Understanding State Lawmaking: The Role of Policy Diffusion

Nicholas Stramp
Graduate Fellow
Center for American Politics and Public Policy
Department of Political Science
University of Washington
stramp@uw.edu

April 9, 2014

Abstract

What differentiates policies by states from that are enacted from those that are not? Do general patterns of policy diffusion persist across time and policy domains? These questions are investigated using a novel dataset consisting of the text of all legislation introduced in all 50 states. Utilizing legislative text allows diffusion to be studied on a macro-level. After reviewing this unique methodology, an example policy idea is examined using the method before demonstrating the type of meta analysis that is possible, though additional validation is still needed before definitive conclusions may be drawn at the macro level.

Prepared for Presentation at the
2014 Western Political Science Association Annual Conference
Seattle, WA

Out of the tens of thousands of policy ideas introduced at the state level around the country every year, very few become law. What is special about the policies that become law? What differentiates the successful policy proposals from unsuccessful ones? This project leverages the text of legislative bills from all fifty states in order to gain a better understanding of the lawmaking process at the state level.

As a part of the larger project, policy diffusion is explored as one element of lawmaking. Instead of adopting the traditional approach of diffusion studies which focus on a single issue, it is examined from a macro perspective by mapping the flow of policy ideas across all areas and all fifty states. This seemingly overwhelming task is accomplished by relying on text analysis methods to identify all instances of policy diffusion for the time period under study. As this is an entirely new approach to diffusion research, the goals of this paper are modest. First, to develop a method that can aggregate and measure all instances of policy diffusion across all states. Second, to demonstrate the validity of this method through a single-policy case study.

This paper begins by making the case for gaining a greater understanding of the state lawmaking process through analysis of legislative text at the level of policy ideas before explaining the method employed to convert text into a dataset amenable for study. The dataset is then leveraged to explore policy diffusion in the innovative method outlined above. Next, a single case of diffusion is identified using this macro-method followed by the presentation of preliminary results of all-issue diffusion. Finally some challenges identified with the method are discussed before concluding with a review of the next steps in the research process.

State Lawmaking and Diffusion Research

Much research has been done at the Congressional level on the lawmaking process, particularly in the areas of agenda setting and decision-making. Do the same rules of the game apply at the state level? What pressures exist that act as positive and negative influences on the agenda setting process? This project has the broad goal of increasing understanding of the state lawmaking process, while this paper focuses on the often studied area of policy diffusion as one piece of the puzzle that is state lawmaking.

Past research of U.S. policy diffusion has primarily focused on diffusion mechanisms at work in specific policy domains such as state lottery adoption (Berry and Berry, 1990), tax policy (Berry and Berry, 1992), anti-smoking policies (Shipan and Volden, 2006), and more. We also know that diffusion generally occurs for one of the following four reasons: learning, competition, coercion, or socialization/imitation (Graham, Shipan and Volden, 2012; Shipan and Volden, 2008). At the same time there is still much that is not understood about diffusion, as evidenced by multitude of articles that critique the field, summarize findings, and call for future research (Shipan and Volden, 2012, 2008; Graham, Shipan and Volden, 2012). Perhaps these concerns are due in part to the rather isolated focus on policy diffusion rather than considering it as part of a larger, multi-dimensional dynamics of the state lawmaking process.

Graham, Shipan and Volden (2012) have identified several goals for future diffusion research:

- Develop a systematic, general understanding of how diffusion works

- Identify robust diffusion processes that persist across multiple policy domains[1]
- Explore the evolution of policies as they diffuse
- Expand the scope of diffusion research beyond adoption to examine how issues arise on the agenda.
- Explain why some areas do not adopt popular policies while others do

These goals can also be applied to the lawmaking process more generally. In order to gain a general understanding of diffusion, it is important to also understand how diffusion interacts with other forces at play in state lawmaking such as party control of the agenda and level of legislative professionalization. In addition to exploring the evolution of diffusing policy ideas, a first step would be to better understand how policy ideas evolve over time within a state. Are more mature ideas more likely to diffuse? Finally, the policies that do not diffuse and are not adopted are just as potentially informative as those that do diffuse and are adopted.

The Policy Idea

Diffusion research generally focuses on either a single policy proposal or action across an entire policy domain. However, traditional studies of the legislative process focus on the bill as the unit of analysis and by starting with bill text there is a natural inclination to conduct the analysis at the bill level. Bills are really just

1 (Boushey, 2010) is a notable exception in that he does offer a general theory of diffusion across a multitude of policy domains.

"vehicles" for policy ideas (Adler and Wilkerson, 2013) and it is these ideas within bills that are target of this analysis.

A policy idea is loosely defined as a policy proposal that a human can easily identify within legislative text and summarize in a simple sentence. Examples include "exempting concealed carry permit records from FOIA requests" and "decreasing the tax on lottery winnings by 10 percent." Policy ideas are independent units that could easily be 'picked-up' in one state and introduced in another. Due to variations in existing state law and bill-writing syntax, it is not expected that states generally mirror entire bills from other jurisdictions but rather pick and chose specific policy provisions that they want to propose for their state.

Tracing Lawmaking and Diffusion in Text

Most diffusion studies utilize some variation of a common research strategy: identify a policy (or policies) to study, gather information about policy proposals and adoptions from a set of nearby governments using keyword searches of legislative and media databases, verify policy congruence through manual reading and coding, and finally explain the pattern found in the case. As is often the case in political science, each study uses slightly different variables, timeframes, definitions, etcetera which makes it very difficult to compare existing research and mechanisms against one another.

The starting point for this research is the text of all legislation introduced across all states over a defined period of time. The text of each individual bill is

evaluated against all bills from other states.[2] Using actual legislative text instead of other methods has several advantages. A layer of analysis is removed compared to keyword searches or relying on bill summaries, patterns of diffusion across traditional policy domains can be identified, and by starting outside of a policy domain it is much easier to test a mechanism found in one domain on another[3] Additionally it is much easier to look at ideas that do not diffuse and ideas that diffuse but are not adopted. Finally, by not limiting the scope to a specific policy or policies, a much more wholesome picture of diffusion processes can be developed. There are also challenges to utilizing text in this way. The volume of text is clearly too much for human coders to handle, so we turn to automated methods of identifying common text within a set of legislative bills. Without extensive human validation and careful methodological planning, automated methods of text analysis may not provide useful results.

Methodology

The research framework outlined above presents two major challenges. The first is to collect and store the text of bills from all fifty states and the second is to identify instances of shared policy ideas within this database.

2 Along with the actual text of a bill, several pieces of metadata accompanying the bill are also important for analysis including who introduced the bill and how far the bill progressed in the legislative process.

3 Ideally domain-specific research would not be necessary under this framework, but it is still possible.

Data Collection

The data collection for this project starts with metadata for recent state legislative bills from all fifty states. This information is aggregated by the Sunlight Foundation as part of their Open States project and is available for download via an API. The Sunlight Foundation collects this data continually through a collaboratively-built set of "scrapers" that extract information directly from the websites of each state legislature. Characteristics of each bill include the bill's title, date of introduction, dates of passage by one or both chambers (if applicable), date of enrollment, sponsor, as well as records for each version of every bill. For each version, a link to the original text of the bill (hosted on each state's servers) is included. Data from 2011 to present is available for all states; for a small number of states (less than 15) the data is available stretching back several additional years.

The first major phase of this project involved downloading the bill text for each state, totaling over 900,000 documents. The raw text for each bill version was cleaned of extraneous formatting, divided into sections, and stored in a MySQL database. A unique challenge encountered during this process is that nearly all states write legislation in an "add and delete" format. When the text of a bill will amend existing law, the existing text is customarily included in the bill, with text formatting used to denote new material to be added as well as old material to be deleted. Most states indicate new material with bolded or underlined text and deleted text as struck-through or enclosed within brackets. In order to avoid working with text that was not actually part of the bill under consideration, the text cleaning process removed all text designated as stricken in accordance with each

state's formatting standards. For now, the text of existing law surrounding the new text is still included for analysis.

About sixty percent of bills were available as .html files, which were parsed relatively easily with Beautiful Soup (python). Roughly 10 percent of bills were available as Microsoft Word (.doc and .docx) files; these documents were first converted to .html using textutil, a command line file conversion toolkit available with OS X from Apple. Once converted to .html, the documents were then processed with BeautifulSoup. The remaining 40 percent of documents were only available as .pdf files. Although many utilities exist for extracting text from pdfs, the need to preserve text formatting in order to remove stricken text proved to make the process more challenging. The pdf format stores underlines and strikethroughs as lines appearing in a specific place in relation to the text, rather than as a formatting attribute of the text itself (such as bold or italics).[4] All third-party pdf conversion utilities tested were not able to recapture this formatting while still producing readable text. Adobe Acrobat Professional was the only program that reliably captured these types of text formatting while converting documents from pdf to html. An Adobe batch sequence was used to convert the .pdf files to .html.[5]

Computational Methodology: Text Reuse Approach

There are several methods available for identifying similar texts. Two main approaches are "bag of words" and "longest common subsequence." The first approach disregards word order and looks for similar word frequencies. The second

4 See http://stackoverflow.com/questions/15577689/scraping-text-from-pdf-with-underlines-and-strikethroughs.

5 Unfortunately this process is quite slow, taking from 1-3 seconds per file. Total processing time was about 200 hours.

preserves word order, but requires exact matches between two texts. Both of these approaches have major drawbacks–the first loses a lot of potentially important information contained in word ordering while the second is rather inflexible. Flexibility is key for this analysis as it is expected that each state will have slightly different syntax and conventions.

An alternative to these common approaches comes from genetic sequencing research. Geneticists have been long been interested in detecting similarities and differences in DNA sequences (i.e. ATCGATTGAGCTCTAGCG). There are two types of alignments that are most commonly used. Global alignments compute a score for the overall similarity between two sequences. The score is computed using a scoring process in which matching characters gain positive points while mismatched characters receive negative points. Importantly, the algorithm allows for "gaps", meaning an extra character in the middle of a sequence will not cause the remainder of the sequence to be mismatched. The other main alignment type is a local alignment. As its name implies, a local alignment algorithm determines the best match between two sequences, adjusting the range included from each sequence until an optimal score is reached. A similar scoring system is used, and users can adjust the weights assigned to matches, mismatches, and gaps.

Both of these methods translate well to text reuse, as the DNA sequences are simply text strings. Because of the nature of legislative texts, a local alignment method is preferred in order to identify specific passages of text that are shared. Specifically, the Smith-Waterman local alignment algorithm is used (Smith and Waterman, 1981). This is the same method used to identify shared policy ideas at the Congressional level (Wilkerson, Stramp and Smith, 2013) and more information on the algorithm is available in the appendix. Because each bill likely contains

multiple policy ideas, bills are divided into sections before being analyzed. All fifty states divide their bills into sections and the general convention is that each bill section "shall contain, as nearly as may be possible, a single proposition of enactment" (Bellis, 2008). By dividing bills into sections, multiple instances of shared policy ideas may be identified between two bills.

Applying the Method

The dataset currently contains text from 545,000 bill versions (306,000 unique bills). In total there are 2.9 million texts, each representing a section from a specific bill. Managing this amount of text is challenging as is, but comparing each section against every section from another state would result in approximately 8 trillion comparisons, a truly unattainable number for political science and a challenge even in computer science. Fortunately, with assistance from David Smith (Department of Computer Science, Northeastern University), a filtering process was developed that indexes each text in the way a search engine would, and then only pairs of sections that have a minimum "n-gram" similarity, in this case a string of 10 words, are scored on the algorithm. This approach is known as "hash-based two pass" (Huston, Moffat and Croft, 2011). The reduction in number of comparisons is staggering, with the number of comparisons to evaluate dropping to under 900,000–approximately 0.00001% of all possible comparisons.

The filtering and comparisons were executed on a cluster of Amazon's Elastic Compute Cloud instances, managed using StarCluster.[6] David Smith's passim

6 A 16-node cluster was utilized, allowing for 64-fold parallelism. In other words, this allows for the speed of computation to be increased by approximately 64-fold, as each parallel core independently works on an assigned task.

workflow[7] automated the process of indexing the documents, reducing the number of candidate section pairs, and producing the comparison results.

From Alignments to Common Policy Ideas

The raw output of this method is a set of metrics for each of the 900,000 identified potential comparisons. For each pair of texts, the entire length of each text was compared against the other and the resulting output is a smaller "alignment" that consists of the largest shared passage between the two texts. The vast majority of comparisons within this set of results are not instances of shared policy ideas, but rather fragments of text that happen to be the same. Based on prior research using this method (Wilkerson, Stramp and Smith, 2013), the score produced by the Smith-Waterman algorithm does a reasonably good job predicting shared policy ideas above a threshold. For this initial analysis, the threshold was set at 2000. For an alignment to achieve this score, a minimum of approximately 150 words must be shared with minimal gaps. Preliminarily human validation of 100 random comparisons above this threshold yielded 86 policy ideas matches.

A constant concern with this approach is the occurrence of "boilerplate" text. This is text that often appears in bills but is not substantive policy language. A common example is enactment clauses such as "this act shall take affect on July 1, 2013." Hundreds of bills likely contain this phrase, but it is not a policy idea. Multiple checks are used to filter out boilerplate text. At the congressional level this is a major challenge because all bills are working under a common set of formal requirements and customs. Initial indications are that this is less of an issue for

7 Available on github at: `https://github.com/dasmiq/passim/`. More information and instructions on replicating this process, including utilizing StarCluster is available on the github page.

comparing legislation from different states, as each state has its own procedure and customs for "wrapping" a policy idea in formal legislative language. As these wrappers are likely different in each state, identifying the common text between states should have the effect of unwrapping the actual policy idea.

Nevertheless, several steps are taken to remove boilerplate comparisons. The indexing phase of the analysis does not consider n-grams that appear more than 100 times, immediately excluding short enactment clauses from consideration such such as the example above. Additionally, setting a high threshold score eliminates many potential boilerplate cases. The next phase of this project will involve more extensive human coding of alignments across a broad range of scores. The alignments identified as boilerplate will be used to train a supervised machine learning algorithm to filter out these results. Once this filtering mechanism is in place and a greater human-coded sample is obtained, the alignment score threshold will likely be lowered without sacrificing accuracy.

An additional filter is applied to restrict the comparisons to the most recent version published of each bill. States vary widely in the number of versions available for each bill. Several states only provide a single version whereas some provide up to 15 versions per bill. By only considering one version of each bill, duplicate comparisons are avoided. The resulting dataset consists of approximately 30,000 pairs of alignments. For each alignment pair, we know its score, which states and bills that are aligned, the aligned text, and differences between the two texts.

Initial Results

Analysis at many levels is possible with this data, but to start the analysis will focus on a single policy idea—establishing the process for transferring guardianship when an individual moves from one state to another. This policy idea was selected by filtering sections by the number of highly scoring matches and then selecting a policy at random that had matches in at least 10 states. Indiana's enacted version of this policy idea was used as the "base", from which first and second order links were aggregated and filtered to only include high scoring alignments. Sixteen states had primary comparisons to the Indiana provision, and one additional state (Hawaii) was detected by examining the comparisons for sections linked to the original Indiana provision.

With very minimal processing, figure 1 was developed which illustrates activity related to this shared policy idea from 2009 to the present[8]. Nine states adopted this proposal during the timeframe and many states introduced it repeatedly. Why did some states not adopt this relatively innocuous policy idea? Why did so many states consider this provision?

8 Data is only available before 2009 for a small number of states. It is possible that additional states introduced and passed this policy idea in 2008 or earlier.

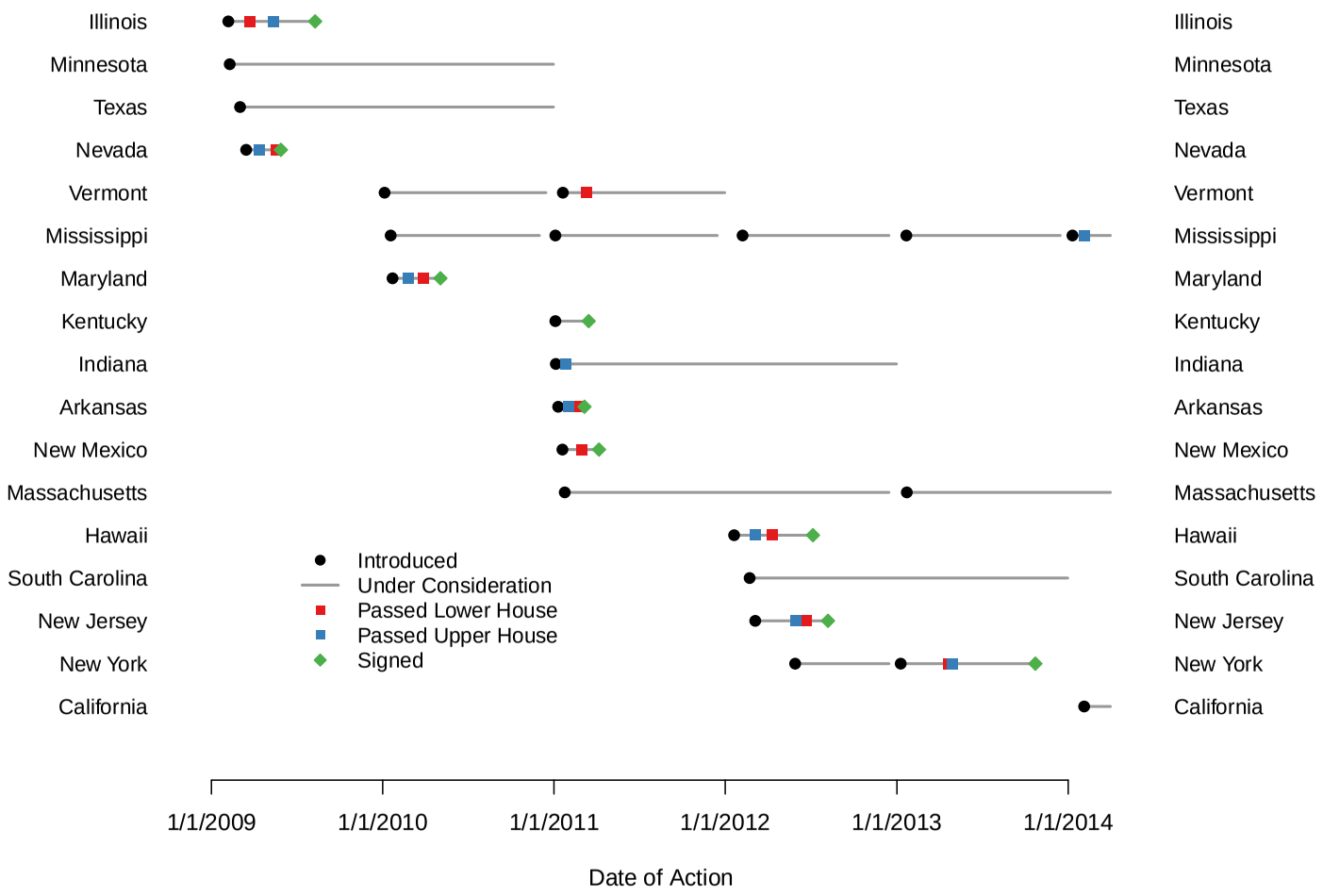


Figure 1. Activity Related to "Guardianship Transfer" Policy Idea

This figure captures the dozens of datapoints related to this policy: introduction dates, chamber passage, final passage, as well as evidence of when the idea was introduced but failed to pass. The answer to the second question is a specific external actor, the Uniform Legal Commission (ULC). After selecting this policy idea as an example, online research determined that this provision is part of a ULC project, the Uniform Probate Code.[9] The ULC is an organization with members from all fifty states that "provides states with non-partisan, well conceived, and well drafted legislation that brings clarity and stability to critical areas of state statutory law."[10].

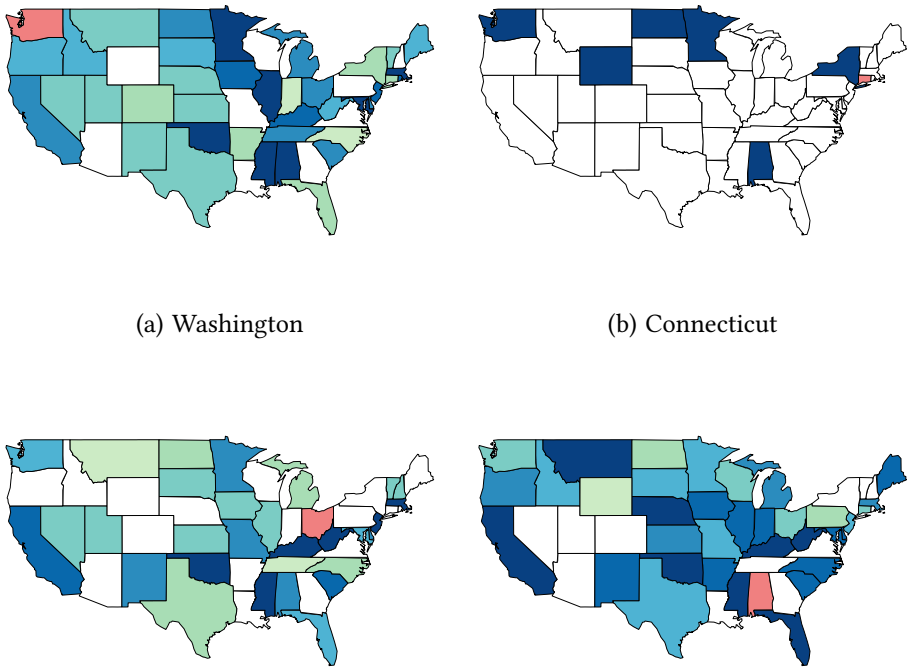
Exploration at the idea level using this method has the potential to uncover new trends in diffusion as well as facilitate the mapping of diffusion processes across a wide range of policy ideas or issues relatively easily. An intermediate goal of this project is to automate and validate the creation of this type of visualization so that it may be drawn from any selected policy idea occurrence within a given state. Along with the useful visual illustration are copious datapoints from each state that considered an idea. This method also has the potential to allow for a macro-level view of overlapping policy ideas around the country as well.

Exploring Macro Level Trends

In addition to examining individual policy ideas, this method also aggregates policy idea matches across all legislative text for all states, allowing for a level of analysis previously not reachable. The maps shown in figure 2 identify shared policy language for each of four states. The darker blue shading indicates that a

9 http://www.uniformlaws.org/Act.aspx?title=Probate%20Code
10 http://www.uniformlaws.org/Default.aspx

greater proportion of a state's enacted legislation (as measured by the total character count of enacted legislation) is shared with another state. This figure is a work in progress, as more human validation is needed. Several trends warrant additional investigation, including the prevalence of Mississippi in all four highlighted states (and most other states as well). As a relatively small, southern state with a mostly part-time legislature (ranked 40th by Squire (2012)), this is not a state that would be expected to be an epicenter for policy ideas. A preliminary investigation uncovered a few potential factors contributing to Mississippi's apparent rise to the top of this category: Mississippi produces a lot of bills (nearly 6000 a year) and has a new session every year, meaning bills must be introduced yearly, while many other states are on a biennial cycle. It is likely the case, as shown in figure 1, that the same bills are introduced each year in Mississippi, artificially inflating its position as a popular source.



(a) Washington (b) Connecticut

(c) Ohio (d) Alabama

Figure 2. Mapping common policy language enacted in four states to all proposed legislation across the country, 2009-2013. Darker colors indicate that a greater portion of a given state's enacted legislation is shared with a specific state.

Challenges

Although this project has already overcome many challenges, more still lay ahead. Three key challenges will be briefly discussed: the need for even more data, finding the true "origin" of policy ideas, and challenges based on bill text structure.

Although more than 300,000 bills is a lot of information, for most states this only captures 2-4 years, an incredibly short period of time from a legislative perspective. Ideas are often introduced again and again, slightly tweaked each time, before they become law. At a minimum, it is anticipated that 6-8 years of data for each state will be necessary to have a more accurate picture of the diffusion of ideas on both macro and micro scales. Most states have additional bill text available on their websites that is not currently aggregated by OpenStates. Additional scrapers will need to be written to collect this additional data. Fortunately the OpenStates dataset is continually expanding as time passes. Additional years of data also allows for an in-depth examination of how a policy idea evolves over time.

The second challenge is closely related to the first. In order to determine the legislative origin of a policy idea, more history is needed. Even with 6-8 years of legislative history it will still likely not be possible to label one state or representative as the legislative origin of an idea. There is also a distinct possibility, as in the guardianship example above, that the origin of a policy idea is a non-governmental body. This method is expected to be much more useful to describe *how* a policy idea travels and evolves rather than where it came from, though it is conceivable that the earliest legislative instances of a policy idea could be compared against contemporary external sources such as interest group literature.

The last challenge is more practical in nature. As noted at the outset, most state write legislation that alters existing legislation by replicating the existing legislation, and then making changes in a "mark-up" fashion, explicitly identified what passages to amend and delete. As a result, the legislative text as interpreted at this point usually includes existing statutory provisions along with the new

changes. This poses a challenge when comparing alignments as sometimes an alignment will pick up two passages of existing law that are similar rather than two proposed changes. This became apparent when researching all firearm legislation that passed in all states in 2013. Initially the more than 130 law provisions relating to firearms passed in 2013 were to be a case study for this paper, but it soon became apparent that the current method needs tweaking in order to avoid matching passages of existing firearm regulations within states. One of the next benchmarks for this project is to modify the process so that it is at least as reliable as a human coder at identifying shared policy ideas across a common policy domain such as firearms.

Conclusion

This paper marks the first step towards achieving the goal of measuring policy diffusion on a macro-scale as an element of the state lawmaking process. It is also particularly well-suited to track the evolution of policies over time. Not only do we know the exact wording of a policy idea at each occurrence, we also know exactly what has changed. It is expected that this evolution, or tinkering, is key to understanding when policies are adopted and how they change over time. Similarly, this method is designed to look beyond policy adoptions to all policies introduced. Future iterations plan on further tracking policy idea progression by identifying which ideas (as part of bills) get committee hearings, are debated, pass one chamber, etc. One of the most understudied areas of diffusion research is why governments *do not* adopt a specific policy idea. These decisions are potentially just as important as adoptions, but get very little attention. As demonstrated in

the guardianship example in figure 1, many states did not even formally consider this policy[11] and about half of those that did failed to enact the legislation.

This initial research establishes that this method may be used to identify singular instances of policy diffusion and presents a framework of how diffusion can be aggregated to the level of all lawmaking activity. More work is still needed to perfect the process and develop valid and reliable meta-level metrics of policy diffusion. This research method has the potential to not only provide a standardized way of evaluating existing theories of diffusion but also the ability to aggregate diffusion to the point of discovering general trends of information flow among the states. Working with large datasets in this manner requires careful consideration of potential challenges and significant investment in human validation of results, but it also has the potential to open an entirely new and more comprehensive strain of policy diffusion and state lawmaking research.

11 In the timeframe under investigation.

References

Adler, E Scott and John D Wilkerson. 2013. *Congress and the Politics of Problem Solving*. Cambridge University Press.

Bellis, M. Douglass. 2008. *Statutory Structure and Legislative Drafting Conventions: A Primer for Judges*. Federal Judicial Center.

Berry, Frances Stokes and William D Berry. 1990. "State Lottery Adoptions as Policy Innovations: An Event History Analysis." *The American Political Science Review* 84(2):395.

Berry, Frances Stokes and William D Berry. 1992. "Tax Innovation in the States: Capitalizing on Political Opportunity." *American Journal of Political Science* 36(3):715.

Boushey, Graeme. 2010. *Policy diffusion dynamics in America*. Cambridge University Press.

Graham, Erin R, Charles R Shipan and Craig Volden. 2012. "The Diffusion of Policy Diffusion Research in Political Science." *British Journal of Political Science* 43(03):673–701.

Huston, Samuel, Alistair Moffat and W. Bruce Croft. 2011. Efficient indexing of repeated n-grams. In *Proceedings of the fourth ACM international conference on Web search and data mining*. WSDM '11 New York, NY, USA: ACM pp. 127–136.

Shipan, C R and C Volden. 2012. "Policy Diffusion: Seven Lessons for Scholars and Practitioners - Shipan - 2012 - Public Administration Review - Wiley Online Library." *Public Administration Review* .

Shipan, Charles R and Craig 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.

Shipan, Charles R and Craig Volden. 2008. "The Mechanisms of Policy Diffusion." *American Journal of Political Science* 52(4):840–857.

Smith, T. F. and M. S. Waterman. 1981. "Identification of common molecular subsequences." *Journal of molecular biology* 147(1):195–197.
URL: *http://view.ncbi.nlm.nih.gov/pubmed/7265238*

Squire, Peverill. 2012. *The Evolution of American Legislatures: Colonies, Territories, and States, 1619-2009*. University of Michigan Press.

Wilkerson, John, Nick Stramp and David Smith. 2013. Tracing the Flow of Policy Ideas in Legislatures: A Computational Approach. In *APSA*.