Identifying Equivalent State Bills through Text Reuse.

Subtitle.

Bruce Desmarais, Matt Burgess, Fridolin Linder, Eugenia Giraudi

12/08/2015

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~~will then assess the scientific value of the text reuse scores in several different ways. First, we use topic modeling on the discovered reused text sequences to assess the content of the matching text. Second, we investigate the relationship between the ideological distance of the legislators that proposed two bills and the text reuse scores between these bills. Third, we assess the how well policy diffusion networks that have been established by previous research can be discovered using the amount of text reuse between states. And fourth, we use data on equivalent bills collected by the National Council of State Legislatures to build an evaluation data set containing true policy overlap to assess the accuracy of text reuse in predicting policy overlap.

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 insurance 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~~ Previous research on policy diffusion relies almost exclusively on small subsamples of legislation and hand coded data. Little work has been invested in investigating automated ways of detecting equivalent bills in order to measure policy diffusion. Wilkerson et al. (2015) use the Smith-Waterman local alignment algorithm ~~(Smith and Waterman, 1981)~~(SW algorithm) (Smith and Waterman, 1981) to detect overlapping language in congressional bills in order to trace policy ideas through the legislative process of the US congress.

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 Detecting Text Reuse

In this study we use the same algorithm as Wilkerson et al. (2015) to detect matching sequences of text in state bills. We also follow their procedure in splitting the bills up in sections as the unit of analysis. However, our procedure differs in the pre-screening of sections that are compared with the Smith-Waterman algorithm and in the procedure to identify boiler plate text.

Wilkerson et al. (2015)'s criterion for pre-screening bills is that they have to match at least 5 10-grams. This means that there has to be a minimum of a 15 character exact match between two bill sections in order to consider them for more in depth analysis with the SW algorithm. We see this as a too restrictive criterion and choose a different approach instead. Our text data is stored in an Elastic Search database. Elastic search is an open source web search engine which is designed to find document that match a search query. We utilize the 'more like this' search engine, which is designed to find documents that are similar to a given document [FILL IN CITATION]. In principle it works in the following way: From the focus document, the k n-grams with the highest tf-idf scores[1] are selected and transformed into a search vector. Then, for each document in the collection, a variation of the cosine similarity[2] is calculated and the n with the highest scores are selected. This procedure is necessary, since the SW algorithm is computationally demanding and an exhaustive comparison is not feasible (we have about $500,000$ bills, assuming each bill consists on average of four sections this would amount to about $2 * 10^{12}$ comparisons). [FILL IN PARAMETER VALUES]

Furthermore, we do not rely on human coders to extract and classify boiler plate text. We identified two types of boiler plate text: 1) Procedural language like bill and section headers, common legislative phrases etc. 2) Topic specific boiler plate, like definitions of tax brackets, [another example], etc. In order to detect these kinds of boiler plate language, we pre-screen all bill and discard $? - grams$ with a tf-idf score below a certain

[1] [Explain these terms in this footnote]

[2] [explain cosine similarity, and reference the exact fromula from the documentation]

threshold. Boiler plate text is expected to have very low tf-idf scores, since it appears often across bills but with low frequency within bills. Additonally, in order to detect the second kind of boiler plate, we exclude ? – *grams* that where passes several times in the same legislature, assuming that the exact same provisions will not be introduced several times in the same legislature.

4 Data

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

Bill text

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:~~ bill text as well as additional information on the legislative process for the bill. We have the first and the last version of the bill text, the sponsor of the bill and all legislative action that was taken on the bill.

~~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 billdocument 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~~

Ideological Scores

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.

4.1 ~~Tasks~~

~~Scrape Google urls for NCSL tables (FL) – lower priority Scraped 64 urls. Can't get around the Google API limitation yet. Also get blocked when trying to scrape the search result.~~

5 Alignments

~~Extract state & bill # from tables (FL) – lower priority 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). 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 alignment scores

are calculated using the affine version of the local alignment algorithm proposed by Smith and Waterman (1981) in order to match genetic sequences. The same algorithm has been used by Wilkerson et al. (2015) to trace bills through the legislative process in congress. When considering two bills, the ~~100 most similar ones) 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). data is on aci 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).~~ algorithm treats each bill as a sequence of words and tries to find the best alignment between these two sequences. Wilkerson et al. (2015) split the bills into sections and search for text reuse across sections. Since this results in about 7 billion possible pairs, first filter out pairs that don't have an exact match of a sequence of at least ten words. We encounter the same problem in our analysis to a greater extent. Assuming each bill consists of ten sections on average, there would be approximately $1.25 * 10^{12}$ possible pairs of sections.

6 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~~ We use three strategies in order 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~~ validity of text reuse as a measure of policy equivalence. First, we apply topic models to the sequences that are identified by the SW algorithm as matching sequences. This provides a rough idea of how many of the alignments are substantial and how much procedural language and not substantively significant language gets picked up by the algorithm. Second, we analyze how well the bill similarity aggregated to the state level corresponds with policy diffusion networks identified in previous research (Desmarais et al., 2015). And third, we assess the correlation of the distance of the sponsors of a bill dyad and the dyads alignment scores. If substantive policy content, and importantly, the same ideological direction of the provisions is detected by the alignment algorithm, we will expect a high negative correlation between these two measures.

6.1 Topic Modeling

We apply statistical topic modeling via Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to the alignment text with two objectives.[3] First, we want 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~~ major topical themes that emerge are consistent with policy domains that we would expect to be prominently featured in texts that capture the diffusion of public policy. Second, we want to see whether ~~the volume of state-to-state textalignment is positively related to the policy diffusion networks inferred in Desmarais et al. (2015).~~ we can identify topics that correspond to boilerplate/procedural text, which we may want to remove from the alignments to focus the analysis on substantive policy language. In our first cut we use a model with 20 topics. The average alignment is short, with little more than 25% exceeding 140 characters – the maximum length of a Tweet. There are

[3] We use the R package `mallet` (Mimno, 2013) for the topic modeling application.

52,735 unique terms in our vocabulary, after removing a standard list of stop words. In Table 3 we present the top 100 terms appearing in our corpus.

state person section states act means health public member general commission department law information provided insurance including board subsection legislature benefit court individual property order services care action business time pursuant united date year subject interstate federal compact school insurer corporation required medical chapter provide party entity vehicle amount notice people assembly agreement tax qualified provisions attorney agency limited made service authority child purposes years commissioner education report days policy employer members company code contract employee interest authorized effective statement district purpose ii application period number credit plan include senate income system amended issued reasonable part program director patient requirements

Table 1: Top 100 words, listed left-to-right, in the alignments corpus.

The topic modeling results are presented in Table 2. We see that most topics correspond to recognizable policy domains. However, some topics, such as Topic 9, which appears to be a mix of automobile policy and abortion policy, appear to consist of more than one distinct policy domain. As such, we should try running this with more topics. Topic 17 appears to be a boilerplate topic.

6.2 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

6	benefit corporation public director officer specific interests directors person purpose
16	state order child section support law application employer enforcement court
14	salts isomers food substances substance means preparation product compound mixture
3	court state order person petition section proceeding law appraisal jurisdiction
8	feet distance medical patient marijuana point qualifying south north west
20	information electronic consumer service section insurance notice portable electronics means
9	vehicle motor person means abortion woman medical child dealer death
18	property interest statement section person trust real transfer agreement financing
5	person offense age sexual years criminal court officer sex guilty
15	states united state presidential election official member vote president popular
17	state enacted legislature general assembly people amended section read senate
2	action person section court state civil violation attorney reasonable degree
7	qualified tax section investment revenue entity credit amount fund code
19	health care patient practice medical physician licensed person means services
12	department health state services act public year federal care program
4	appointed members house attorney senate power principal member authority representatives
1	individual employer year years taxable service income employee period january
10	insurance insurer policy commissioner company section coverage contract state plan
11	state commission interstate compact member states rules compacting effective law
13	school student board district education public charter state students county

Table 2: Results from 200 iterations of LDA with 10 iterations of hyper parameter optimization. Top 10 words in each topic listed for each topic. Topics listed in order of frequency of appearance within documents, beginning in the first row of the table with the most frequent topic.

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.[4]

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

state person section states act means health public member general commission department law information provided insurance including board subsection legislature benefit court individual property order services care action business time pursuant united date year subject interstate federal compact school insurer corporation required medical chapter provide party entity vehicle amount notice people assembly agreement tax qualified provisions attorney agency limited made service authority child purposes years commissioner education report days policy employer members company code contract employee interest authorized effective statement district purpose ii application period number credit plan include senate income system amended issued reasonable part program director patient requirements

Table 3: Top 100 words, listed left-to-right, in the alignments corpus.

	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 4: 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 4. 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 the log scale, the addition of a diffusion tie corresponds to a 40% increase in the expected number of cross-state alignments.[5].

6.3 ~~Tasks~~Ideological Distance of Aligned Bills

~~Develop bill metadata dataset~~

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

The MALP dataset contains ideal points for about 20,000 state legislators. The `openstates` API contains data and identifiers on 12,000 legislators. Of these we where able to uniquely match about 8,000 legislators. This allowed us to obtain ideal points for 395,474 (69%) bills and 1,745,336 ~~(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~~ 50%) pairs of bills.

In order to assess the validity of text reuse as a measure of substantive policy overlap, we expect the quality of the alignments to be inversely correlated to the distance of the bills sponsors' ideal points. The MALP ideal points are located on a common scale across all states, the distance between sponsors from different states is therefore meaningful.

In the following sections we present several analyses to assess this correlation. When we are generating the alignments the bills are split up into sections. Table 5 displays results for several methods of combining multiple alignments across sections of the same bill dyad as well as treating sections as the unit of analysis. For the case where we aggregate to the bill level, simple bivariate linear regression is used to test for a correlation, since the data has network structure we calculate standard errors with the quadratic assignment procedure (QAP) [note, Table 5 still has normal standard errors, we are runnign the QAP at the moment] (Krackardt, 1987). There are six models in Table 5. The unit of analysis of the first model is the bill section, the other five models are on the bill level. Model (1) has the alignment score of the section dyad as the dependent variable, models (2) - (5) use different methods to aggregate the scores on the section - dyad level to the bill - dyad

level.

Table 5: Regression results for log-linear models for ideology analysis. The models from left to right: 1) Alignment score by section (dependence not considered yet), 2) Sum of alignment scores of all sections of bill dyads, 3) Number of alignments of bill dyad, 4) mean alignment score of dyad, 5) penalized mean alignment score of dyad, 5) maximum alignment score of dyad.

	Dependent variable:					
	Score	Sum	Number	Mean	Penal	Max
	(1)	(2)	(3)	(4)	(5)	(6)
ideology_dist	-0.029^{***}	-0.034^{***}	-0.009^{***}	-0.025^{***}	0.002	-0.031^{***}
	(0.0003)	(0.0005)	(0.0003)	(0.0004)	(0.002)	(0.0004)
Constant	3.745^{***}	4.300^{***}	0.486^{***}	3.814^{***}	2.982^{***}	3.972^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.001)
Observations	3,957,875	1,745,336	1,745,336	1,745,336	1,745,336	1,745,336
Adjusted R^2	0.003	0.003	0.001	0.003	0.00000	0.003
Note:						*p<0.1; **p<0.05; ***p<0.01

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.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *the Journal of machine Learning research 3*, 993–1022.

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.

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.

Krackardt, D. (1987). Qap partialling as a test of spuriousness. *Social networks 9*(2), 171–186.

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

Mimno, D. (2013). *mallet: A wrapper around the Java machine learning tool MALLET*. R package version 1.0.

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*.

7 Appendix

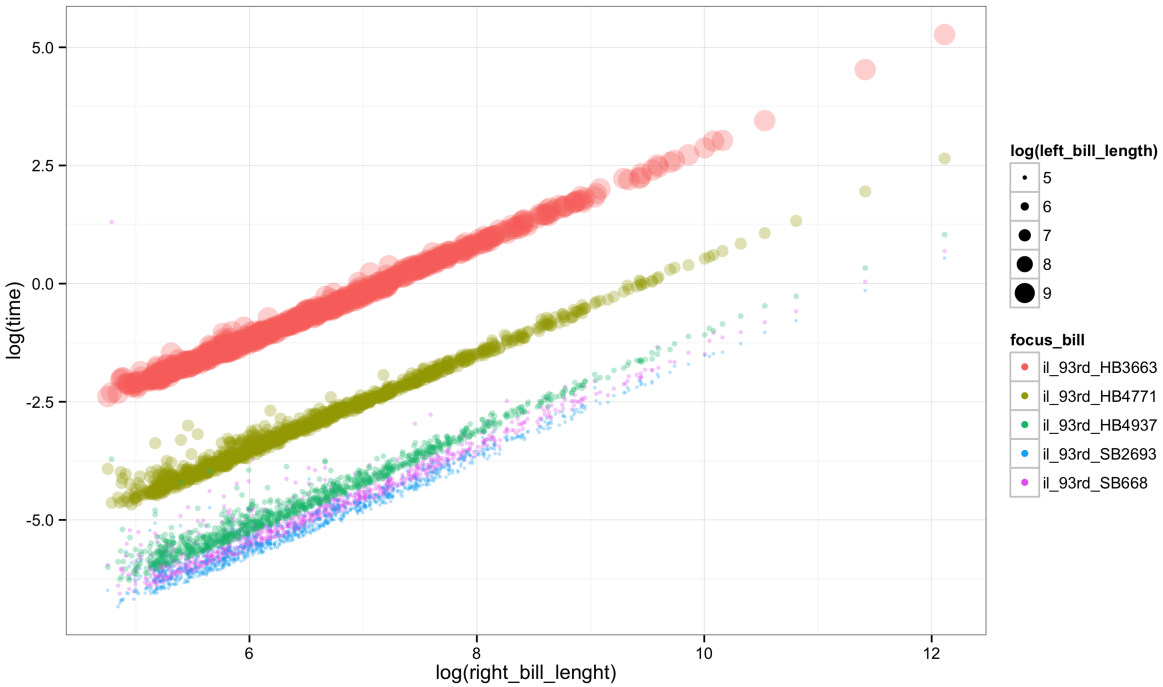

Figure 1: Computation time for selection of bills. The Colors correspond to five selected bills (focus bill or left bill). For each of the bills, the best alignment with 1000 randomly selected bills (right bills) are calculated. The size of the dots corresponds to the logarithm of the length of the focus bill (in number of words), the x-axis corresponds to the logarithm of the size of the comparison bill and the y-axis corresponds to the logarithm of the computation time (in seconds).

In order to assess the computational burden the local alignment algorithm poses we test the implementation of the alignment algorithm (Burgess et al. 2015) on a set of 1000 randomly selected bills. In a first pass we calculate the alignments between all pairs of the full bills (not divided into sections). Figure 1 displays the relationship between bill length and computation time. As expected the time increases exponentially with the length of the bills. Figure 2 displays a histogram of the logarithm of the computation time (in seconds). Computation time varies between $0.001s$ and ~~ideological similarity between sponsors~~ $1,650s$ with 75% of the distribution below $0.1s$ and a median time of $0.03s$. This shows that an exhaustive comparison of just the full bills (not separated into sections) would take about 100 years on a single machine.

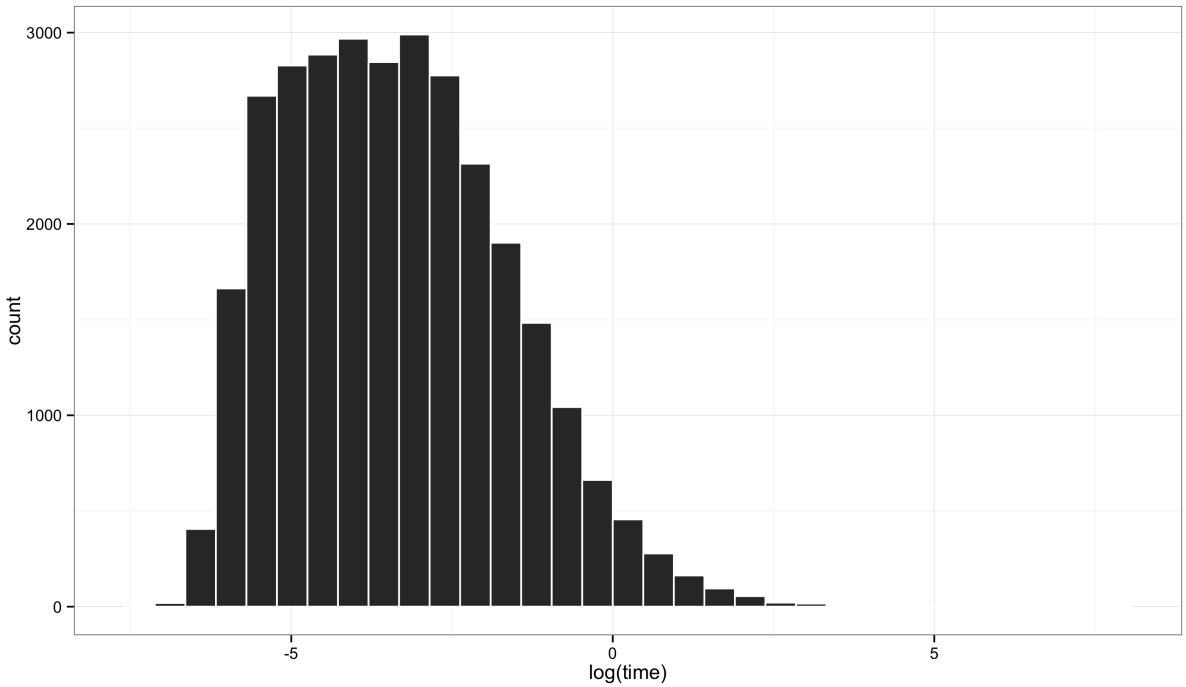

Figure 2: Histogram of computation time for all analyzed bills.

Since for this calculation there is no minimum threshold (as in Wilkerson et al. (2015) minimum of 5 matching 10-grams). Alignments between almost all bills are found and most of the alignments are procedural language. This is apparent in Figure 3 which displays the average alignment score for bills from Illinois (left panel) and Washington (right panel) for all other states in the sampled data set (x-axis). Especially for Illinois there is a much higher alignment score for bills from the same state. This shows that a minimum threshold and a classifier for procedural language are necessary.

In order to make the computation more feasible, we rely on a similar strategy as Wilkerson et al. (2015). Our bills are stored in an elastic search database wich provides a variety of search algorithms that are designed to find similar documents. Since seaching the whole database for the full document is very time consuming the following search algorithm is used to first find a pre-selection of similar bills that are then exhaustively checked for text reuse:

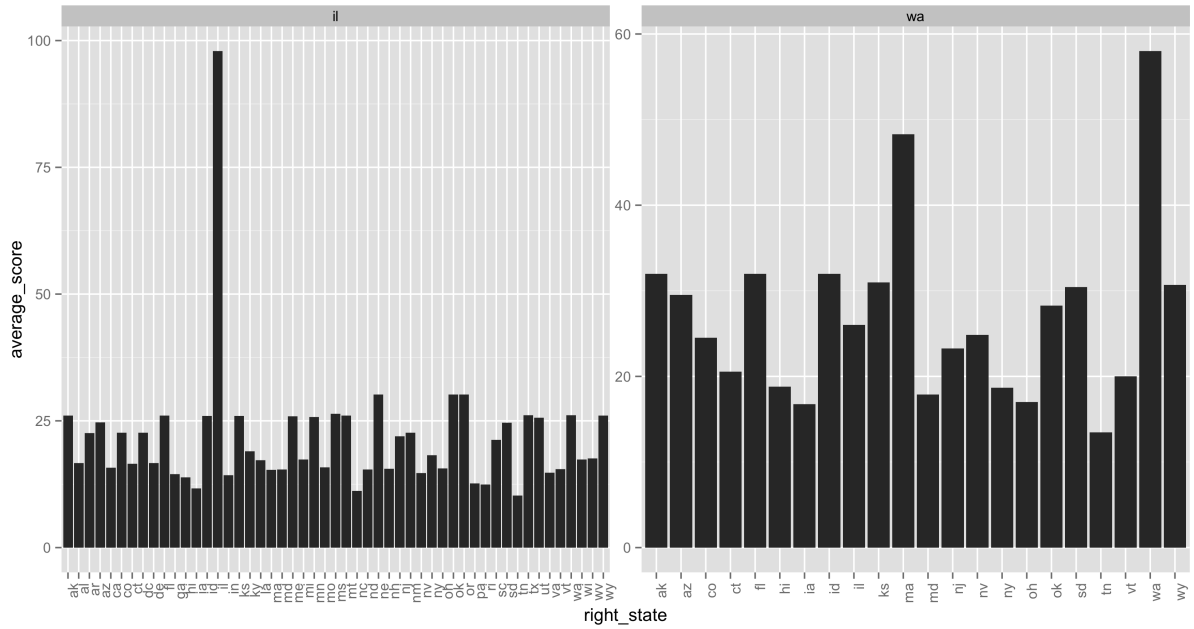

Figure 3: Average alignment scores between states. The left panel corresponds to 'left bills' from Illinois, the right panel to bills from Washington. Each bar displays the average score of the best alignment found between bills of the states.

1. ~~Analysis of alignment aggregated at the state level.~~ All bills are stored in the database in n-grams of size 2-5

2. ~~How do we scoreoverlap at the bill and state level?~~ Select the 25 n-grams with the highest tf-idf score in the query document

3. Calculate a score slightly adapted cosine similarity score between the selected n-gram vector and all full document n-gram vectors in the database[6]

In order to evaluate how well the pre screening algorithms performs we select 1000 focus bills at random and retrieve the 1000 most similar bills according to the lucene score (described above). We then calculate find the best alignments between the sections of the focus bill and the section of each of the retrieved bills. This procedure generates a dataset

[6] See `https://lucene.apache.org/core/4_9_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html` for details on the similarity scoring in Elastic Search

of approximately 4 million alignments.

For a first assessment of the usefulness of the lucene scores to select document that align well, we investigate the relationship of these scores with the alignment scores. Figure 4 displays the lucene scores against the alignment scores for a random sample of 500,000 alignments. The color indicates if the bills from which the alignment was genrated are from the same state, the larger triangles indicate scores for a bill compared to itself. In order to assess which alignments would get selected if thresholding, that is selecting for example the 100 closest bills according to the lucene score, we plotted all alignments for a sample of 9 focus bills. The green points indicate alignments from the 100 bills with the highest lucene scores for the respective focus bill.

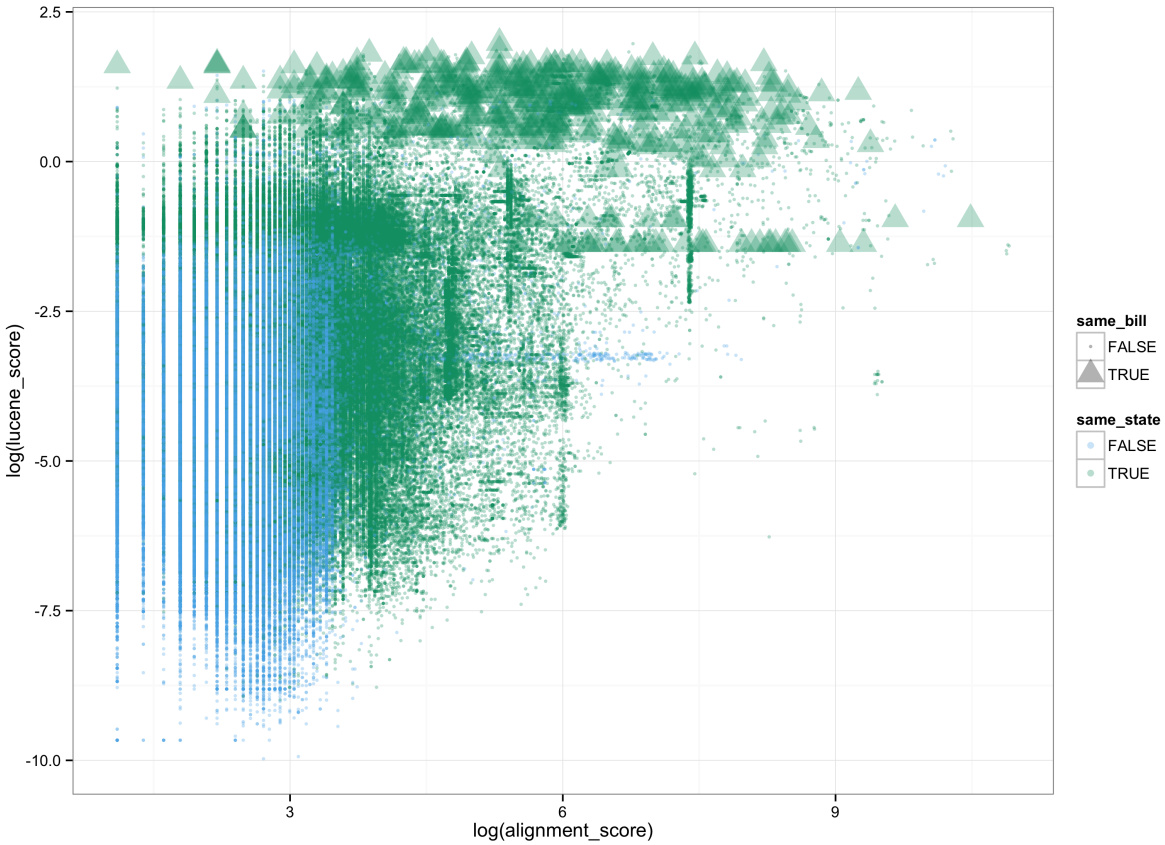

Figure 4: Alignment scores and lucene scores for a sample of 500,000 alignments (sections). The color indicates if the bills of the alignment are from the same state. The larger triangles indicate alignment and lucene scores for sections of bills compared to themselves.

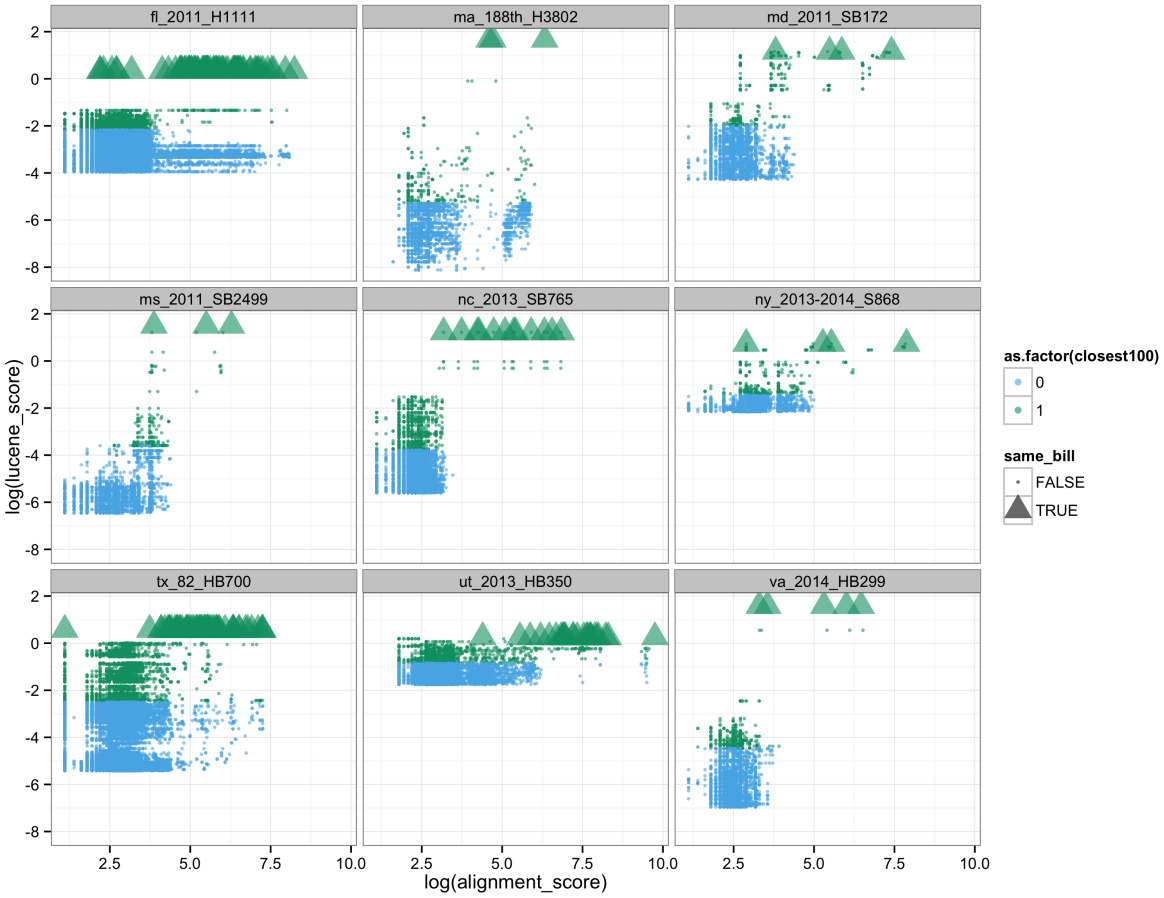

Figure 5: Alignment and lucene scores for a sample of nine focus bills (panels). Each point in the plot is an alignment for a section of one comparison bill. The color indicates, if the section comes from a bill that is among the 100 bills with the highest lucene score for the specific focus bill. The large triangles indicate that an alignment is from a comparison of the focus bill with itself.