The Legislative Influence Detector: Finding Text Reuse in State Legislation

Matthew Burgess
University of Michigan
mattburg@umich.edu

Eugenia Giraudy
UC Berkeley
egiraudy@berkeley.edu

Julian Katz-Samuels
University of Michigan
jkatzsam@umich.edu

Joe Walsh
University of Chicago
jtwalsh@uchicago.edu

Derek Willis
ProPublica
Derek.Willis@propublica.org

Lauren Haynes
University of Chicago
lnhaynes@uchicago.edu

Rayid Ghani
University of Chicago
rayid@uchicago.edu

ABSTRACT

State legislatures introduce at least 45,000 bills each year. However, we lack a clear understanding of who is actually writing those bills. As legislators often lack the time and staff to draft each bill, they frequently copy text written by other states or interest groups.

However, existing approaches to detect text reuse are slow, biased, and incomplete. Journalists or researchers who want to know where a particular bill originated must perform a largely manual search. Watchdog organizations even hire armies of volunteers to monitor legislation for matches. Given the time-consuming nature of the analysis, journalists and researchers tend to limit their analysis to a subset of topics (e.g. abortion or gun control) or a few interest groups.

This paper presents the Legislative Influence Detector (LID). LID uses the Smith-Waterman local alignment algorithm to detect sequences of text that occur in model legislation and state bills. As it is computationally too expensive to run this algorithm on a large corpus of data, we use a search engine built using Elasticsearch to limit the number of comparisons. We show how LID has found 45,405 instances of bill-to-bill text reuse and 14,137 instances of model-legislation-to-bill text reuse. LID reduces the time it takes to manually find text reuse from days to seconds.

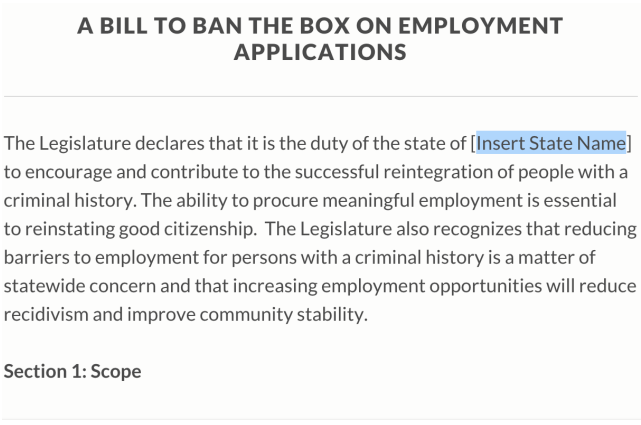
Keywords

Social Good; Government Transparency; Text Reuse; Machine Learning

1. INTRODUCTION

State governments have a central yet undervalued role in the United States. Each year, states spend more than $1.5 trillion on programs and services [14] and pass 75 times more bills than Congress [7]. State legislators play a large role by setting budgets, providing oversight, and passing laws. It is a heavy workload given that they lack the time, staff, and expertise compared to their congressional counterparts. According to the National Conference on State Legislatures, only three states have full-time legislatures with large staffs [17].

Rather than writing bills from scratch, legislators increase their productivity by using text written by others. The two most common sources state legislators draw on are legislation from other states and model legislation written by interest groups. Legislation from other states needs adaptation –for example, changing references to existing codes. Of the two, model legislation is especially easy to use: it's "legislative Mad Libs," providing blanks for legislators to fill in (see Figure 1). By lowering the cost of legislating, text reuse can help policies spread.



A BILL TO BAN THE BOX ON EMPLOYMENT APPLICATIONS

The Legislature declares that it is the duty of the state of [Insert State Name] to encourage and contribute to the successful reintegration of people with a criminal history. The ability to procure meaningful employment is essential to reinstating good citizenship. The Legislature also recognizes that reducing barriers to employment for persons with a criminal history is a matter of statewide concern and that increasing employment opportunities will reduce recidivism and improve community stability.

Section 1: Scope

Figure 1: Example model legislation from ALEC

This type of policy diffusion—whether from other states or from lobbyists—can be good. One example might be seat-belt laws, which have been credited with saving thousands of lives. In 1984, New York became the first state to pass a seat-belt law, and over the next 11 years the other 49 states followed. Policy diffusion can also be bad. For example, a relatively small number of powerful people might obtain

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939697

favorable policies at the expense of the majority by writing model legislation for state senators and representatives to turn in to law.

Americans rely on watchdogs to find legislation of questionable origin, but with at least 45,000 bills introduced in state capitols each year (see Figure 2), there is too much legislation for journalists to read, let alone analyze. Even using Google can be slow: to find good matches, the user has to read at least part of the bill to find a short piece of text to search for and then look through the results, many of which will not be legislation. The time-consuming nature of this approach makes it hard for journalists to fully understand where state legislation ideas come from, even for closely watched bills. For example, the media wrote nearly 1,000 stories about the bill Wisconsin Governor Scott Walker signed banning abortions past the 19th week of pregnancy. Some identified the lobby group behind the bill while others noted similar laws had already passed elsewhere, but we have not seen a full legislative accounting in the media, which includes all versions of the bill being introduced in 19 states. This lack of information about who is influencing state laws limits transparency and hinders citizen control, which is the defining feature of democracy.

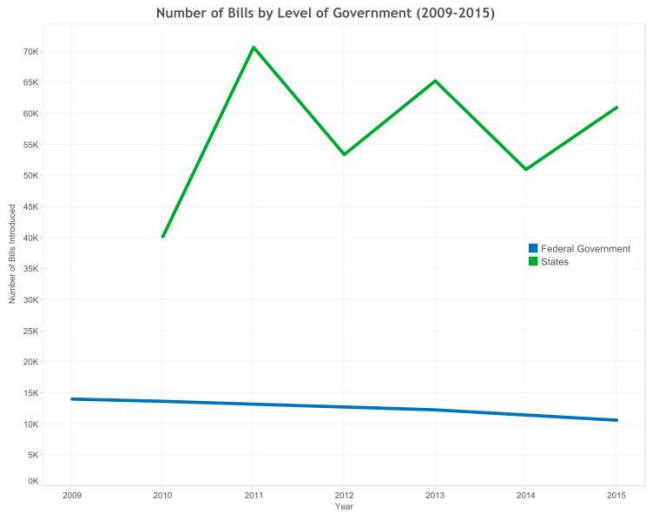


Figure 2: Legislative bills introduced each year

The high cost of analyzing text reuse in state legislative bills also harms social science. Most social scientists limit their inquiries to one or two salient topics (e.g. abortion) or interest groups (e.g. the American Legislative Exchange Council [ALEC]) as one-time studies. At its best, science reproduces findings on new data [4], but reproducibility is difficult when the cost of manual data collection is this high.

This paper presents the Legislative Influence Detector (LID), a tool that helps solve the time-consuming, human-intensive nature of existing approaches by rapidly detecting cases of text reuse in model legislation and state bills. LID uses the Smith-Waterman local alignment algorithm to detect sequences of text that occur in model legislation and state bills. As it is computationally too expensive to run this algorithm on large sets of text, we use a customized search engine built using Elasticsearch to limit the number of documents compared. LID can help journalists, scholars, and concerned citizens shed light on legislative influence by finding legislative text reuse across states and from interest groups.

We ran LID on 550,000 bills collected by the Sunlight Foundation [1] and 2,400 pieces of model legislation collected by our team and found 45,405 instances of bill-to-bill text reuse and 14,137 instances of model-legislation-to-bill text reuse. LID reduces the time it takes to manually find text reuse from minutes, hours, or even days to seconds. We have posted the code[2] and data[3] publicly for download and analysis.

2. SOURCES OF LEGISLATION

Thanks to federalism, states can act as "laboratories of democracy," trying "novel social and economic experiments without risk to the rest of the country" [3]. While this can aid the discovery and adoption of good policies, practical challenges often prevent the dissemination from happening. First, it's rarely clear what good policy is because different people prefer different outcomes [24]. Second, it can be difficult to tell whether a policy is effective. Jurisdictions rarely conduct randomized controlled trials to test policy effectiveness, instead adopting policies from states located nearby [16] or with similar ideology [10] or culture [1], from bills with symbolic value [8], or from bills backed by federal incentives [15]. As the US grew, conducting business and other affairs across state lines grew more difficult, and lawyers began to write model legislation to aid legal standardization across the country [23].

Standardization of law is not the only reason to draft legislation; lobbyists also write model bills to get their preferred policies adopted. Lobbyists have drafted model bills on a range of topics, such as crime, abortion, tax, education, the environment, and health. The goal of lobbying groups is to influence legislators, both at the national and state levels, to introduce and pass bills that resemble their model legislation as completely as possible. To make it easy for legislators and their staffs, lobbying groups provide model legislation that can be easily adapted to each state, as shown in Figure 1.

Whether it comes from state bills or from lobbyists, prewritten legislation makes passing laws easier, especially at the state level. Unlike members of Congress, few state legislators have the expertise, time, and staffs to draft legislation. It is far easier for a legislator to adapt existing legislative text than to write a bill from scratch. Because it is inefficient, if not impossible, to manually find legislative text reuse, legislators can often claim full credit for authoring bills. As lobbyists care more about implementing their preferred policies than getting credit for legislative authorship, and in many cases prefer to remain anonymous, they are often happy to let the legislator claim credit.

Legislative staff also use model legislation and bills from other states. All state legislatures have an office staffed with lawyers through which all legislation must pass before introduction. Senators and representatives can give proposed legislation to the office or ask the office to draft legislation on their behalf. A small staff of lawyers handles these requests,

[1]The Sunlight Foundation is a national, nonpartisan, nonprofit organization that uses the tools of civic tech, open data, policy analysis and journalism to make our government and politics more accountable and transparent to all.
[2]https://github.com/dssg/policy_diffusion
[3]http://dssg.uchicago.edu/lid

and each is tasked with handling a portfolio of legislation that covers several major areas of policy. These legislative staffers do not have the time to draft all original legislation, so they often borrow text from other sources, including lobbyists [2].

A bill's origin provides useful information about its content. For example, a bill written by the banks likely contains bank-friendly policy, and a bill copied from Massachusetts is probably more liberal than a bill copied from Alabama. By arming citizens and journalists with this information, they can exercise more control over their government. This is especially true at the state level, where political knowledge is lesser than at the federal level, and involves less contested issues, such as state-level land zoning.

There is evidence that the importance of model legislation is growing. Through the American Legislative Exchange Council (ALEC) and other professional groups, state-level politicians are organizing and learning from one another, and one of the tools they commonly learn to use is model legislation. In addition, congressional gridlock is encouraging lobbyists to pursue legislation at the state level instead of at the federal level [26].

3. CURRENT APPROACHES

Today, a journalist or member of the public who would like to know where a particular bill originated must perform a largely manual search. The difficult approach would be to read the bill of interest and manually skim and read hundreds of thousands of bills written elsewhere. A smarter approach would use internet search engines, but even those would be slow and tedious. For example, Google limits queries to short text strings, so the user needs to skim the bill to find a text string to search for and then parse through all the results, few of which would be legislation. Some groups use armies of volunteers to monitor legislation for matches, but given this labor-intensive process, they struggle to find all matches in a timely manner. Watchdogs and scholars can further limit the workload by focusing on a single topic (e.g. abortion [19] or crime [13]) or lobbying groups.[4] In addition, manual searches are also unreliable: while one person might flag two bills as a match, another might not. As a consequence, existing analysis tends to be limited and biased.

Political scientists have taken three approaches to automatically detecting legislative text reuse. The first uses a bag of words model. To study whether states are more likely to use model legislation or legislation from other states, Garrett and Jansa [9] use cosine similarity over bags of words to build social networks, finding that model legislation tends to have a larger influence on state legislation than legislation from other states. However, bags of words will often miss the many subtler examples of text reuse that occur in legislation. By the time some bills come to a vote, they can be called "Franken-bills" because they have stitched together legislative text from many sources. In this case, the bills—or even sections of the bills—could have dissimilar bags of words even though parts of the bills match perfectly.

The second approach uses supervised machine learning. Hertel-Fernández and Kashin [11] collected ALEC model legislation, labeled matches from ALEC Exposed—a group dedicated to finding legislative ALEC's legislative fingerprints—and text features to build a match/no match classifier. This approach is more ambitious—it offers the possibility of matching documents with similar intent but dissimilar text—but is also more difficult to use. Their model is ALEC-specific, so it would need to be retrained every time one wants to include new organizations or to compare state-to-state bills. And because it takes a holistic view toward documents, it struggles to find matches in Franken-bills.

The third approach uses a local-alignment algorithm. Local-alignment algorithms were developed to find similar subsequences within longer genetic strings, so they could find short pieces of matching text between two documents, such as those found in legislation stitched together from multiple sources. To study whether congressional bills reuse text from other congressional bills, Wilkerson et al. [27] used the Smith-Waterman algorithm to compare strings of text from congressional bills introduced since the 1990s, finding that the Affordable Care Act borrowed large pieces of text from Republican legislation introduced during the Clinton administration.

The Smith-Waterman algorithm is a good choice for comparing a relatively small number of bills, where matches may be common at the sub-document level. Because the algorithm is $O(n^2)$ costly, it can take a long time to run on a large corpus, such as state legislation. The corpus of federal bills is far smaller and so more amenable to this type of analysis. Including model legislation in the corpus exacerbates the problem. The Smith-Waterman algorithm is impractical for comprehensive analysis at the state level: by our estimates, it would take over 5,000 years for the algorithm to run on the set of bills introduced between 2010 and 2015. A number of researchers have tried this approach and quit when they realized how long it would take.

Our approach aims to overcome the shortcomings of existing approaches. First, we aim to have a generalizable approach that could be used to detect text reuse on all state bills and model legislation without having to limit our analyses to a few topics or to a specific interest group. Second, we aim to have system that will analyze all available state bills and easily update the corpus so it can provide results in near real time. Last, we aim to have our system open and available for all researchers and journalists, so that they no longer need to rely on manual processes.

4. OUR SOLUTION

LID consists of three major components as pictured in Figure 3. Given a query, LID first uses the *search module* to identify documents most likely to contain text reuse. The *alignment module* uses the Smith-Waterman local alignment algorithm to extract parallel passages of text between the query and each of the documents returned from the search module. Alignments are then scored by their probability of being substantive text via the *classification module*. We first describe our process for cleaning and chunking bills into sections and then each of main modules of LID.

All state bills share a similar format, consisting of sections that denote statements of a proposed law. It is often the case that when two documents exhibit text reuse, only certain sections of one bill are copied from the other. Since the format of state legislation is mostly uniform for a given state, our section chunker consists of a regular expression for each state. Most of the regular expressions consist of variations

[4]For example, ALEC Exposed only tries to find matches for American Legislative Exchange Council (ALEC) model legislation.

of the pattern: *\n section*. By splitting the bills by section, LID allows the user to specify a finer granularity at which to identify text reuse since we can run the alignment algorithm on specific sections of a bill.

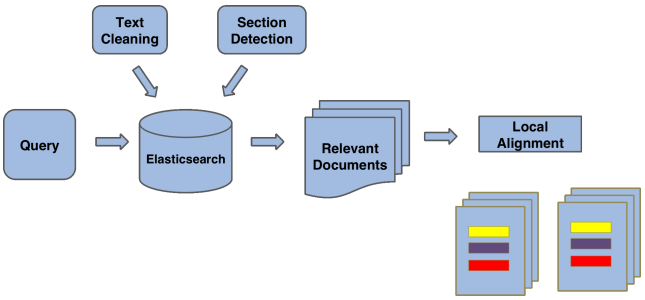


Figure 3: LID system overview.

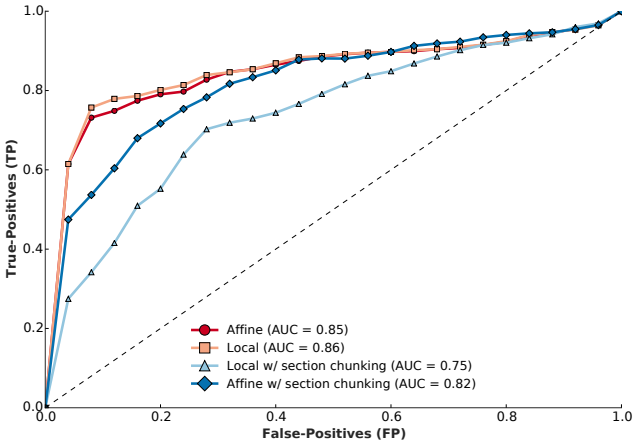


Figure 4: ROC curves for the four variants of the local alignment algorithms.

4.1 Search Module

The local alignment algorithms used to extract the similar bill text is $O(n^2)$ (where n is the length of the longest document), making it computationally too expensive to run on the entire corpus. In order to decrease the computational cost for a given query LID filters the set of documents that we run the local alignment algorithm on. This filter step can be thought of as an information retrieval task where the goal is to identify a subset of bills in the corpus most likely to contain text reuse with the query document.

In the context of this paper, text reuse occurs when two state bills share:

1. Long passages of text, i.e (sections of bills) that can differ in details.

2. Passages which contain text of substantive nature to the topic of the bill.

In addition to text that describe legal directives, state bills also contain boilerplate text that is common to all bills from a particular state or to a particular topic. Examples of legislative boilerplate include: "Read first time 01/29/16. Referred to Committee on Higher Education." and "Safety clause. The general assembly hereby finds, determines, and declares...". The first example is metadata describing where a bill is in legislative process. The second example is a clause included in almost all legislation from Colorado, stating the eligibility of a bill to be petitioned with a referendum.

A common approach for identifying documents that have text reuse is to index the documents using "shingles." Shingles are way of representing a document as a bag of words that comprises of all n-grams of a fixed size. We constructed an inverted index that contains all n-grams of length 3-5. State bills can be long documents, therefore making the computation of similarity between two documents slow. As with previous approaches for detecting text reuse [21] we down-sample the n-grams when computing the similarity between documents. Since our goal is to identify documents that share substantive text, our down-sampling technique is based on ranking the n-grams by their TF-IDF score. Filtering out common terms with high document frequency has been shown to increase efficiency [5] without sacrificing accuracy in document similarity tasks.

Our query construction using down-sampling works as follows: for a given query document, we rank the n-grams by their TF-IDF score. We then construct a query of the top 25 n-grams and search our inverted index with this query. We implemented our inverted index using the open-source search engine, Elasticsearch, configured with the standard Lucene scoring function to score the documents for a given query. The resulting list of documents is limited to be at most 300 and to have a Lucene score over 0.01. We chose the values of 300, 0.01, and 25 for the document limit, Lucene threshold and top-k n-grams as we found these values to achieve good results while still being fast.

4.2 Alignment Module

The alignment module uses the Smith-Waterman local alignment algorithm [22] to extract similar passages of text between the query document and each of the documents returned by the search module. The alignment algorithm works by identifying a sub-sequence of words in each document that has the highest alignment score. An alignment score is based on three parameters: word matches, word mismatches (when a word appears in one sub-sequence but not the other), and gaps (when the algorithm inserts a space in one of the sub-sequences). The optimal score is calculated using a dynamic program running in $O(n^2)$, where n is the length of the largest document.

We experimented with many different parameter values, as explained in Section 6. Similar to Smith *et al.* [21], we also implemented a variant of the algorithm with an affine gap penalty that incorporates two gap parameters: An initial *gap* score, and a *extended-gap* score that penalizes all contiguous gaps that follow an initial gap. Using the section chunker, we also implemented a variant of the alignment algorithm where we split the query bill into sections and used the alignment algorithm to extract alignments for each section separately.

4.3 Classification Module

As described previously, state legislation contains a lot of boilerplate text, but the alignment algorithm is indifferent to the substance of the text. An alignment produced by the local alignment algorithm consists of two sequences of text, one that is extracted from the query document and

another from the document the query is being aligned to. The classifier module uses a logistic regression classifier to predict the probability that a pair of aligned sequences contains substantive text. The classifier uses features based on the metadata of the alignments such as the mean position of where an two aligned sequences start in a document, and content features such as the Jaccard similarity between the two aligned sequences. We describe each feature used in the classifier below:

We denote a_l and a_r as the left and right sequences respectively and l_{start} and r_{start}

- **alignment length**: Number of words in the alignment
- **jaccard similarity**: Jaccard similarity between the sets of words in a_l and a_r respectively.
- **# matches**: The number of matching words between a_l and a_r.
- **# mismatches**: The number of non-matching words between a_l and a_r.
- **mean distance from top**: average starting position of the sequences, $\frac{l_{\text{start}}+r_{\text{start}}}{2}$
- **n-gram inverted frequency score**: Let N be the total number of n-gram counts in the state legislation corpus and $C(n_i)$ be the number of times the n-gram n_i occurs in the corpus. Let A be the combined set of n-grams that occur in a_l and a_r. We take the average inverted frequency score as a measure of how common the alignment text is.

$$IF(a_l, a_r) = \frac{\sum_{a \in A} \log\left(\frac{N}{C(a)}\right)}{|A|} \quad (1)$$

Our LID implementation uses $n = 4$.

Once the scores are computed, the interface to LID can present a ranking of the alignments to the user, making it easier for the user to find meaningful alignments.

5. DATA SOURCES

We use two main data sources. First, we use the Sunlight Foundation's corpus of state legislation, which includes 550,000 bills and 200,000 resolutions for all 50 states, ranging from 2007 to 2015. While for some states this corpus includes data since 2007, for the majority of states we have data as early as 2010. We do not include in our analysis data from Puerto Rico, where the text is in Spanish, and from DC, whose data includes many idiosyncrasies (e.g. correspondence from city commissions introduced as bills). On average, each state introduced 10,524 bills, with an average length of 1205 words.

Second, we have scraped more than 2400 pieces of model legislation from groups across the political spectrum, including ALEC, the best-known conservative lobbying group; the State Innovation Exchange, the best-known liberal lobbying group; the Council of State Governments; and the Uniform Law Commission.

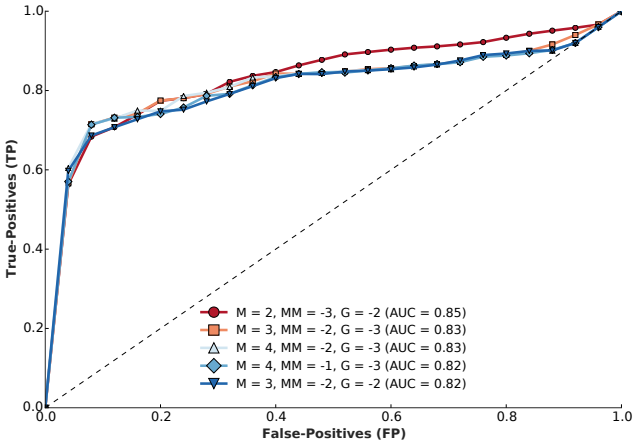


Figure 5: varying the parameters of local alignment algorithm. We denote M for the match score, MM for the missmatch score and G for the gap penalty.

6. RESULTS

6.1 Evaluating Alignment

To evaluate the alignment module we created an evaluation set that includes 165 bills with legislative text reuse. To build this evaluation set we manually read through bills that were highlighted by journalists and experts as cases of text reuse. We then grouped bills into sets with matching text. All pairs of bills within a set are given a positive label, while pairs of documents from different sets are given a negative label. We evaluate the efficacy of the alignment algorithm with ROC analysis using the alignment score as a threshold.

Figures 5 and 4 show the results for the alignment module. Figure 4 provides a comparison of four different kinds of alignment algorithms. On one hand, we vary whether the algorithm has an affine penalty. On the other hand, we vary whether an alignment algorithm finds the alignment with the largest score in a pair of documents or finds alignments by breaking up each document into sections and averaging the score found for each section. The local alignment algorithm performed marginally better than the affine local alignment algorithm. We can also see that sectioning decreased performance as compared to not sectioning. We hypothesize that this is because only certain sections of a document match one another, therefore by averaging the alignment score we may be artificially lowering the score of the few sections that actually match between two documents. Despite these results, we still think that sectioning is an important feature, enabling users to focus the granularity of their intended analysis.

Figure 5 shows how the ROC curve varies for different parameter values for the standard local alignment algorithm. As shown, the algorithms performance does not vary much with respect to the parameter settings. We chose to implement LID with the optimal parameter values: match (3), mismatch(-2), and gap (-3).

6.2 Alignment Classifier Evaluation

To train and evaluate the alignment classifier we labeled a set of alignments produced by LID. Since boilerplate text can come in the form of alignments that have a large score

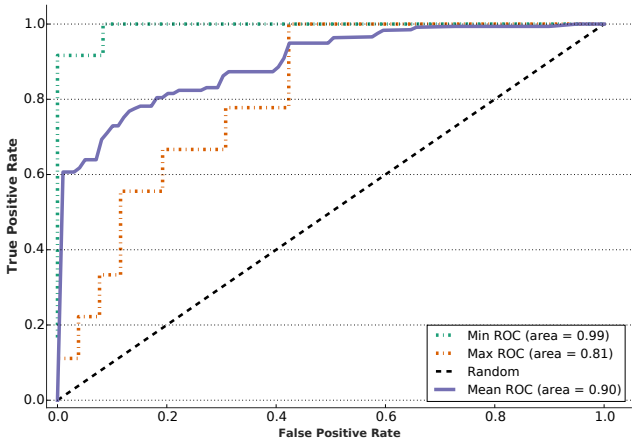


Figure 6: ROC curves for the alignment classifier. The average curve is the mean ROC curve resulting from a 10-fold cross validation.

(e.g. boilerplate sections) or alignments in the form of partial sentences, we wanted to train the classifier on training data that represented both of these scenarios. We obtained these alignments by sampling a set of 400 alignments, 200 with a score above 50 and 200 with a score below 50. We then obtained labels for each training example from 3 human experts (two political science PhD students and one political science post-doc). We only used training examples where all three annotators agreed on the label. Our final training set consisted of 354 examples. The classifier was evaluated using 10-fold cross validation on the training set. Figure 6 shows the mean, min, and max ROC curves resulting from the cross validation. The boilerplate classifier does well on average, and in the worst case still performs substantially better than random guess.

6.3 Interest Group Influence

Using LID we were able to track the influence of interest groups in each of the 50 states. For this, we collected 2400 model bills from websites of different interest groups. These pieces of model legislation can be used as inputs in our system to find state bills that share portions of text.

For the current analysis, we focus on the two largest interest group associations: ALEC (on the conservative side) and ALICE (on the liberal side). Each of these are umbrella associations that represent a large number of interest groups. For instance, ALEC represents the interests of the National Rifle Association (NRA), while ALICE represents several major labor associations. Both ALEC and ALICE have thousands of model bills, on a wide array of topics: labor rights, voting regulations, environmental issues, and economic issues. One of the main goals of both associations is to have a database of model legislation for politicians and activists to enact in state legislatures.

With LID we were able to estimate influence for each interest group. We found 5,557 ALEC bills and 2,307 ALICE bills have been introduced; however, state senators and representatives often introduce slightly different versions of the same bill multiple times. We are not sure how other researchers count these bills, so we conservatively continue the analysis only counting one version per state per year.

Using our stricter definition, LID finds state legislatures have considered 1,816 ALEC-written bills and passed 163 of them (9% success rate). In contrast, the Brookings Institute's manual effort [12] found 132 ALEC bills introduced, 12 of which were enacted (9% success rate). While these differences are large, it is worth noting that Brookings benefited from data precompiled by ALEC Exposed, including copies of ALEC model legislation and a list of ALEC's "most significant" legislation. ALEC Exposed has volunteer researchers in every state to monitor local legislation for ALEC influence.

Hertel-Fernández and Kashin [11] found ALEC introduced 10,370 bills, 1,573 of which were enacted (15%). These larger numbers should come as no surprise: their corpus of state bills is four times larger, their model is trained specifically to find ALEC-written legislation, and they use data labels that include more than word-for-word matches.

Using our stricter definition, LID finds 960 ALICE-written bills that were introduced, 84 of which passed (success rate 9%). We are unaware of another large-scale count of ALICE bills introduced to compare this to, but ALICE's success rate—which is nearly equal to ALEC's—suggests that perhaps this group should receive more attention.

Figures 7 and 8 show how many bills were introduced by state. The darker the color, the more bills introduced. These figures show that states vary in the numbers of bills introduced. For instance, Illinois and Mississippi are both highly influenced by ALEC and ALICE, while Georgia has a great influence from ALEC but a low influence from ALICE.

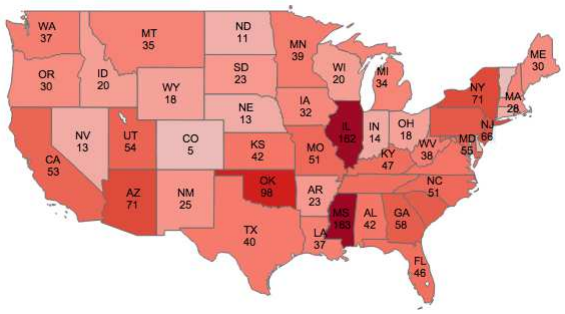


Figure 7: Introduced bills by state from ALEC model legislation

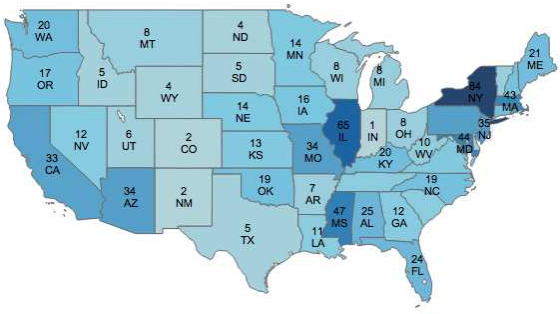


Figure 8: Introduced bills by state from ALICE model legislation

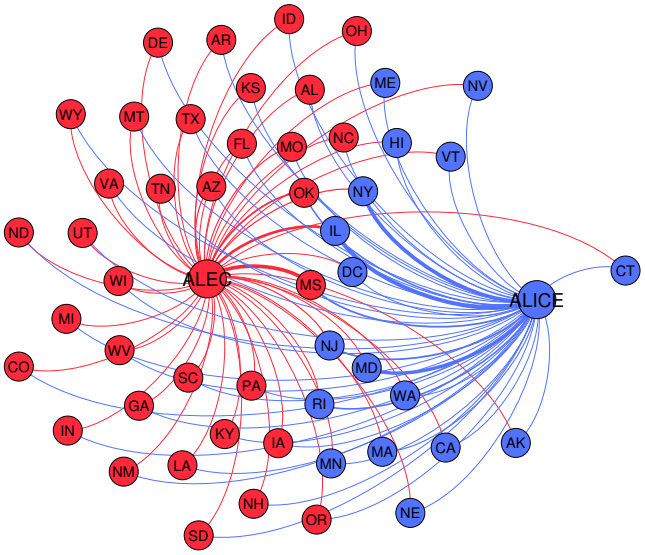


Figure 9: Influence of ALEC and ALICE across States

Figure 9 is another useful way to visualize our analysis. The network shows the influence of both ALEC and ALICE to each state. The thicker the line, the greater the influence. Circle color indicates which group successfully introduced more bills in that state. States more influenced by ALICE are in blue, while the ones more influenced by ALEC are in red. As this figure shows, most conservative states, such as Alabama, Louisiana, and North Dakota, are more influenced by ALEC, while liberal states, as Vermont, Massachusetts, and Connecticut, are more influenced by ALICE. This finding confirms that LID is working accurately.

The analysis here provided can easily be extended to other interest groups. As long as journalist or researchers have model legislation, they can, in a matter of minutes, find all the states' bills that share similar text to the model legislation. One benefit of using the Smith-Waterman algorithm is that it is does not learn from a specific lobbying group. This differs from other existing approaches, resulting in a more useful tool for users to trace the influence of interest groups.

Similarly, researchers can use LID output to go into greater detail on the type of bills that interest groups are pushing across states. LID output allows researchers to analyze the content of those bills that have matches to model legislation. For instance, researchers or journalists can use topic models or cluster analysis to understand what topics (e.g. abortion, taxes, voting rights) are most relevant for each interest group.

6.4 Case Study: Exploring Bills

Last summer, Wisconsin governor Scott Walker signed into law a bill banning non-emergency abortions past the 19th week of pregnancy. Unsurprisingly, Walker's move garnered support from one side, derision from the other, and media attention from both. However, journalists faced a big hurdle when trying to provide context for a story such as this: it is time-consuming to figure out how many states have introduced similar legislation and where it originated.

Abortion is a hotly contested issue, and abortion-related legislation gets far more attention than most. When Walker

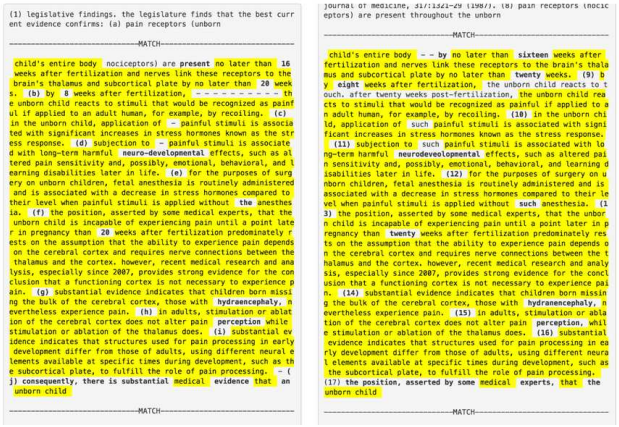


Figure 10: Match between Scott Walker's bill and a highly similar bill from Louisiana. For a detailed view, please visit http://dssg.uchicago.edu/lid/.

signed this legislation, journalists wrote nearly 1,000 articles about it, with most only reporting how many other states had similar laws on the books.[5] Few gave a more complete account of the bill's legislative history, such as where similar legislation had been considered and where it originated.

Journalists face this challenge every time they want to see if a particular controversial bill has passed in other state. For these cases, LID can be incredibly useful as it can provide a list of bills that share similar text across states in a matter of minutes.

LID allows the user to enter the text of a bill and returns documents that potentially match. The tool highlights similar sections in those documents, allowing the user to quickly evaluate the similarities. With the information about bill number, its state, and its legislative session, users can analyze which bills have passed, which ones have died, and which ones are still under consideration.

Figure 10 is a screen shot of the interactive web application of LID showing the highest-rated match for Walker's bill. The left-hand side displays text from Wisconsin Senate Bill 179 (2015), and the right-hand side displays text from Louisiana Senate Bill 593 (2012). The highlighting shows that these sections match almost perfectly. Where differences exist, they are usually numbers versus spelling (e.g. "16" versus "sixteen") or section identifiers (e.g. "(b)" versus "(9)"). Thanks to LID, we learned that similar bills had passed in Arkansas, Georgia, Idaho, Kansas, Oklahoma, Texas, and West Virginia; that similar bills were under consideration in Illinois, Iowa, Kentucky, Maryland, Oregon, South Carolina, and Virginia; and that similar bills had died in Florida, Michigan, Minnesota, Mississippi, and New Mexico. In summary, different to what all previously mentioned media outlets claimed, 19 states have at least introduced similar bills to Scott Walker's abortion law.

The interactive application of LID can be of particular usefulness for journalists. In a matter of minutes any user can find whether a bill has been introduced in any of the other 49 states in the past 5 years. In addition, the highlighting nature of LID allows user to quickly evaluate simi-

[5]Interestingly, the media gave competing counts: some claimed that 12 [25] states had passed similar measures, others claimed 14 [20], and yet others reported 20 [18].

larities across bills. This tool will be an invaluable resource for journalists covering state politics, who are normally faced with tight budgets to conduct research. As a consequence, LID will (i) increase transparency in state politics by revealing where bills come from and (ii) democratize the process of keeping state lawmakers accountable by enabling any individual to do this work.

7. LIMITATIONS

This tool sheds light on the origins of otherwise untraced state legislation, but it will not find all matches. If the wording changes enough from the first to the second document, LID will not find the match. While this is more likely on very short strings of text, which are perhaps likely to be less important policy ideas, it is a bias to keep in mind.

Many lobbyists intentionally work behind closed doors, and we are unlikely to find their model legislation in the public domain. LID cannot find pieces of model legislation written by clever lobbyists who wish to be invisible, but it can encourage transparency by raising the cost of behind-the-scenes lobbying. Before this tool, a lobbyist might be able to write one piece of model legislation to get a law passed in all 50 states. After this tool, a lobbyist would need to write a separate piece of legislation for each state to avoid detection. By identifying repeating pieces of legislative text across states, this tool can alert journalists to investigate whether there is a common source.

8. IMPLEMENTATION

This section describes how LID can be used by journalists, and what the implications are for coverage of state legislatures and policy. As previously discussed, state legislatures produce more legislation than Congress, yet the number of reporters covering state houses has declined dramatically during the past two decades, according to research by the Pew Research Center [6]. In one state, just two reporters are assigned to cover the state house.

A dwindling press corps demands an approach that maximizes the opportunities for automating portions of the reporting process. The only way for a smaller number of journalists to be able to review and research large amounts of legislation is with the assistance of software and other computational tools. There are several potential uses of LID for journalism, among them detecting similarities within a set of legislation, identifying common phrases or policies across legislatures, and outlier detection, in which the absence of similar legislation can provide insight into the origins and nature of proposed legislation.

The example of model legislation is a primary use of LID, particularly as interest groups seek to make progress in state legislatures (or to block their opponents). A journalist could monitor the spread of model legislation across state lines, enabling her to provide context for how a proposal originated and to explore how it has changed policy outcomes in other states. Ideally, a LID user could rely on model legislation already collected and add new examples to it.

Journalists also could use LID to compare different bills within a topic, whether introduced in the same legislative session or over time. Many legislative proposals are reintroduced repeatedly over time, with the goal of finding the right combination of legislative language and political context to pass the bill. LID could help determine whether proposals are identical or contain differences that might point a journalist toward important additions or deletions compared to earlier versions.

Stories can be found by looking for outliers within a collection of data, or even by the absence of data. If a journalist does not find matching legislative language in a particular state or chamber, that in itself could be a story: is a specific proposal being blocked, or saved for the right moment?

Although the focus of LID is on the legislative text, the tool also could help reveal patterns about legislators who introduce similar bills both within a legislature or across states. LID could be used to determine to what degree the language a lawmaker introduces matches that from other lawmakers of the same party in other legislatures. Combining this information with existing ideological measures and campaign contributions would provide a broader picture of a lawmaker's activity and interests.

In these ways, LID could serve as a general purpose tool for journalists covering topics such as environmental regulation or gun rights, as a resource for reporters seeking to examine a single legislator's interests and devotion to causes, and as a way to identify and trace the activity and success of interest groups seeking to influence public policy.

This tool also has wider applications than state legislation: finding matches among constitutions, referenda, and political speeches, especially during the 2016 campaign. It could be applied to press releases within a state's delegation or among an ideological cohort, or to identify informal coalitions based on the subjects they speak and write about. When political leaders introduce new phrases into a state's policy discussions, LID could be used to track the spread (or lack of it) both inside statehouses and in the wider public sphere.

9. CONCLUSION AND FUTURE WORK

This paper has shown how LID can shed light on a generally neglected topic: who is writing state legislation. In particular, LID can enable researchers, citizens, and journalists to rapidly compare state legislation across states and from model legislation. As a consequence, users will be able to understand who is writing and influencing state policies.

LID solves the shortcomings of existing approaches. First, it creates a generalizable approach that that can be use to detect text reuse on all state bills and model legislation, without limiting the analysis to a subset of topics or interest groups. Second, LID can find sequences of text across documents, which provides more accurate results than a bag of words approach. Third, LID overcomes the computational cost of using the Smith-Waterman local alignment algorithm in a corpus of more than 500,000 state bills. Last, LID is an open system, available for all researchers and journalists, so that they no longer need to rely on manual processes.

Although LID has demonstrated its value, we are planning several improvements that will increase the tool's usefulness:

- We are continuing to develop LID to make it more useful and accessible. First, we are working with journalists and interest groups to produce daily updates. As state legislators introduce bills nearly every day, we would like to run those bills through LID every night and flag text similarities for review. We will set up an alert system to notify users of potential matches.

- Our corpus of model legislation is limited even for the handful of lobbying groups included because it does not include future model legislation. In addition to scraping known lobbying websites for model bills, we are building a tool that automatically queries internet search engines for unique pieces of text in documents that do not match to model legislation in our existing corpus.

- We are improving the quality of the matches. This version of LID only credits exact matches. We would like to credit synonyms and other small differences, which LID currently treats as mismatches. This will help uncover dark-horse lobbying, where people can submit pieces of legislation that they find but are otherwise not available to the public.

- We will make the interactive tool public. At the moment, users can download datasets of potential matches but they can neither perform their own searches nor submit their own documents. We are developing a tool that enables users to provide feedback on the quality of results and then uses that information to weight each user's feedback. For example, this tool gives more weight to a user whose feedback is consistent with other users than a user whose feedback is inconsistent with other users. The tool will also enable LID to weigh user-submitted documents using feedback from other users.

All in all, the main goal of LID is to enable users to better understand who is influencing politics at the state level. We believe that LID can be an essential tool to increase government transparency, enhance accountability, and strengthen democracy.

10. ACKNOWLEDGMENTS

We thank the Eric & Wendy Schmidt Data Science for Social Good Fellowship for generously supporting this work. We also thank Irina Matveeva, James Turk, Paul Tagliamonte, Rachel Shorey, Miles Watkins, and Rob Mitchum for useful discussions. We extend special thanks to those who shared the expertise they gained working on related projects: John Wilkerson, David Smith, Nicholas Stramp, Alexander Hertel-Fernandez, and Konstantin Kashin.

References

[1] A. Abbott and S. DeViney. The welfare state as transnational event: evidence from sequences of policy adoption. *Social Science History*, 16(02):245–274, 1992.

[2] J. Bollman. Michigan legislative process. personal communication, August 12 2015.

[3] L. D. Brandeis. Dissenting opinion. *New State Ice Co. v. Liebmann*, 285, 1932.

[4] C. Drummond. Replicability is not reproducibility: nor is it good science. 2009.

[5] T. Elsayed, J. Lin, and D. W. Oard. Pairwise document similarity in large collections with mapreduce. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 265–268, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[6] J. Enda, K. E. Matsa, and J. L. Boyles. *America's Shifting Statehouse Press: Can New Players Compensate for Lost Legacy Reporters?* July 10 2014.

[7] S. I. Exchange. Top ten progressive victories of 2015. https://stateinnovation.org. Accessed January 14, 2016.

[8] R. C. Fording. Laboratories of democracy or symbolic politics. *Race and the politics of welfare reform*, pages 298–319, 2003.

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

[10] L. J. Grossback, S. Nicholson-Crotty, and D. A. Peterson. Ideology and learning in policy diffusion. *American Politics Research*, 32(5):521–545, 2004.

[11] A. Hertel-Fernandez and K. Kashin. Capturing business power across the states with text reuse. In *annual conference of the Midwest Political Science Association, Chicago, April*, pages 16–19, 2015.

[12] M. Jackman. Alec's influence over lawmaking in state legislatures. *Brookings*, 2013.

[13] S. L. Kent and J. T. Carmichael. Legislative responses to wrongful conviction: Do partisan principals and advocacy efforts influence state-level criminal justice policy? *Social science research*, 52:147–160, 2015.

[14] C. H. Lee, T. Dyson, M. Park, C. Handy, and M. B. Reynolds. State and local government finances summary: 2013. *Governments Division Briefs*, 2015.

[15] S. F. Liebschutz. The national minimum drinking-age law. *Publius: The Journal of Federalism*, 15(3):39–52, 1985.

[16] C. Z. Mooney. Modeling regional effects on state policy diffusion. *Political Research Quarterly*, 54(1):103–124, 2001.

[17] NCSL. Full- and part-time legislatures, 2014.

[18] F. News. Gop presidential hopeful scott walker signs abortion ban bill, 2015.

[19] D. J. Patton. The effect of united states supreme court intervention on the innovation and diffusion of post-roe abortion policies in the american states-university of kentucky. *National Catholic Bioethics Quarterly*, 2004.

[20] M. J. Sentinel. Scott walker signs 20-week abortion ban, 2015.

[21] D. A. Smith, R. Cordell, E. M. Dillon, N. Stramp, and J. Wilkerson. Detecting and modeling local text reuse. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 183–192, Piscataway, NJ, USA, 2014. IEEE Press.

[22] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197, 1981.

[23] R. A. Stein. *Forming a More Perfect Union: A History of the Uniform Law Commission*. Uniform Law Commission, 2013.

[24] D. A. Stone. *Policy paradox: The art of political decision making*. Norton New York, 2002.

[25] N. R. to Life. 12 states have passed measures similar to walker's, 2015.

[26] L. E. Whyte and B. Wieder. Amid federal gridlock, lobbying rises in the states. *The Center for Public Integrity*, February 11 2016.

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

.