Identifying Equivalent State Bills through Text Reuse.

Subtitle.

11/02/2015

Abstract

Much research has been focused on the diffusion of policy ideas in US state legislatures. Most of this research uses hand coded data sets that identify equivalent bills and analyze patterns of adoption of these bills. Bills on equivalent policies often contain the same language, since legislators use past legislation from other states or model legislation from interest groups as templates when drafting new bills or model legislation from interest groups. In this paper we evaluate the effectives of text reuse measures to detect bills that address the same policy issues. We find that...

1 Introduction

1.1 Tasks

-

Research on public policy adoption and diffusion has conventionally relied on modest hand-coded datasets encode when one or a handful of policies were adopted by each jurisdiction (Boehmke and Skinner, 2012). Scholars of public policy diffusion have recently turned their attention to the empirical identification of emulation ties connecting policy jurisdictions (Volden, 2006; Boehmke, 2009; DESMARAIS et al., 2015; Garrett and Jansa, 2015). Identifying instances of policy emulation and explicit diffusion ties opens policy diffusion scholarship to an entirely new set of questions and theories that can be approached empirically. In this paper we draw upon a massive and nearly all-encompassing source of data from which to infer diffusion ties – the text of legislation. If the computer-based analysis of bill text can be tuned to successfully identify policy adoption and emulation, we will be able to vastly scale up both the scale of data gathered and the precision of diffusion inferences.

How ever, it is not clear, how well text reuse is suited to to detect real policy diffusion. There are several complications that we are addressing in this paper. First, every bill has procedural content that is not related to the policy content of the bill. Since every

state legislature has a set of such standard or boiler plate text in each bill, there will be significant text reuse between bills in the same state and possibly also between bills from different states. Second, it is not obvious how much text reuse will mean substantive policy overlap. Given that each bill pair has a continuous proportion of overlapping text, setting the threshold too low might mean to classify bills as equivalent that are only in the same policy area, or are on a similar issue, but are opposed in content. On the other hand, setting the threshold too high could mean overlooking equivalent bills because of small insignificant changes to the text.

In this paper we address these issues and evaluate how much policy overlap can be detected using text reuse measures. Following Wilkerson et al. (2015) we use supervised machine learning to separate boiler plate from substantive text. We further more use an original data set on policy diffusion to evaluate how much equivalent policy can be detected using measures of text reuse. Using the system developed by Burges et al. (2015) we calculate text reuse scores for all pairs of bills in our dataset. We then use these scores in a model that classifies bills as equivalent and evaluate its performance with the validation data set.

This work has several important implications. First, it allows us to estimate, how policies are transferred between states. Do state legislators work mainly from templates from other states or interest groups, or to what extent do legislators draft their own bill text. Furthermore, text reuse is a relatively simple metric to calculate for large amounts of text. Previous scholars of policy diffusion mainly relied on case studies or manually coded data sets of policy diffusion in few policy areas. If working copying text forms a significant portion of how legislators adopt policies from other states, text reuse can be used to easily gather comprehensive data sets on policy diffusion.

[This might be overlapping a bit with your part]

2 Background

Public policy diffusion – the process by which policymakers emulate the policies implemented outside of their jurisdictions – is a firmly established area of research in both Comparative (Simmons and Elkins, 2004; Gilardi et al., 2009) and American politics (Walker, 1969; Berry and Berry, 1990; Shipan and Volden, 2006; Nicholson-Crotty, 2009). Until recently, policy diffusion studies have deployed quantitative research designs in which the relational component of policy diffusion (i.e., which jurisdiction is being emulated in a given diffusion instance), has been treated as completely unobservable or assumed to align with the geographic adjacency network (i.e., jurisdictions only emulate their geographic neighbors) (Volden, 2006; Boehmke, 2009). Recognizing the limitations in this approach, scholars have recently taken on the task of directly measuring the latent networks through which policies diffuse. DESMARAIS et al. (2015) apply network inference algorithms to US state adoption sequences in over 100 policy domains to empirically infer the underlying network through which policies diffuse. Garrett and Jansa (2015) analyze the text of US state legislation to measure the diffusion of policy in one domain – restrictions on insur-

ance coverage for abortion – and identify the influence of model legislation, as introduced by interest groups.

In order to detect text reuse between bills we follow Wilkerson et al. (2015) and use the Smith-Waterman local alignment algorithm (Smith and Waterman, 1981).

2.1 Tasks

- Policy diffusion (FL expand paragraph)
- Text re-use in CBP in bills
- Text re-use in general
  - Plaigirism
  - Alignment (FL expand review of Wilkerson et al)
- Discussion of text-based classification

3 Data

We will rely on two main data sources to assess the reliability of text reuse to identify substantively equivalent bills.

In order to calculate the alignment scores between bills, we rely on a database collected by Burgess et al. (2015) and the Sunlight foundation. This data base contains approximately 500,000 bills from 2008 to 2015. The available metadata for the bills includes a timestamp for introduction and approval of the bill, the name and party affiliation of the sponsor(s), the state and the bill id.

This collection of bills is based on all bills that are available through the `openstates.org` API. Open states is a website maintained by the Sunlight Foundation, in order to increase transparency in state politics. The Sunlight Foundation use web scrapers to access all bills that are available on the websites of legislatures in all US states. This includes enacted legislation as well as bill that is still in the legislative process, or where not enacted. The database contains the full bill text and the following additional metadata:

- action_dates:
  - first
  - last
  - passed lower
  - passed upper
  - signed

- actions (a list of legislative actions. E.g. referred to a committee)
  - actor
  - date
  - related_entities
  - type
- bill document first
- bill document last
- bill id
- bill title
- date created
- date introduced
- date signed
- date updated
- session
- short title
- state
- summary
- sunlight_id
- unique id

For the ideological matching, we rely on latent ideology scores measured by (SHOR and McCARTY, 2011). The data set contains scores for 20738 legislators from 50 state legislatures as well as their party affiliation and their time in office.

We additionally construct a validation dataset of equivalent bills from information obtained from the National Conference of State Legislatures (NCSL). The NCSL publishes summary tables on specific policies, citing the relevant bills or sections in the state statutes of the states that have implemented regulations on this policy. We collected all these tables, and extracted all bills that address the same policy measure and that overlap with the time frame covered by our bill database.

3.1 Tasks

- Scrape Google urls for NCSL tables (FL) – lower priority
    1. Scraped 64 urls.
    2. Can't get around the Google API limitation yet. Also get blocked when trying to scrape the search result.
- Extract state & bill # from tables (FL) – lower priority
    1. Didn't do this yet, let's first check how many are potentially suitable and then probably better to do by hand
- See tasks in analysis – create metadata file and dyadic file, then put them on the ACI (FL).
    1. We have a preliminary dyadic file for all similar bill pairs and their alignment scores (this is from the LID approach: each bill is only compared to the 100 most similar ones)
    2. Matt transfering the database to a new server at the moment, it is therefore unavailable. Should be online again sometime today.
- Store MALP data on ACI and make sure the identifiers match those in the bill data (FL).
    1. data is on aci
    2. Legislators can be matched by Name, State, Pary and Term. This information should be contained in the database. If not, the open state api has a legislator search method, that returns all necessary information.
- Put policy diffusion data in project folder on ACI (BD).

4 Analysis

Below we list X separate analyses designed to test the degree to which measuring text reuse measures policy overlap/diffusion.

- We use statistical topic modeling to assess the major content areas represented by the text identified in the alignment algorithm. We consider whether the resultant topics align with policy areas.
- The Measuring American Legislators Project (MALP) provides data that we will use to assess the significance of text re-use in US state legislation SHOR and McCARTY (2011). The most recent release of the MALP data covers 1993-2014. The MALP

data provide ideological scores of legislators on an annual basis. We conduct a bill-level statistical network analysis to see whether the rate of bill-to-bill alignment is positively related to the ideological similarity of their sponsors.

- We conduct a state-level analysis to see whether the volume of state-to-state text alignment is positively related to the policy diffusion networks inferred in DESMARAIS et al. (2015).

4.1 Diffusion networks and Text re-use

To evaluate whether text re-use corresponds to the transfer of policy, we test whether the presence of a diffusion network tie between two states is a predictor of text reuse. We use the policy diffusion networks inferred in DESMARAIS et al. (2015). We use [FL FILL IN] to measure the incidence of dyadic text alignment between all state bills covering the time period 200–[??]. For each bill, we first gather the 100 closest bills based on cosine similarity in the term vectors characterizing the bills. We then represent a dyad of bills by the text alignments between them. For each state-pair we calculate the number of alignments between bills from each state. The median number of alignments between states is 3,491 with a mean of 6,309 and standard deviation of 8,809. The "Diffusion Ties" variable measures the number of diffusion edges between states in the 2008 diffusion network, as measured by DESMARAIS et al. (2015). The number of diffusion edges between two states is either 0, 1, or 2. The diffusion network in 2008 is inferred using policy adoptions in the 35 years preceding (and excluding) 2008. As such, we do not risk double-counting bills in the policy adoption and text re-use data. There is one observation in the analysis for each of the 1,225 unique state-pairs. Since this is dyadic data, we use a matrix permutation method, quadratic assignment procedure, to calculate p-values (Krackhardt, 1988). As a robustness check, we run the model with both the identity and log link.[1]

	Identity Link		Log Link	
	Coefficient	p-value	Coefficient	p-value
Intercept	5717.18	0.0000	8.0431	0.0000
Diffusion Ties	2417.00	0.0308	0.3288	0.0452

Table 1: Predicting number of alignments in legislation across states with diffusion ties. Coefficients calculated with OLS regression. p-values based on 5,000 QAP permuations.

Results of the simple dyadic regression are presented in Table 1. In both specifications there is a positive relationship between the number of diffusion ties and the number of alignments, and the relationship is statistically significant at the 0.05 level (two-tailed). Furthermore, the magnitudes of the relationships are substantively significant. A shift from the minimum to the maximum number of diffusion ties corresponds to more than a half of a standard deviation increase in the expected number of cross-state alignments. On

[1]The p-values were calculated using 5,000 random matrix permutations.

the log scale, the addition of a diffusion tie corresponds to a 40% increase in the expected number of cross-state alignments.[2].

4.2 Tasks

- Develop bill metadata dataset (FL)
  - unique bill identifier
  - sponsor identifier from MALP data
  - state identifier
  - chamber identifier
  - Introduction date
  - Anything else, even if some is missing (status, committees, etc)
- Develop bill-to-bill edgelists (FL)
  - Sparse dyadic dataset (i.e., no observations when there is 0 alignment/overlap)
  - alignment score(s)
  - name of a text file in which aligned text is stored
- Topic models, possibly of a relational form, applied to aligned text
- Computationally intensive analysis of bill-to-bill alignment and sponsor ideology.
  - Get big sample of bills
  - Assess relationship between alignment between bills and ideological similarity between sponsors
- Analysis of alignment aggregated at the state level.
  - How do we score overlap at the bill and state level?

References

Berry, F. S. and W. D. Berry (1990). State lottery adoptions as policy innovations: An event history analysis. *American Political Science Review 84*(2), 395–415.

Boehmke, F. J. (2009). Policy emulation or policy convergence? potential ambiguities in the dyadic event history approach to state policy emulation. *Journal of Politics 71*(3), 1125–1140.

[2]Calculated as $100 \times [\exp(0.3288) - 1] = 38.93$

Boehmke, F. J. and P. Skinner (2012). State policy innovativeness revisited. *State Politics & Policy Quarterly 12*(3), 303–329.

DESMARAIS, B. A., J. J. HARDEN, and F. J. BOEHMKE (2015, 5). Persistent policy pathways: Inferring diffusion networks in the american states. *American Political Science Review 109*, 392–406.

Garrett, K. N. and J. M. Jansa (2015). Interest group influence in policy diffusion networks. *State Politics & Policy Quarterly*, 1532440015592776.

Gilardi, F., K. Füglister, and S. Luyet (2009). Learning from others: The diffusion of hospital financing reforms in oecd countries. *Comparative Political Studies 42*(4), 549–573.

Krackhardt, D. (1988). Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social networks 10*(4), 359–381.

Nicholson-Crotty, S. (2009). The politics of diffusion: Public policy in the american states. *Journal of Politics 71*(1), 192–205.

Shipan, C. R. and C. Volden (2006). Bottom-up federalism: The diffusion of antismoking policies from u.s. cities to states. *American Journal of Political Science 50*(4), 825–843.

SHOR, B. and N. McCARTY (2011, 8). The ideological mapping of american legislatures. *American Political Science Review 105*, 530–551.

Simmons, B. A. and Z. Elkins (2004). The globalization of liberalization: Policy diffusion in the international political economy. *American political science review 98*(01), 171–189.

Smith, T. F. and M. S. Waterman (1981). Identification of common molecular subsequences. *Journal of molecular biology 147*(1), 195–197.

Volden, C. (2006). States as policy laboratories: Emulating success in the children's health insurance program. *American Journal of Political Science 50*(2), 294–312.

Walker, J. L. (1969). The diffusion of innovations among the american states. *American Political Science Review 63*(3), 880–899.

Wilkerson, J., D. Smith, and N. Stramp (2015). Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science*.