Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach

John Wilkerson[1]
David Smith[2]
Nick Stramp[1]

September 17, 2013

Abstract

We propose a method for tracking policy ideas in legislation. In the US Congress, only a very small proportion of bills become law. Surviving bills likely serve as vehicles for policy ideas originating in other bills, but there is currently no reliable way to learn when this occurs. Using the legislative history of the Patient Protection and Affordable Care Act as our test bed, we investigate whether ”text reuse” methods can help to shed additional light on policy development and lawmaking. In particular we ask whether lawmaking is more inclusive when judged in terms of the progress of ideas rather than the progress of bills.

[1]University of Washington, [2]Northeastern University. Prepared for delivery at the 2013 New Directions in Text as Data workshop, London School of Economics, Sept. 27-29. Please send comments or questions to jwilker@uw.edu.

1 Introduction

An irony of the Patient Protection and Affordable Care Act (Obamacare) is that one of its key provisions, the individual insurance mandate, has conservative origins.[1] In Congress, the requirement that individuals to purchase health insurance first emerged in Republican health care reform bills introduced in 1993 as alternatives to the Clinton plan. The mandate was also a prominent feature of the Massachusetts plan passed under Governor Mitt Romney in 2006. According to Romney, "we got the idea of an individual mandate from [Newt Gingrich], and [Newt] got it from the Heritage Foundation."

To date, systematic studies of lawmaking in Congress have focused almost exclusively on the progress of bills. Yet most bills never become law, and the versions of those that do often evolve to the point that the enacted version of a bill is substantially different from the one that was introduced. What is needed is an approach to tracing the progress of the policy ideas lawmakers propose, rather than just the progress of their bills.

We begin by discussing why attention to policy ideas can be informative. We then approach the tracing of policy ideas as scholars in other fields approach plagiarism detection and genetic matching. Whether these methods will also work to capture policy ideas is an unanswered question. A central focus is on validation. We develop a gold standard dataset of human-labeled cases, and

[1] http://www.forbes.com/sites/aroy/2012/02/07/the-tortuous-conservative-history-of-the-individual-mandate

then assess performance using out-of-sample prediction. Lastly we illustrate the method's potential by tracing the legislative origins of ideas found in the Patient Protection and Affordable Care Act (PPACA).

2 Policy Ideas in Legislation

How often do laws incorporate policy ideas originally proposed in other bills? How often do laws sponsored by members of one party include ideas proposed by members of the other party? Are ideas more likely to be shared in the House or Senate, or in some committees? Which legislators' ideas get picked up and why? How have political developments, such as increasing partisan polarization altered this policy development process? These questions have never been investigated because they require a method for systematically tracing policy ideas in legislation.

Examples of the potential value of such a method are easy to come by:

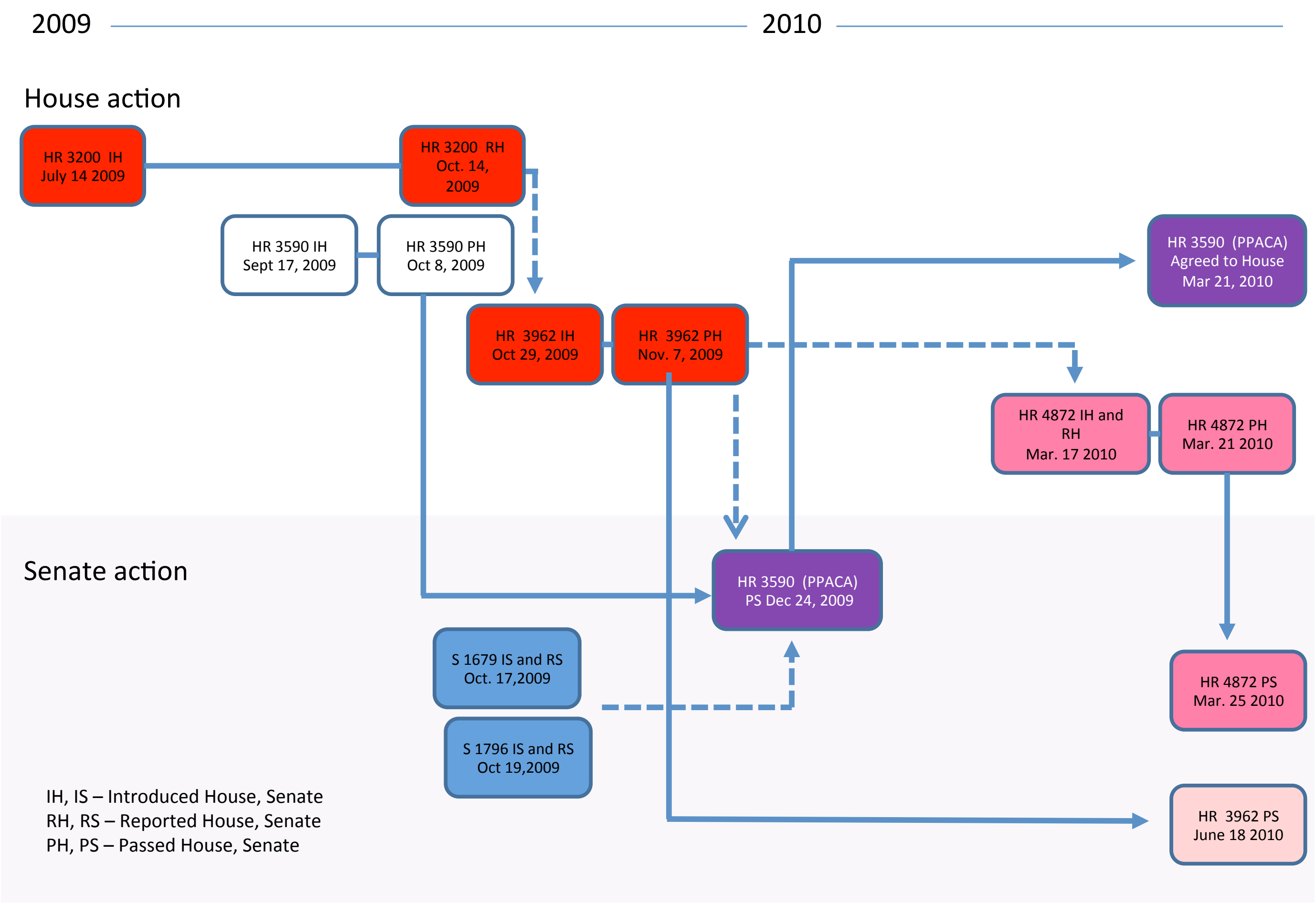
- As introduced in 2009 HR 3590 was six pages long and promoted veteran home ownership. As enacted in 2010, HR 3590 was over 900 pages in length and reformed the US health care system.
- The USA Patriot Act Reauthorization of 2005 contained numerous provisions having nothing to do with terrorism, such as Title VII, the Combat Methamphetamine Epidemic Act, and section 121, modifying the definition of contraband cigarettes.
- A bill proposing a tariff on imported wool trousers, sponsored by

> Democratic Representative (Martin Meehan-MA), died in the House Ways and Means Committee in 1999. Later the same year, Congress passed a law sponsored by the chairman of Ways and Means that included an identical provision.

Bills are best thought of as “vehicles” for policy ideas (Adler & Wilkerson 2012). Many if not most important bills are largely drafted by committee staff under the supervision of the committee or subcommittee chair (Evans 1991, Kaiser 2013). The substance of these bills reflects input from a variety of sources other than the bills named sponsor. Furthermore, bills evolve. The popular view of lawmaking sees this evolution as a policy refinement process the original proposal is scrutinised and edited to address its shortcomings. But bills evolve for other reasons as well. Provisions are added as lawmakers try to address vexing policy problems, engage in logrolling, or capitalize on ”must act” procedural opportunities (Krutz 2001).

The linear “legislative process” perspective taught in civics courses fails to capture the twists and turns of many policy proposals in Congress. The bill that served as the vehicle for the Patient Protection and Affordable Care Act (aka Obamacare) was HR 3590 (Figure 1). However, most of the markup activity related to health care reform occurred in the House over a three month period (July-September 2009) as three committees considered and reported a different bill, HR 3200. But instead of passing this bill, the House passed HR 3962, a bill subsequently introduced by Speaker Pelosi. The Senate never considered HR 3962 however. Its relevant committees reported their own bills (S1679 and S 1796) which also never made it to the floor. Ultimately the Senate took up a bill relating to veterans home ownership passed earlier by the House (HR 3590), stripped its content, and inserted health care reform legislation. The House accepted HR 3590 as amended by the Senate without changes and the PPACA became law. However then House then followed up by passing another bill (HR 4872) that included desired changes that did not make it into HR 3590.

Clearly, approaches that focus on the progress of individual bills miss important dynamics of lawmaking. Our goal is to begin to explore whether it is possible to systematically study the progress of ideas in legislation.



Note: The shaded cells indicate bills substantively related to health care reform. For example, HR 3590 as introduced was about home loans for veterans. After the House passed HR 3590, the replaced the bill's content with its version of the PPACA. Similarly, the Senate used HR 3962, originally the House version of health care reform, as the vehicle for a different set of policies once HR 3590 was enacted.

Figure 1: Bills Providing Major Policy Contributions to Health Care Reform in the 111th Congress

2.1 Operationalizing the Policy Idea

Policies are a legislature's most important products. Yet policy substance has received limited attention in legislative studies.[2] Legislative scholars study ideology, voting patterns, agenda setting and many other topics because of their implications for policy outputs (Shepsle & Weingast 1987, Krehbiel 1998, Poole & Rosenthal 2000, DeGregorio 1999, Cox & McCubbins 2005, Baumgartner et al. 2009). Even research on legislative productivity and policy reversals devote limited attention to policy substance (Mayhew 1991, Krehbiel 1998, Binder 2003, Clinton & Lapinski 2006, Maltzman & Shipan 2008, Berry, Burden & Howell 2010). Although scholars have also long been interested in legislator effectiveness (Volden, Wiseman & Wittmer 2013, Hasecke & Mycoff 2007, Krutz 2005, Anderson, Box-Steffensmeier & Sinclair-Chapman 2003, Ainsworth & Hanson 1996, Schiller 1995), nearly all of their studies gauge effectiveness in terms of bill progress, leading to the not surprising finding that committee chairs and majority party members are by far the most effective legislators. [3] Policy agenda setting research portrays policy change as an event driven process where entrepreneurs capitalize onlegislative "windows of opportunity." Yet little of this research systematically traces how much opportunity different windows afford (Walker 1977, Kingdon 1995, Baumgartner & Jones 1993).

We hope to be able to shed new light on policymaking by tracing the progress

[2] for recent exceptions see (Burstein, Bauldry & Froese 2005, Ryan & Wojcik 2013, Huberty 2013)

[3] Miquel and Snyder (2006) is a noteworthy exception that relies on surveys of effectiveness for the North Carolina legislature.

of policy ideas in legislation. The term ”policy idea” is admittedly amorphous. In this paper, we operationalize it in a limited, technical way. We also do not address the broader question of where ideas come from, an important area of existing research (Schulman 1988, Quirk 1988, Legro 2000). Finally, we limit our attention to tracing ideas within the legislative process, and do not consider the important question of where lawmakers get their ideas.

Our motivation for tracing policy ideas is nicely illustrated in Walker's (1977) description of the development of automobile safety legislation in the 1960s, when two junior senators sponsored legislation that attracted the interest of a senior senator:

> *By the time traffic safety legislation reached the stage of serious formulation and debate in 1966, its original sponsors had been pushed aside by Senators better placed to create a winning coalition. Senators Ribicoff and Gaylord Nelson, both of whom had pressed for the legislation in the early stages, were displaced by Warren Magnuson, the powerful chairman of the Senate Commerce Committee. Under his leadership a legislative victory was achieved against the determined opposition of powerful industrial interests - a result that few Senate insiders would have predicted when debate began on the issue (Walker 1977, 435).*

Such idea recycling seems to be common enough that norms govern when it is acceptable: “Often, members will pick up old pieces of legislation and sponsor them anew, especially if a member has retired or the bill has been gathering dust for years.” But, said one staffer “’[y]ou don’t take someone else’s stuff without asking.”[4]

The challenge is to systematically detect when idea recycling occurs. Burstein et al (2005) use the bill summaries prepared by the Congressional Research Service (CRS) (`http://thomas.loc.gov`) to trace the progress of a pre-selected set of 40 policy proposals across time. They assume that two bills propose the same thing when their summaries are “virtually the same.”

CRS also lists “related” bills deemed to be important sources of substance in a bill or law. Unfortunately, it is not always evident which version of a bill is being related or summarized, given that bills evolve. For the PPACA, half of the related bills are about home mortgages, while the other half are about health care.

Our approach is to focus on legislative text. We assume that two bills share a policy idea when they share similar text. Of course, this raises many questions about whether similar text does actually capture shared policy ideas. This paper constitutes an early cut at the question.

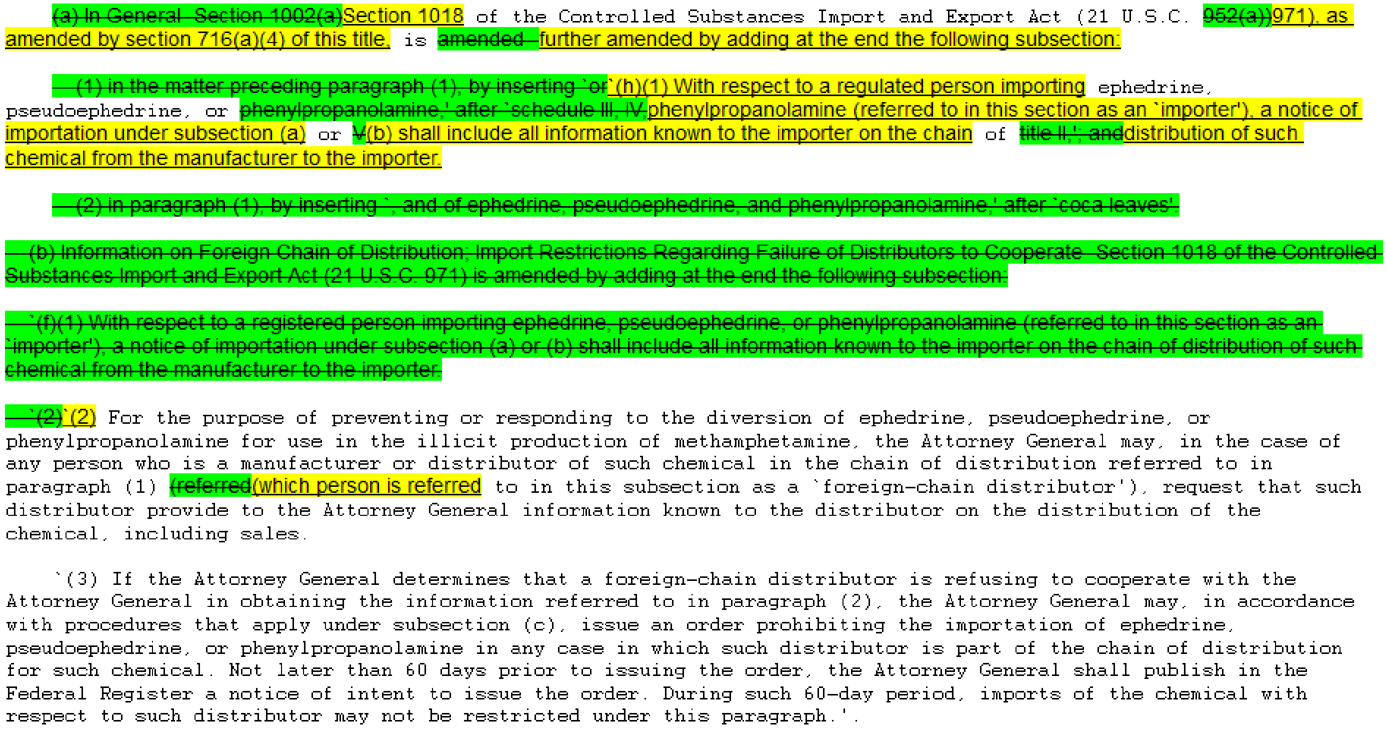
[4] “Joe Walsh Takes Without Asking” `http://www.politico.com/news/stories/0712/79101.html`

2.2 Bill Sections as ideas

Congressional bills have a structure that closely matches the objectives of this project: “almost always, from the earliest days of the Republic, the text of a law, if divided at all, has been divided into sections” (Bellis 2008). According to convention, each bill section “shall contain, as nearly as may be possible, a single proposition of enactment.” Focusing on the bill section makes conceptual sense. It also helps to address the computational challenges associated with comparing the text of so many bills.

Figure 2 illustrates a textual comparison of two sections from two bills. The text in the figure is a merger of Section 2 of the Methamphetamine Precursor Control Act of 2005 (H.R. 1056 - introduced March 2, 2005 by Democratic Representative Darlene Hooley (OR)); and Section 721 of the USA Patriot Act Improvement and Reauthorization of 2005 (HR 3199 - introduced July 11, 2005 by Republican Representative James Sensenbrenner (R-WI)). The highlights indicate text that is unique to one section (the green text is found only in H.R. 1056; the yellow in HR 3199). The non-highlighted text is shared.

In our view the shared text clearly conveys a shared policy idea, giving the Attorney General greater regulatory control over the importation of chemicals used to manufacture meth. This seems to be an example of an idea sponsored by a minority party member being incorporated into a majority-sponsored law. The only other bill containing similar language (HR 3889) was introduced more than three months after H.R. 1056.



~~(a) In General. Section 1002(a)~~Section 1018 of the Controlled Substances Import and Export Act (21 U.S.C. ~~952(a))~~971), as amended by section 716(a)(4) of this title, is ~~amended--~~further amended by adding at the end the following subsection:

~~--(1) in the matter preceding paragraph (1), by inserting `or~~`(h)(1) With respect to a regulated person importing ephedrine, pseudoephedrine, or ~~phenylpropanolamine,' after `schedule III, IV,~~phenylpropanolamine (referred to in this section as an `importer'), a notice of importation under subsection (a) or ~~V~~(b) shall include all information known to the importer on the chain of ~~title II,'; and~~distribution of such chemical from the manufacturer to the importer.

~~--(2) in paragraph (1), by inserting `, and of ephedrine, pseudoephedrine, and phenylpropanolamine,' after `coca leaves'.~~

~~--(b) Information on Foreign Chain of Distribution; Import Restrictions Regarding Failure of Distributors to Cooperate. Section 1018 of the Controlled Substances Import and Export Act (21 U.S.C. 971) is amended by adding at the end the following subsection:~~

~~--`(f)(1) With respect to a registered person importing ephedrine, pseudoephedrine, or phenylpropanolamine (referred to in this section as an `importer'), a notice of importation under subsection (a) or (b) shall include all information known to the importer on the chain of distribution of such chemical from the manufacturer to the importer.~~

~~--`(2)~~`(2) For the purpose of preventing or responding to the diversion of ephedrine, pseudoephedrine, or phenylpropanolamine for use in the illicit production of methamphetamine, the Attorney General may, in the case of any person who is a manufacturer or distributor of such chemical in the chain of distribution referred to in paragraph (1) ~~(referred~~(which person is referred to in this subsection as a `foreign-chain distributor'), request that such distributor provide to the Attorney General information known to the distributor on the distribution of the chemical, including sales.

`(3) If the Attorney General determines that a foreign-chain distributor is refusing to cooperate with the Attorney General in obtaining the information referred to in paragraph (2), the Attorney General may, in accordance with procedures that apply under subsection (c), issue an order prohibiting the importation of ephedrine, pseudoephedrine, or phenylpropanolamine in any case in which such distributor is part of the chain of distribution for such chemical. Not later than 60 days prior to issuing the order, the Attorney General shall publish in the Federal Register a notice of intent to issue the order. During such 60-day period, imports of the chemical with respect to such distributor may not be restricted under this paragraph.'.

Figure 2: Textual Comparison of Different Versions of the Same Policy Idea

Congressional bill texts are digitally available from 1989 to the present, over 100,000 introduced bills along with multiple versions for many of them. Our goal in this paper is to ask whether provisions of the PPACA can be traced to other bills introduced in the 111th Congress (2009-10). We first scraped all versions of all 10,000 plus bills in this Congress,[5] before parsing the substantive sections (excluding other textual elements as short titles, tables of contents, authorization of appropriations etc.). These 119,704 bill sections are the focus of the current paper.

3 A Text Reuse Approach to Tracing Policy Ideas

The Combat Meth Act example illustrates that what is needed is a method that can reliably detect shared policy ideas without requiring exact language matching. As Burstein et al. (2005) succinctly put it: "How similar in content must bills be, to be viewed as manifestations of the same policy proposal? When bills are identical, this is no problem. And when they are very different, it is no problem either" (Burstein, Bauldry & Froese 2005, 296). We draw on a longstanding computer science research agenda examining "text reuse" in documents (Brin, Davis & García-Molina 1995, Büchler et al. 2010). A familiar application is the plagiarism software used to compare students submissions to other documents (Hoad & Zobel 2003) but text reuse methods have broad application.[6]

[5] `http://www.gpo.gov/fdsys/browse/collection.action?collectionCode=BILLS`

[6] They are used in informational retrieval (search algorithms) to identify duplicate queries. Communications scholars use them to study the diffusion of memes. Digital humanities scholars use them to trace references to authors or classic texts in literature. Music scholars use them to investigate rhythmic patterns and (potentially) to identify

Importantly however, whereas plagiarism software is typically used to isolate cases warranting human inspection, we eventually hope to avoid human follow-up by demonstrating that automatically generated similarity scores are predictive of shared policy ideas. This will allow us to study policy ideas in ”big data” fashion.

3.1 Detecting Candidate Section Pairs

A brute-force approach to text reuse detection would entail comparing the text of every bill section to every section of every other bill. For the 10 Congresses we hope to analyze, this would mean more than 500 billion comparisons of sometimes lengthy sections. With collections of this size, achieving the right balance between accuracy and speed is essential. We first describe a more efficient method for reducing candidate section pairs, and then a less computationally expensive method for comparing the substance of those paired sections.

When applied to text, many machine learning algorithms perform quite well with “bag of words” features. The input to the classifier is simply the presence or frequency of each distinct word in the text, regardless of the order in which those words occur. For text reuse detection, more sequence information is generally helpful, but sequence information is also computationally costly.

One compromise solution we explored was a “bag of n-grams” representa-

composers.

tion, which records the count of sequences of 2, 3, 4, etc., characters or words in a section. In a pilot experiment, we computed Dice coefficients (Dice 1945) for each section pair. This measure of similarity is based on the proportion of character bigrams two sections share and required just a day on an ordinary laptop to complete all 10 Congresses. In contrast, a pilot experiment using an off the shelf repeated n-gram plagiarism package (WCopyFind) was estimated to require more than 2,000 hours to complete the 111th Congress on an Amazon EC-2 single instance server, and 30 years to complete all ten.[7]

Algorithmic shortcuts and parallel computing can help to address these computational challenges. We first reduce the number of section pairs to be compared to those that pass a minimal overlap threshold.

We build an inverted index of all the repeated word n-grams in the corpus. We then use the two-pass space-efficient algorithm described by Huston, Moffat —& Croft to filter sections that are unlikely to share substantive content with other sections.(2011) In a first pass, n-grams above a threshold are hashed into a fixed number of bins. On the second pass, n-grams that hash to bins with just one occupant are discarded.[8] For this experiment, we limit consideration to n-grams equal to or greater than 10.

Once we have a list of sections for each distinct n-gram (10 or greater), we

[7] WCopyFind offers corner cutting options pairs) using an off-the-shelf was estimated to to speed up processing, but after selecting these options the 111th Congress job was still running after a week. `http://plagiarism.bloomfieldmedia.com`

[8] Due to hash collisions, there may still be a small number of singleton n-grams that reach this stage. These singletons are filtered out as the index is written.

output all section pairs in each list.[9] We then sort the list of repeated n-grams by document pair to assign a score based on the number of overlapping n-grams. Lastly, we retain only those section pairs that (in the current experiment) contain at least 5 overlapping n-grams.

For the 119,704 bill sections in the 111th Congress corpus, this n-gram indexing process reduced the candidate section pairs from about 7 billion to 1.6 million, or to just 0.02% of the original number.

3.2 Aligning Bill Sections

In terms of how the similarity of bill section pairs might be judged using a repeated n-gram approach, there are two main options. One is to compare the overall similarity of the two sections. The other is to identify and compare the similarity of shared subsequences within the two sections. The former corresponds to a "global" alignment approach(Needleman & Wunsch 1970) where the boundaries of the aligned text are preset. The latter corresponds to a "local" alignment approach where the boundaries are endogenously determined (Smith & Waterman 1981).

We opt for a local alignment approach because it seems best suited to our objective of isolating shared policy ideas. In the meth example of Figure 2, the overall sections of the two bills differ in many respects, including over length, suggesting a low global alignment score even though they contain a

[9] For the current analysis, we suppress repeated n-grams that appear in different sections of the same version of a bill, and n-grams that occur more than 100 times (on the assumption that frequent n-grams are common idiomatic expressions).

shared policy idea.[10] A local alignment approach, in contrast, focuses attention on what the two sections texts share in common - the non-highlighted text that is shared across the two documents. In addition, if a section contains two distinct policy ideas, a local alignment approach is also more likely to treat them as distinct. This allows for the detection of instances where text from two or more sections of an earlier bill ends up in a single section of a later bill (or vice versa). The appendix to this paper offers a detailed description of the Smith Waterman local alignment algorith and how SW alignment $SWalign$ scores are computed.

Table 1 provides an example of a local alignment capturing a policy idea. The dashes in each text indicate a relatively small number of character mismatches or gaps (places where one of the texts contains additional characters) within the aligned portions of the two bill sections. The text on the right comes from the enrolled version of the PPACA (HR 3590). The text on the left comes from a bill that never made it out of committee (S. 1244, The Breastfeeding Promotion Act of 2009, sponsored by Jeff Merkley (D-OR)). In this case the aligned text spanned just .59 and .55 of the respective sections.

To speed up the process of aligning 1.6 million section pairs, we rely on "massive parallelism." On a five-year-old cluster of commodity servers, indexing repeated 10-grams for the 111th Congress took just 10 minutes with 12-fold parallelism, detecting the 1.6 million candidate section pairs took 23 minutes with 8-fold parallelism, and performing Smith-Waterman alignment

[10] the Patriot Act section is also much larger than shown

ing mothers a in general section 7 of the fair labor standards act—— 29 usc 207 is amended by adding at the end the following r 1 an employer shall provide—- reasonable break time for an employee to express breast milk for her nursing child for 1 year after the childs birth each time such employee has need to express the milk the employer shall make reasonable efforts to provide a place other than a bathroom that is shielded from view and free from intrusion from coworkers and the public which may be used by an employee to express breast milk– an employer shall not be required to compensate an employee—————————————— for any work time spent for such purpose 2 for purposes of this subsection the term employer means an employ	ing mothers————- section 7 of the fair labor standards act of 1938 29 usc 207 is amended by adding at the end the following r 1 an employer shall provide a a reasonable break time for an employee to express breast milk for her nursing child for 1 year after the childs birth each time such employee has need to express the milk and ———————————————b a place other than a bathroom that is shielded from view and free from intrusion from coworkers and the public which may be used by an employee to express breast milk 2 an employer shall not be required to compensate an employee receiving reasonable break time under paragraph 1 for any work time spent for such purpose 3 ——————————an employer —-that employ

Table 1: A Local Alignment Example with a passage from S 1244 (Breastfeeding Promotion Act of 2009) on the left and the PPACA on the right.

on these sections took 28 minutes with 50-fold parallelism.

3.3 Building the Train/Test Corpus

We are interested in whether SW alignment scores can be used to systematically predict the presence of shared policy ideas. As is standard practice in supervised machine learning research, we construct a "gold standard" human-labeled dataset to test prediction performance. Our train/test corpus includes 3400 of the 19,241 alignments related to the PPACA, drawn from alignments that make up at least 7 percent of one or both sections (these are the top 50 percent of cases). The corpus also includes the *SWalign* score for each section pair, the aligned texts, the differences between the aligned texts, and details about each bill (such as date of introduction, sponsor characteristics, committee of referral, etc).[11]

There are many potential reasons why objectively similar passages may be either substantively unimportant or different. As we began to code, it became clear that bills share a lot of language about mundane but necessary things, such as defining terms, establishing effective dates, creating commissions, calling for reports, and adjusting for inflation, etc. We call this non-policy substance ”boilerplate.” Bill sections also often share similar short snippets of text such as preambles that lack policy substance. Finally, two aligned texts can differ in small but critical respects: both might propose similar medical education programs but for different professions (e.g. pediatrics vs dentistry).

Of interest is whether it is possible to systematically reduce false positives and false negatives to the point where SWalign scores accurately predict the presence of a shared policy idea. We (two of the co-authors and a graduate research assistant) classified each alignment in the sample to only one of

[11] This ancillary information was concealed during the process of coding for shared policy ideas.

6 categories (see figure 3 below).[12] Most of the alignments we classified were boilerplate (category 5) or preambles and incomplete snippets (junk) (category 6). Our working definition of a policy idea is based on human judgment: The alignment must include a comprehensible description of a policy objective.[13] Instances of shared policy ideas (categories 1 and 2) make up about 16 percent of the sample, while cases where the aligned passages address different policy ideas make up slightly less. Finally, it is important to note that we are labeling alignments, not overall bill sections.

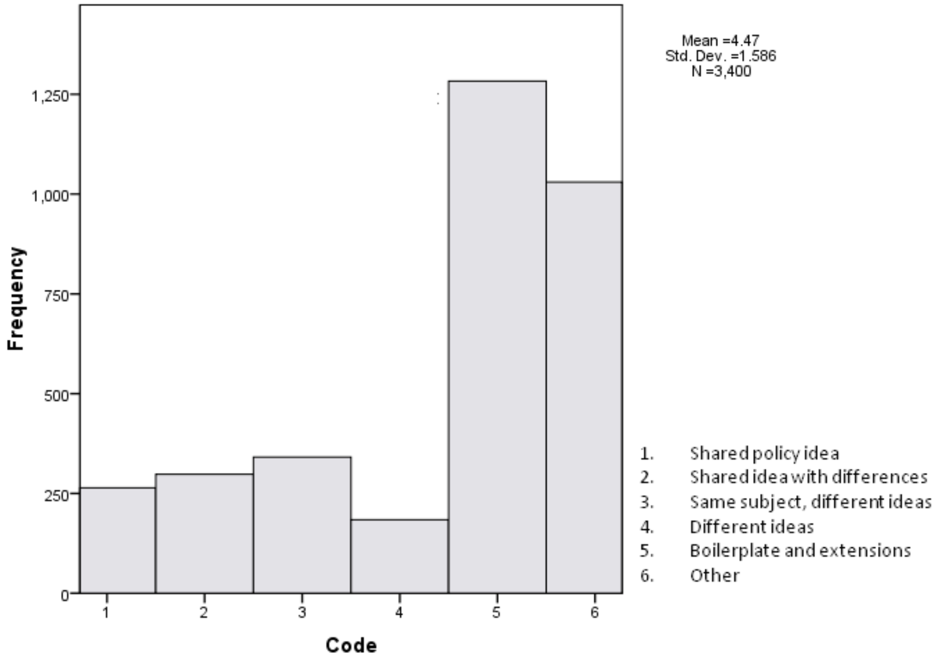


Figure 3: Histogram of Human-labeled Alignments by Category

[12] Across the 6 categories interrater reliability (3 coders) was approximately 75 percent. Our primary goal is to differentiate cases of shared policy ideas (categories 1 and 2) from other cases (categories 3-6). For this task, interrater reliability was above 90 percent.

[13] This definition almost certainly omits some policy changes, such as those that simply delete existing statutory language (e.g."section 824 g of the foreign service act of 1980 22 usc 4064 g is amended in paragraph 1b by striking to facilitate the and all that follows."

3.4 Predicting Shared Policy Ideas

We anticipate that the main challenge is filtering false positives - cases with high SWalign scores that are not policy ideas. We first train a supervised machine learning algorithm (SVM) to predict boilerplate cases using the 3400 human labeled cases (Joachims 2002). Next we partition the human-labeled cases into those containing shared policy ideas (1,2) and those that don't (3,4,6). We then predict shared policy ideas via logistic regression using n-fold cross validation (including and then excluding predicted boilerplate cases). The only independent variable in the logistic regression is the Smith-Waterman alignment score.

Table 2 reports the results of three experiments. The last set of results are for the most complete method that first uses SVM to exclude predicted boilerplate cases. Overall accuracy is 92 percent.[14] Recall is 97.3 percent, which means that true shared policy ideas are missed only 2.7 percent of the time. Precision is lower, but false negatives are of less concern because it is much easier to review the much smaller number of predicted cases of shared ideas.[15]

[14] Because the sample is biased in favor of higher alignment scores, actual accuracy will be higher.

[15] Many of these false positives are boilerplate cases that the initial SVM method failed to detect. More training examples should help to address this.

	Point Estimate	**95% Confidence Interval**
Predicting Boilerplate language (SVM):		
Percent Correctly Predicted	85.6	(82.4-88.2)
Precision:	76.5	(70.8-82.3)
Recall:	91.1	(87.8-94.0)
Predicting Shared Policy Ideas (Logistic Regression):		
Percent Correctly Predicted:	87.4	(84.6-89.8)
Precision:	69.2	(59.0-78.4)
Recall:	90.9	(87.9-93.5)
Predicting Shared Policy Ideas Excluding Predicted Boilerplate (SVM and Logistic Reg.):		
Percent Correctly Predicted:	92.0	(89.6-94)
Precision:	65.0	(55.1-74.3)
Recall:	97.3	(95.4-98.8)

Table 2: Results from N-fold cross validation (2900 train, 500 test)

4 Findings: Tracing the Policy Development of the PPACA

In this section we use the methodological approach described above to identify policy ideas in the enacted version of the Patient Protection and Affordable Care Act that are also found in other bills introduced earlier. Applying the method to the 19,241 alignments related to PPACA resulted in 1081 predicted cases of shared policy ideas. We then reviewed and excluded false positives, resulting in a final dataset of 1023 shared policy ideas.

As discussed and illustrated in Figure 1, most of the work on health care reform in the House centered on two other bills (HR 3200 and HR 3962), while the Senate had its own two markup bills (S1679 and S1796). We would therefore expect that a substantial number of policy ideas in the PPACA are also found in these four bills. We know less about which of these bills has the most in common with the final version of the PPACA. And we have no information at this point about whether provisions of the PPACA can be traced to other bills. We are particularly interested in whether Republican-sponsored ideas made their way into what has been described as a highly partisan law. Put another way, how ”inclusive” was the process?

A total of 166 different bills introduced in the 111th Congress contain policy ideas also present in the final PPACA (Figure 4). Each dot is a bill, where size is the log of the number of matching bill sections. The four large blue dots are the two House and two Senate bills broadly recognized as precursors to HR 3590 (S 1976 appears to be the most similar to the PPACA as enacted). The smaller dots indicate other bills, where red dots are bills sponsored by Republicans. Overall, one-fourth of the policy ideas linked to the PPACA are found in bills introduced before the main markup bills, and many of these bills were Republican-sponsored.[16]

[16] Only four of these 166 bills became law on their own and all were Democratically sponsored (not shown).

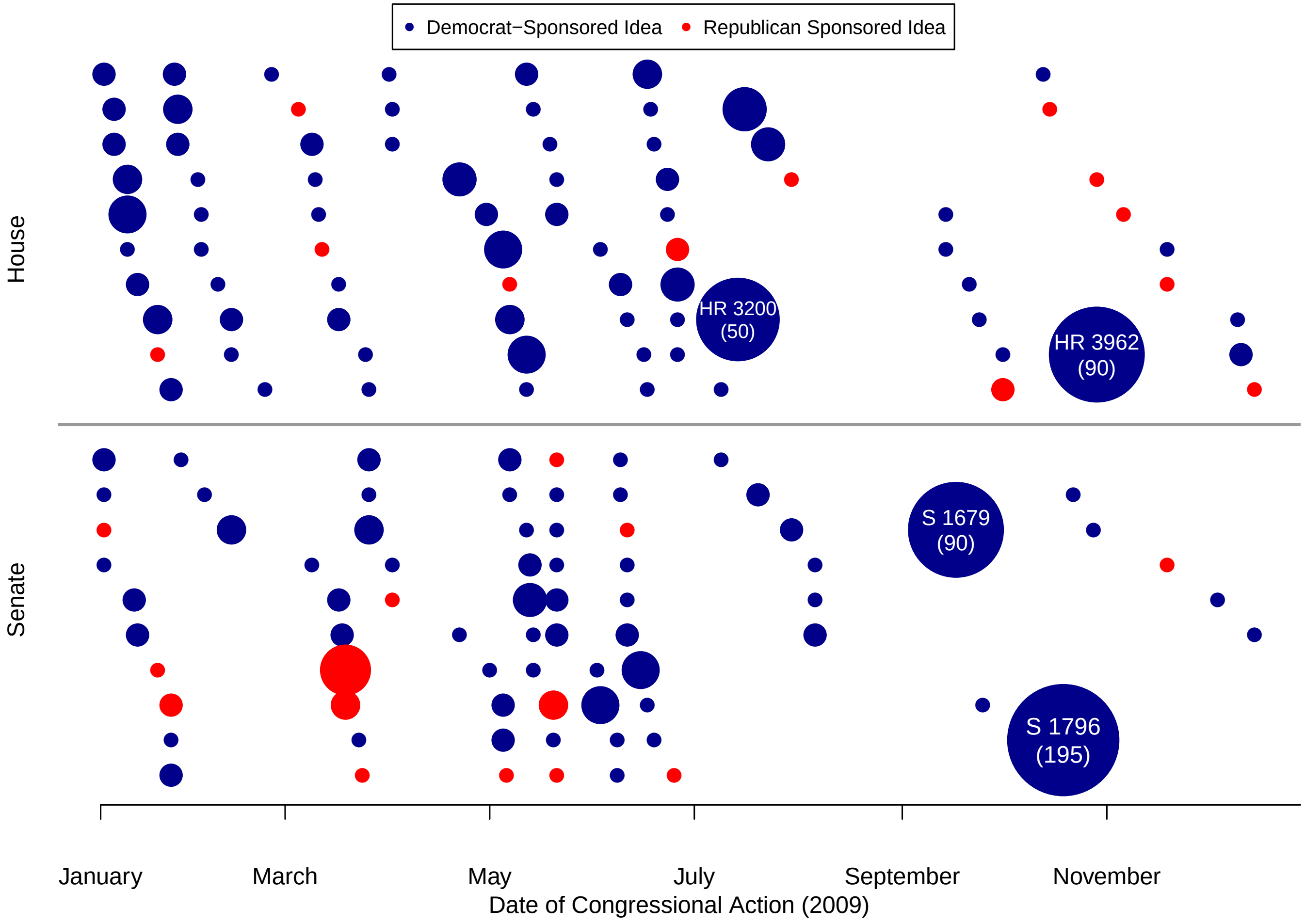


Figure 4: Dates of Introduction for PPACA Ideas. Sized by number of alignments in each bill. Bills in the known history of the PPACA are labeled by bill number and number of alignments.

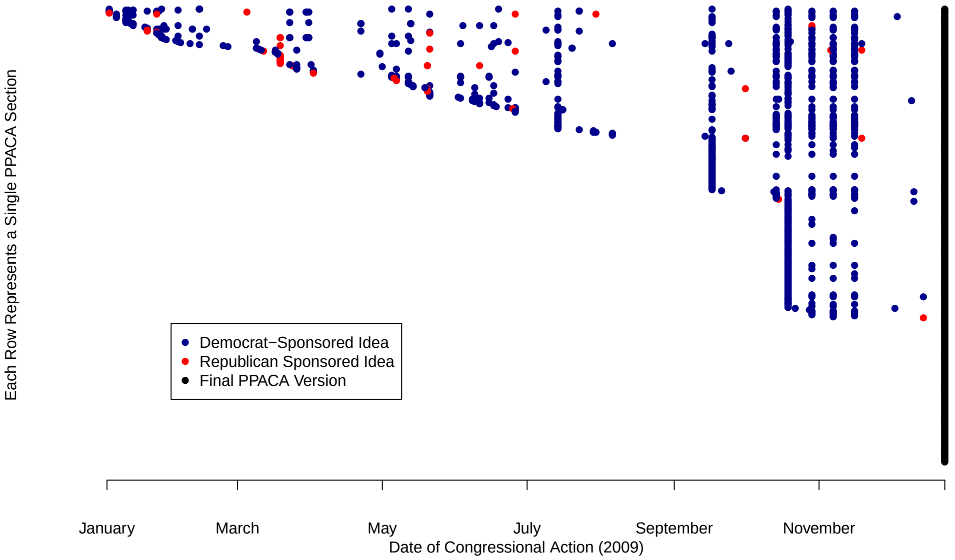


Figure 5: Traces of PPACA Sections. Each substantive ACA section represented by one row.

Figure 5 provides a a different perspective on the chronological development of health care reform. The black line on the right corresponds to the total number of sections in the law. Each row of the figure is a section of the PPACA and each column is a point in time. At the top are PPACA sections that include alignments that can be traced chronologically to a bill introduced near the beginning of the 111th Congress. Further down are sections that have more recent origins, followed by sections that cannot be traced to any previous bills.[17] What the figures appear to show, as we hoped to find at the outset, is that policy development is a collective evolutionary process. Even highly partisan bills appear to reflect input from a broad spectrum of lawmakers.

[17] many of these sections may not include substantive language (they may be boilerplate).

4.1 Inclusiveness

Scholars rely almost exclusively on floor roll call voting patterns to assess partisan cooperation in Congress. By this assessment, the PPACA was a highly partisan bill. No Republican in the House or Senate supported it. One interpretation of this outcome might be that Republicans were excluded from the process. A policy ideas approach provides an opportunity to revisit bipartisan cooperation at the bill construction level. Specifically, is lawmaking is more inclusive when judged in terms of the progress of policy ideas, instead of in terms of roll call voting patterns or bill success?

Table 3 summarizes information about the sponsors of policy ideas later found in the PPACA.[18] The first table includes the Democratic sponsors of the 4 major markup bills. The second table excludes those sponsors. While there is limited evidence of inclusiveness in the House, there is substantial evidence in the Senate. More than one-fourth of the linked ideas are found in bills sponsored by Republicans. For example, an entire subtitle (B) relating to nursing home fraud prevention in the PPACA appears to be drawn primarily from a bill (S.647) sponsored by Charles Grassley (R-IA) much earlier in the process. Because the Congressional Service does not designate the Grassley bill as ”related” to the PPACA, this is a case of Republican influence that would otherwise be invisible.

[18] the tables focus on only the clearest cases of shared policy ideas (category 1).

All aligned sections

	House	Senate
Minority	2.8%	8.0%
Majority	97.9%	92.2%
N	468	376

Excluding markup bills

	House	Senate
Minority	10.7%	27.8%
Majority	89.3%	72.2%
N	122	108

Table 3: ”Inclusiveness” by Chamber and Party Status

5 Discussion

This paper proposes a text reuse approach to tracing the movement of policy ideas through the legislative process. Using the case of the voluminous 906 page Affordable Care Act (HR 3590), we developed and tested whether a ‘local alignment’ algorithmic approach can predict shared policy ideas. Using a human-labeled ‘gold standard’ set of 3400 alignments, we were able to predict known cases with 92 percent accuracy and 97.3 percent recall using a combination of supervised machine learning and logistic regression. The same modeling strategy applied to the broader dataset of 19,241 alignments detected 1080 instances of ideas found in the PPACA that could be traced to bills introduced at earlier points in time. We learned substantially more about the history of the PPACA than the alternative approach of relying on the ‘related bills’ designated by the Congressional Research Service.

Important limitations of this paper should be noted. We have operationalized the policy idea in a technical way - ideas must be shared nearly verbatim. We do not know the extent to which our local alignment captures all of the policy ideas shared between the PPACA and other bills. We have also not demonstrated that a particular bill provided the inspiration for a provision found in the PPACA, nor have we attempted to trace the specific events that led to a provision's incorporation into the PPACA. Sometimes ideas found in laws are only found in one other bill. At other times, the same idea is found in multiple bills. At what stage in the evolution of the PPACA did the provision get added, how and why? Are multiple occurrences evidence of an uptake process or mere mimicking?

The current paper is also limited to a specific law at a particular point in time. Over the longer term one of our goals is to use this method to reevaluate existing understandings of lawmaking that are currently based primarily on voting patterns and the progress of bills. A text reuse approach offers a window into policymaking that allows us to explore the markup process in far greater detail. We anticipate that policymaking will be found to be much more inclusive when assessed in terms of the progress of ideas. We also look forward to assessing inclusiveness across committees, members, issue areas, and time.

Although there is substantial work to be done, text reuse methods have already taught us quite a bit about the evolution of one important law. We look forward to comments and suggestions as we move forward.

6 Appendix: Computing local alignments

The Smith-Waterman algorithm employs dynamic programming to reuse calculations when comparing all possible subsequences of the two input documents.

In our case, two sections would be treated as sequences of text X and Y whose individual characters are indexed as X_i and Y_j. Let $W(X_i, Y_j)$ be the score of aligning character X_i to character Y_j. Higher scores are better. We use a scoring function where only exact character matches get a positive score and any other pair gets a negative score. We also account for additional text appearing on either X or Y. Let W_g be the score, which is negative, of starting a "gap", where one sequence includes text not in the other. Let W_c be the cost for continuing a gap for one more character. This "affine gap" model assigns a lower cost to continuing a gap than to starting one, which has the effect of making the gaps more contiguous. We use an assignment of weights fairly standard in genetic sequences where matching characters score 2, mismatched characters score -1, beginning a gap costs -5, and continuing a gap costs -0.5. We leave for future work the optimization of these weights for the task of capturing shared policy ideas.

As with other dynamic programming algorithms such as Levenshtein distance, the Smith-Waterman algorithm operates by filling in a "chart" of partial results. The chart in this case is a set of cells indexed by the characters in X and Y, and we initialize it as follows:

$$H(0,0) = 0$$

$$H(i,0) = E(i,0) = W_g + i \cdot W_c$$

$$H(0,j) = F(0,j) = W_g + j \cdot W_c$$

The algorithm is then defined by the following recurrence relations:

$$H(i,j) = \max \begin{cases} 0 \\ E(i,j) \\ F(i,j) \\ H(i-1,j-1) + W(X_i, Y_j) \end{cases}$$

$$E(i,j) = \max \begin{cases} E(i,j-1) + W_c \\ H(i,j-1) + W_g + W_c \end{cases}$$

$$F(i,j) = \max \begin{cases} F(i-1,j) + W_c \\ H(i-1,j) + W_g + W_c \end{cases}$$

The main entry in each cell $H(i,j)$ represents the score of the best alignment that terminates at position i and j in each sequence. The intermediate quantities E and F are used for evaluating gaps. Due to taking a max with 0, $H(i,j)$ cannot be negative. This is what allows Smith-Waterman to ignore text before and after the locally aligned substrings of each input.

After completing the chart, we then find the optimum alignment by tracing back from the cell with the highest cumulative value $H(i, j)$ until a cell with a value of 0 is reached. These two cells represent the bounds of the sequence, and the overall SW alignment score reflects the extent to which the characters in the sequences align and the overall length of the sequence.[19]

In our implementation, we include one further speedup: since in a previous step we identified n-grams that are shared between the two bill sections, we assume that any alignment of those sections must include those n-grams as matches. In some cases, this anchoring of the alignment might lead to suboptimal SW alignment scores.

[19] see also `http://www.cs.kent.edu/ssteinfa/files/PDSEC08_handouts.pdf`

References

Adler, E. Scott & John D. Wilkerson. 2012. *Congress and the Politics of Problem Solving.* London: Cambridge University Press.

Ainsworth, Scott & Douglas Hanson. 1996. "Bill sponsorship and legislative success among freshmen senators, 1954–1986." *The Social Science Journal* 33(2):211–221.

Anderson, William J., Jane M. Box-Steffensmeier & Valeria Sinclair-Chapman. 2003. "The Keys to Legislative Success in the U.S. House of Representatives." *Legislative Studies Quarterly* 28(3):357–386.

Baumgartner, F.R., C. Breunig, C. Green Pedersen, B.D. Jones, P.B. Mortensen, M. Nuytemans & S. Walgrave. 2009. "Punctuated equilibrium in comparative perspective." *American Journal of Political Science* 53(3):603–620.

Baumgartner, Frank & Bryan Jones. 1993. *Agendas and Instability in American Politics.* Chicago: University of Chicago Press.

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

Berry, Christopher R., Barry C. Burden & William G. Howell. 2010. "After Enactment: The Lives and Deaths of Federal Programs." *American Journal of Political Science* 54(1).

Binder, Sarah. 2003. *Stalemate: Causes and Consequences of Legislative Gridlock.* Washington, D.C.: Brookings Institution Press.

Brin, Sergey, James Davis & Héctor García-Molina. 1995. "Copy detection mechanisms for digital documents." *SIGMOD Rec.* 24(2):398–409.

Büchler, Marco, Annette Geßner, Gerhard Heyer & Thomas Eckart. 2010. Detection of Citations and Textual Reuse on Ancient Greek Texts and its Applications in the Classical Studies: eAQUA Project. In *Proceedings of the Digital Humanities Conference 2010.*

Burstein, Paul, Shawn Bauldry & Paul Froese. 2005. "Bill Sponsorship and Congressional Support for Policy Proposals, from Introduction to Enactment or Disappearance." *Political Research Quarterly* 58(2):295–302.

Clinton, Joshua D. & John S. Lapinski. 2006. "Measuring Legislative Accomplishment, 18771994." *American Journal of Political Science* 50(1):232–249.

Cox, Gary W. & Matthew McCubbins. 2005. *Setting the Agenda: Responsible Party Government in the U.S. House of Representatives.* New York: Cambridge University Press.

DeGregorio, Christine A. 1999. *Networks of Champions: Leadership, Access, and Advocacy in the U.S. House of Representatives.* Ann Arbor, MI: University of Michigan Press.

Dice, Lee R. 1945. "Measures of the Amount of Ecologic Association Between Species." *Ecology* 26(3):pp. 297–302.
URL: *http://www.jstor.org/stable/1932409*

Evans, C. Lawrence. 1991. *Leadership in Committee: A Comparative Analysis of Leadership Behavior in the U.S. Senate.* Ann Arbor, MI: University of Michigan Press.

Hasecke, Edward B. & Jason D. Mycoff. 2007. "Party Loyalty and Legislative Success: Are Loyal Majority Party Members More Successful in the U.S. House of Representatives?" *Political Research Quarterly* 60(4):607–617.

Hoad, Timothy C. & Justin Zobel. 2003. "Methods for identifying versioned and plagiarized documents." *J. Am. Soc. Inf. Sci. Technol.* 54(3):203–215.
URL: *http://dx.doi.org/10.1002/asi.10170*

Huberty, Mark. 2013. Applying Natural Language Processing for Computer Assisted Analysis of Legislative History: The LegHist Package for R. In *unpublished paper.* Presented at the 2013 Meetings of the American Political Science Association, August 2–Sept. 1, Chicago IL.

Huston, Samuel, Alistair Moffat & 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.

Joachims, T. 2002. *Learning to Classify Text using Support Vector Machines.* Kluwer/Springer.

Kaiser, Robert. 2013. *Act of Congress: How America's Essential Institution Works, and How It Doesn't.* New York: Knopf.

Kingdon, John. 1995. *Agendas, Alternative, and Public Policies.* New York: Harper Collins.

Krehbiel, Keith. 1998. *Pivotal Politics: A Theory of U.S. Lawmaking.* Chicago: University of Chicago Press.

Krutz, Glen. 2001. *Hitching a Ride: Omnibus Legislating in the U.S. Congress.* Columbus, OH: Ohio State University Press.

Krutz, Glen S. 2005. "Issues and Institutions: Winnowing in the U.S. Congress." *American Journal of Political Science* 49(2):313–326.
URL: *http://dx.doi.org/10.1111/j.0092-5853.2005.00125.x*

Legro, Jeffrey W. 2000. "The Transformation of Policy Ideas." *American Journal of Political Science* 44(3):pp. 419–432.

Maltzman, Forrest & Charles R. Shipan. 2008. "Change, Continuity, and the Evolution of the Law." *American Journal of Political Science* 52(2):252–267.

Mayhew, David. 1991. *Divided We Govern: Party Control, Lawmaking, and Investigations, 1946-1990.* New Haven: Yale University Press.

Miquel, Gerard Padro I & James M Snyder. 2006. "Legislative effectiveness and legislative careers." *Legislative Studies Quarterly* 31(3):347–381.

Needleman, Saul B. & Christian D. Wunsch. 1970. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." *Journal of Molecular Biology* 48(3):443–453.
URL: *http://dx.doi.org/10.1016/0022-2836(70)90057-4*

Poole, Keith T & Howard Rosenthal. 2000. *Congress: A political-economic history of roll call voting.* Oxford University Press.

Quirk, Paul J. 1988. "In Defense of the Politics of Ideas." *The Journal of Politics* 50(1):pp. 31–41.

Ryan, Josh, Anand Sokhey Gilad Wilkenfeld & Stefan Wojcik. 2013. Tracing the Legislative Process: A Network Approach. In *unpublished manuscript.* Prepared forthe Legislative Networks in Transatlantic Perspectives Workshop, April 15, 2013, Madison, WI.

Schiller, Wendy J. 1995. “Senators as Political Entrepreneurs: Using Bill Sponsorship to Shape Legislative Agendas.” *American Journal of Political Science* 39(1):pp. 186–203.

Schulman, Paul R. 1988. “The politics of ideational policy.” *Journal of Politics* 50(2):263–291.

Shepsle, Kenneth A. & Barry R. Weingast. 1987. “The Institutional Foundations of Committee Power.” *The American Political Science Review* 81(1):pp. 85–104.

Smith, T. F. & 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*

Volden, Craig, Alan E. Wiseman & Dana E. Wittmer. 2013. “When Are Women More Effective Lawmakers Than Men?” *American Journal of Political Science* 57(2):326–341.
URL: *http://dx.doi.org/10.1111/ajps.12010*

Walker, Jack L. 1977. “Setting the Agenda in the U.S. Senate: A Theory of Problem Selection.” *British Journal of Political Science* 7(4):pp. 423–445.
URL: *http://www.jstor.org/stable/193298*