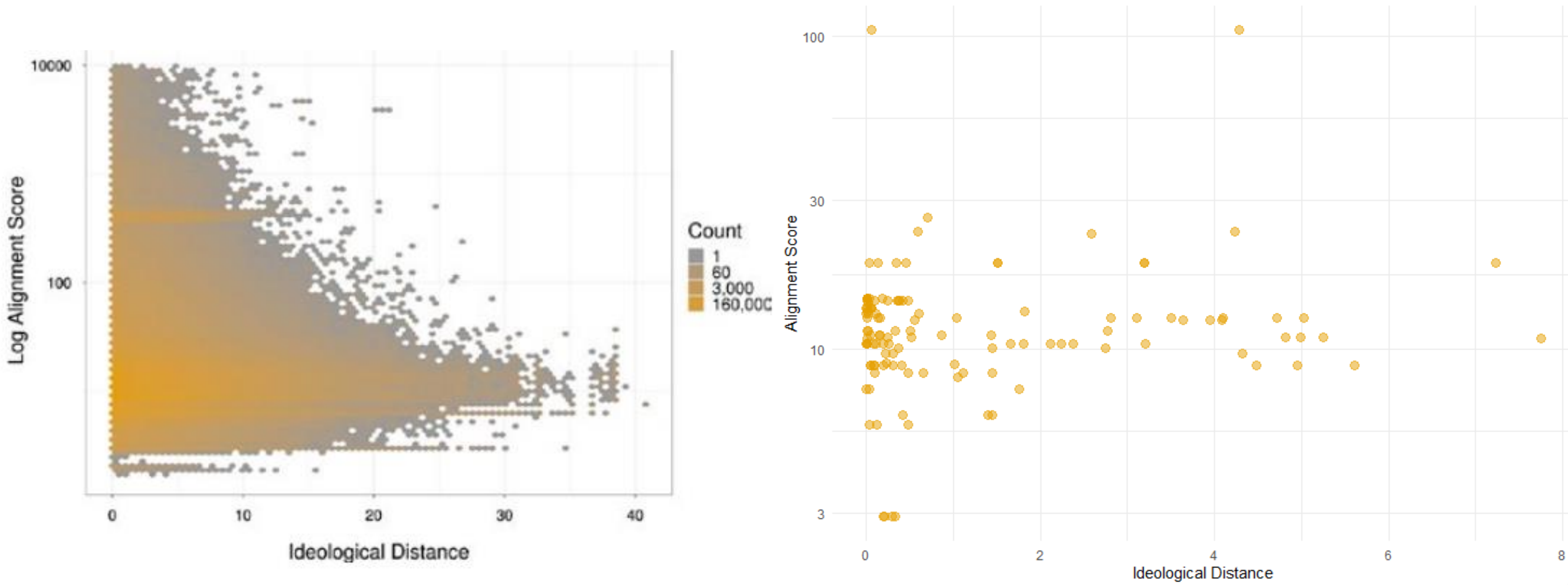
I find that for the data available, the direction of coefficients and the significance of the primary variable of interest which is, whether a tie exists between states in the diffusion network, aligns with the study's results with some differences. Specifically, I find significant evidence that the presence of a diffusion tie is strongly correlated with higher alignment scores, aligning with the study's results. I find much greater congruence in coefficient strength, direction and significance for both diffusion ties and coverage, across mine and the study's results when the log of alignment scores is used as the dependent variable.

In the identity link regression however, larger differences are observed in the replication results for coefficient strength for the intercept and policy coverage. This maybe due to differences in the version of data that the authors make publicly available, relative to what they use (as discussed in the first replication results). The difference in significance however can most likely be attributed to the replication utilizing 1000 permutations, due to computational limitations, instead of 5000, as the authors do in the study. Further the difference in random state can also affect permutation results.

3) Next, I replicated the robustness check of whether differences between ideology scores of the sponsors of a bill pair are predictors of the alignment score of the bill pair. Following the author's methodology, I use the ideology score estimates (ideal points) as made available by Short and McCarty (2011) . However, matching the two datasets proved to be a difficult endeavor. Matching required full names of a bill's sponsor, their party and state, which was not available in the raw bills dataset. The author's method of finding names of sponsors through open states API was also not replicable. By sampling bills from text alignments and finding names of sponsors of sampled bills through alternative APIs and manual matching, I was able to match 123 alignments with the ideology scores dataset. The small size of the dataset compared to the author's sample of 88 million alignments, led my replication to arrive at statistically weak results. The differences I show below are primarily attributable to my highly unrepresentative sample.

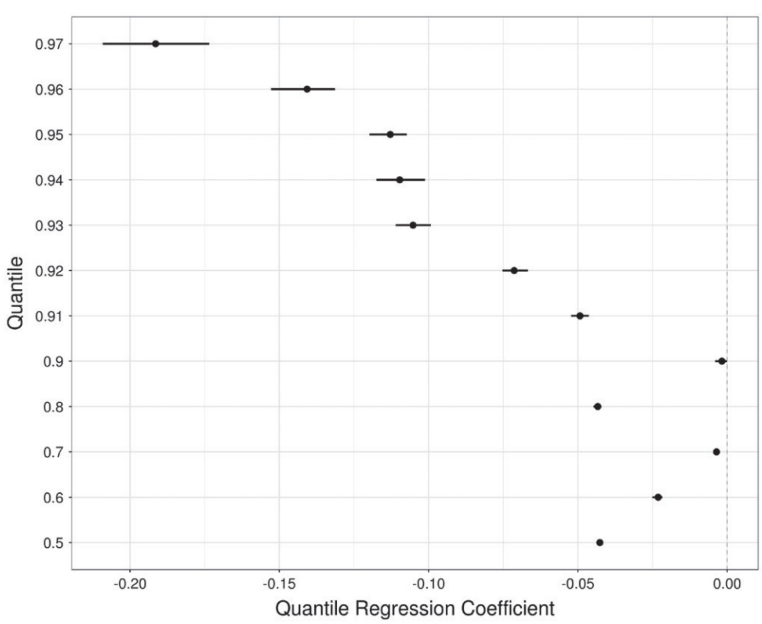


Distribution of Alignment Scores and Ideology Scores

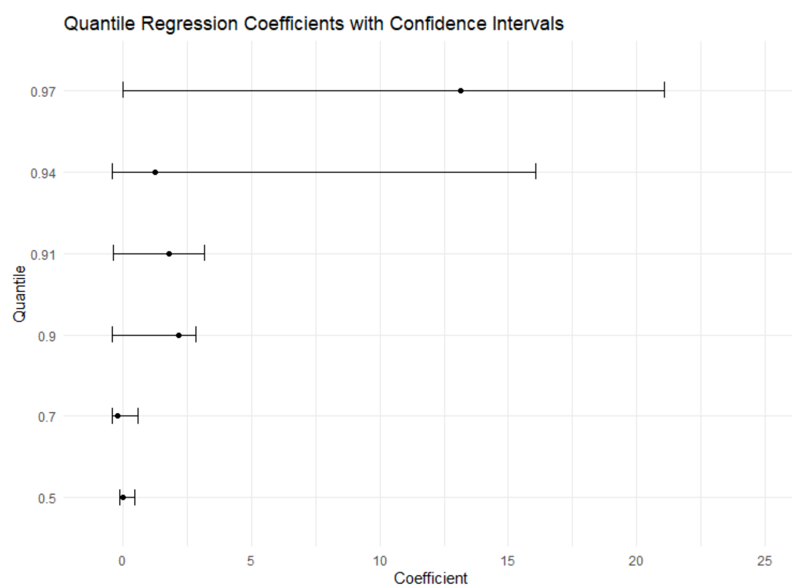


Distribution of Log Alignment Scores and Ideology Scores

Study Visualization



Replication Visualization



Quantile regression of Ideology Score on Alignment Score result

I find that my sample distribution is expected given the distribution of all documents. Most documents are concentrated in the 0 to 100 score range (log). However, this leads to my sample not being representative of the outlier cases of low ideological distances where there are high text alignment scores. Subsequently, my quantile regression results, regressing ideology difference on alignment score, are statistically insignificant with coefficients' direction being opposite of what the authors found. This is expected given my sample, doesn't truly have quantile wise distribution and statistically, the entire pipeline of my sampled replication is not valid.

Autopsy

1) I was unable to replicate the pipeline from the raw data to the final bill pairs with boilerplate adjusted alignment scores. This was due to resource constraints in the implementation of ElasticSearch. The most recently available implementation of ElasticSearch in python and R, requires access to the ElasticCloud API service which is a paid service and also requires separate cloud deployment; also a paid resource. Unable to find the 500 closest bills to

each bill using elastic search I was thus unable to implement the SW algorithm, since it is a computationally expensive algorithm to run on all 500,000 bills.

2) I was also unable to replicate the first robustness check of assessing the accuracy of text alignment scores in predicting co-placement of bills in the NCSL database. This is due to a combination of resource constraints and major changes in the source, i.e. NCSL website, being scraped. The NCSL database has undergone significant change since the study was published in 2018 and no longer reports similar bills in separate tables. NCSL now classifies its bills by policy topic, making it unfeasible to scrape tables that provide us information on bills' policy similarity regardless of policy topic. Secondly, scraping NCSL, requires access to the Bing search API within a Microsoft Azure cloud unit both of which are paid services, like in the case of ElasticNet.

3) For the robustness check on ideology score, the authors method of using open states API, to reverse search bills and find sponsors and their political affiliation was also not replicable due to changes in open states API structure since the study was published. I undertook a mix of stratified sampling from text alignment scores, finding bills in the raw data as per the alignments found in the sample. I then utilize the LegiScan API instead of Open States to reverse search bills by title to find full sponsor names. I then supplement still unmatched bills with manual matching to create a sample of alignment scores with sponsor name and party information. This sample was then matched with Shor and McCarty's dataset to arrive at 123 alignment pairs with all requisite information. This required significant experimentation with the LegiScan API structure and further significant effort in manual matching.

4) Finally, I was unable to undertake the analysis of distribution of alignment scores for same party dyads since this required me to sample bills from the text alignment dataset with ex ante knowledge of sponsor ideology. Given the significantly large size of the dataset - sampling and querying on LegiScan was not feasible due to computational resources and API call limits, especially due to the sparse results for a bill as was found in replicating the ideology scores.

Extension

1. The use of Bidirectional Encoder Representations from Transformers (BERT) for topic modeling presents a more modern approach to find policy topics of bills on a more detailed level, such as income tax vs corporate tax. Through such thematically accurate topics, an area of interest for me would be to observe how text alignment across same or different party dyads are changing over time. For e.g. aligned bills have higher scores when the alignment is between democrats on the topic of healthcare, vs republican bill dyads showing higher scores on taxes. Such ideology and political dis(agreement) can be of key interest in understanding more nuanced phenomena extracted from text such as party line towing on specific policy topics which may be influenced by popular policy topic of the time.

2. LLMs, such as GPT, can be utilized to assess the similarity between policy texts, offering a more nuanced approach to understanding thematic alignments that might not be apparent through more direct text comparisons as undertaken in this paper. The underlying transformer architecture is highly successful in finding and remembering context in large pieces of text, which can allow for similarity to be assessed on a more nuanced level, even when text matches may lead to lo alignment scores.

3. The methodology can also be extended to examine policy alignment within different parliamentary systems, such as the Indian Parliament. In India, the dynamic between federal and state legislation is particularly significant, given the limited policy purview at the state level, which is substantially influenced by federal grants. This analysis aims to explore whether state-level public initiatives, which serve as a common method for implementing state policies while navigating the legal hierarchical constraints on policy areas, align with federal-level bills on similar topics. By assessing the correspondence between state initiatives and federal politics, it would be interesting to assess which state level parties 'unofficially' align with the ruling party in the federal government, due to high diversity in state level parties dominated by local legacy political groups, unlike at the federal level dominated by two historic national level parties. Subsequently, this would allow an assessment of how national level parties may be consolidating power in a more nuanced manner.