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

John Wilkerson University of Washington
David Smith Northeastern University
Nicholas Stramp University of Washington

This article proposes a new approach to investigating the substance of lawmaking. Only a very small proportion of bills become law in the U.S. Congress. However, the bills that do become law often serve as vehicles for language originating in other bills. We investigate "text reuse" methods as a means for tracing the progress of policy ideas in legislation. We then show how a focus on policy ideas leads to new insights into the lawmaking process. Although our focus is on relating content found within bills, the same methods can be used to study policy substance across many research domains.

An irony of the Patient Protection and Affordable Care Act (PL 111-148) is that one of its key provisions, the individual insurance mandate, has conservative origins.[1] In Congress, the requirement that individuals 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 reform passed with the support of then 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."

Like many laws, the 906-page Patient Protection and Affordable Care Act (PPACA, or Obamacare) is a product of inputs from many sources. Yet systematic approaches to tracing how laws develop are virtually nonexistent. We propose a shift from the traditional research focus on the progress of bills to what most scholars ultimately care about—the progress of policy ideas.

We provide a definition of policy ideas in legislation and use genetic sequencing methods to discover when two bills share a policy idea. We investigate the PPACA's legislative history by comparing its final provisions to the content of more than 29,000 bill versions published in the 111th Congress (2009–10). We find that the law includes many provisions originally advanced in other (failed) bills, including bills sponsored by minority Republicans. Turning to the 111th Congress as a whole, we observe similar patterns, as well as important variations across issues and lawmakers. We conclude that moving beyond the current focus on bills to investigate the progress of policy ideas more directly is feasible.

From Bills to Policy Ideas

The conventional approach to legislative history for both government librarians and research scholars is to trace the progress of individual bills as they move through the legislative process. Such an approach makes sense when researchers care about bill progress. It makes less sense when the goal is to understand how policies progress in

John Wilkerson is Professor, Department of Political Science, University of Washington, Seattle, WA 98195-353530 (jwilker@u.washington.edu). David Smith is Assistant Professor, College of Computer and Information Science, Northeastern University, 440 Huntington Avenue, West Village H, Room 356, Boston, MA 02115 (dasmith@ccs.neu.edu). Nicholas Stramp is a Ph.D. Student, Department of Political Science, University of Washington, Box 353530, Seattle, WA 98195 (stramp@u.washington.edu).

The authors thank the reviewers, Kevin Shotwell, and colleagues at talks at the University of Washington, Northeastern University, the University of Pennsylvania, the Hertie School of Governance, and the New Directions in Analyzing Text as Data and Comparative Agendas Project workshops. This article is based upon work supported by the National Science Foundation under Grant No. 1224173. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation. Replication files are available on DataVerse (http://dx.doi.org/10.7910/DVN/27584).

[1] "The Tortuous History of Conservatives and the Individual Mandate," http://www.forbes.com/sites/aroy/2012/02/07/the-tortuous-conservative-history-of-the-individual-mandate.

American Journal of Political Science, Vol. 59, No. 4, October 2015, Pp. 943–956

DOI: 10.1111/ajps.12175

Congress, particularly in recent years. Congress is passing fewer laws, but more "omnibus" laws that pull together ideas from many different sources (Krutz 2001; Sinclair 2011).

Omnibus lawmaking suggests that "the history of any legislation is more likely to be a tapestry of many histories woven together than a single thread" (Cannan 2013, 135). Idea borrowing appears to happen a lot in Congress—so much so that norms dictate that members should ask permission when they borrow ideas from other sitting members.[2] In the 111th Congress, 62% of bills were longer at enactment compared to when they were introduced. The average law was two to four times longer.[3] It is also not a new phenomenon. Examining the development of the National Traffic and Motor Vehicle Safety Act (NTMVSA) in the 1960s, political scientist Jack Walker noted that some of its best ideas originated with other bills (and bill sponsors):

> 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.(Walker 1977, 435)

The history of the NTMVSA should probably include the (failed) bills of Ribicoff and Nelson. Similarly, assessments of the effectiveness of these two lawmakers should probably consider not just the bills they sponsored that became law, but also the policy ideas they advanced that became law as provisions in bills sponsored by other members. Yet scholars know little about how bills evolve between introduction and enactment.

"Obamacare" is another excellent example. The reform is actually two laws with little legislative history in themselves (HR 3590 and HR 4872). Most of the action centered on other bills that did not become law. A more complete discussion of the history of this issue would include several "markup" bills (Figure 1). In the House, three committees considered and reported versions of HR 3200, sponsored by House Energy and Commerce Committee Chair John Dingell (D-MI). Dingell later introduced a new bill reflecting informal negotiations with Speaker Nancy Pelosi and others (HR 3962). This was the health care reform bill the House sent over to the Senate. Two Senate committees also reported major reform bills, S 1679 and S 1796. But instead of taking up one of these bills or the House bill (HR 3962), Majority Leader Harry Reid (D-NV) proposed comprehensive health care reform as a substitute amendment to HR 3590 (a six-page bill proposing mortgage subsidies for service members turned into a 906-page bill that had nothing to do with mortgages). The House then passed HR 3590 without additional changes (to avoid having to return HR 3590 to the Senate and an expected filibuster). Three months later, the House and Senate made 55 pages of additional changes via a budget reconciliation bill, HR 4872, that could not be filibustered.[4]

Operationalizing the Policy Idea

The tangled history of the PPACA underscores the need for a different approach to studying how laws are constructed. Prior efforts to link bills based on their policy substance have relied centrally on the judgment of experts. In a very original study, Burstein, Bauldry, and Froese (2005) trace the progress of 40 policy proposals over several congresses using expert-prepared bill summaries from the Congressional Research Service (CRS).[5] They assume that bills proposed in different congresses are the same if their CRS summaries are "virtually the same."

This method will not work for identifying linkages when a bill (e.g., the PPACA) shares only partial content with other bills. CRS offers another option, the "related" bill designation, but cautions that "although every attempt is made to identify related measures, it is not always possible that all related measures will be captured because of the complexity of such relationships."[6]Our research confirms that the CRS "related" bill designation is not only incomplete but also unreliable. For example, Congress regularly passes omnibus miscellaneous tariff bills (MTBs) that aggregate hundreds of temporary duty

[2] "Joe Walsh Takes without Asking," http://www.politico.com/news/stories/0712/79101.html.

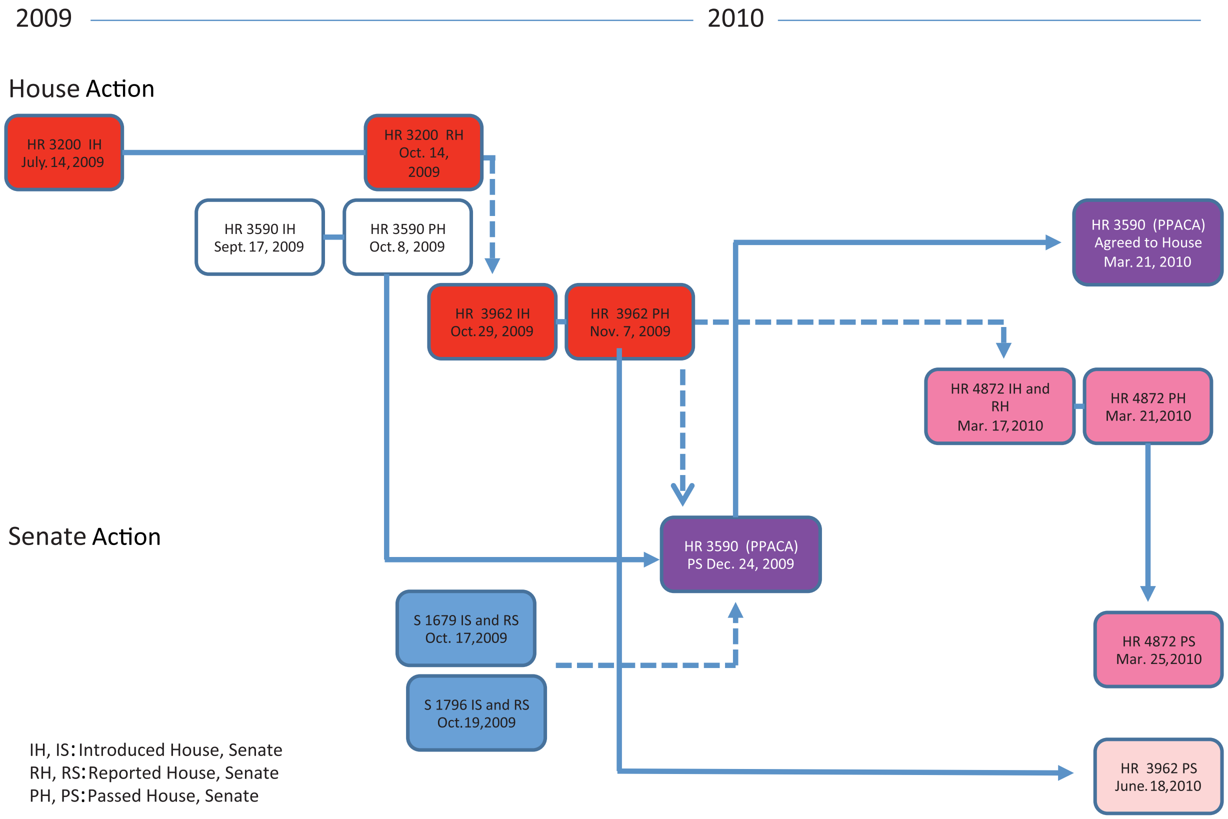
[3] This is based on the size of the text introduced and enrolled versions (there was no introduced version for 58 of 383 laws). The average bill was four times bigger by the time it became law, excluding minor laws (e.g., building namings, land transfers, and commemorative coin issuances) and appropriations. Excluding a small number of outlier omnibus bills that are hundreds of times larger (e.g., the PPACA and Dodd-Frank), the average bill is more than two times longer at the end of the process.

[4] This sequence of events was prompted by the Democrats' loss of their 60-vote majority in the Senate with the special election of Senator Scott Brown (R-MA).

[5] Made available via the Library of Congress's THOMAS (http://thomas.loc.gov) and Congress.gov websites (http://congress.gov).

[6] http://thomas.loc.gov/bss/abt˙related.html.

FIGURE 1 **Bills Providing Major Policy Contributions to Health Care Reform in the 111th Congress**


2009
2010
House Action
HR 3200 IH July. 14, 2009
HR 3200 RH Oct. 14, 2009
HR 3590 IH Sept. 17, 2009
HR 3590 PH Oct. 8, 2009
HR 3962 IH Oct. 29, 2009
HR 3962 PH Nov. 7, 2009
HR 3590 (PPACA) Agreed to House Mar. 21, 2010
HR 4872 IH and RH Mar. 17, 2010
HR 4872 PH Mar. 21,2010
Senate Action
HR 3590 (PPACA) PS Dec. 24, 2009
S 1679 IS and RS Oct. 17,2009
S 1796 IS and RS Oct.19,2009
HR 4872 PS Mar. 25,2010
HR 3962 PS June. 18,2010
IH, IS: Introduced House, Senate
RH, RS: Reported House, Senate
PH, PS: Passed House, Senate

Note: The shaded cells indicate bills substantively related to health care reform. For example, HR 3590 as introduced was about home loans for veterans. AVer the House passed HR 3590, the bill's content had been replace 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.

suspension bills. The committees involved typically issue a report cross-referencing the individual bills with specific provisions of the law. Yet the last time CRS related a significant number of individual duty suspension bills to an MTB on the THOMAS website was in 1990.[7] For the most recent MTB law (HR 4380, 111th Congress), no related bills are indicated. CRS also does not appear to update related bills as the substance of a bill evolves. Two of the bills that are officially related to the PPACA (as of August 2014) propose mortgage credits for service members, whereas the law itself contains no such language.

Our approach focuses on the language of legislation. We consider two bills to be related if they share a policy idea—an admittedly ambiguous concept. For some, policy idea refers to a general policy objective (e.g., universal health care), whereas for others it refers to specific policy provisions in laws. The policy ideas we have in mind specify, in statutory language, what governments, private entities, or citizens can (or cannot) do. A policy idea in legislation is a conferral of substantive legal authority. The text below provides an example of a conferral of authority—in this case, mandating that large employers "shall provide reasonable break time" or face penalties.

> **Section 501. Privacy For Breastfeeding Mothers** *Section 7 of the Fair Labor Standards Act (29 U.S.C. 207) is amended by adding at the end the following*:
>
> (1) An employer shall provide reasonable break time for an employee to express breast milk for her nursing child for 1 year after the child's 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,

[7] HR 1594.

> 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 employer as defined in section 3(d) who employs 50 or more employees for each working day during each of 20 or more calendar workweeks in the current or preceding calendar year.
>
> *Penalty*
>
> Section 16(b) of such Act (29 U.S.C. 216(b)) is amended by inserting after the first sentence the following: "In lieu of any other remedy under this section or section 17, any employee who is harmed by a violation of section 7(r) may bring an action to enjoin such violation and to recover such equitable relief as may be appropriate to effectuate the purposes of such section."

We are interested in when a policy idea proposed in one bill ends up becoming law as part of another bill. Two bills share a policy idea when they include similar conferrals of authority, as in the example above. By this definition, a single law can contain many policy ideas and can be related to many other bills in different ways.

Several challenges remain, however. The first is to systematically identify shared language across thousands of bills and laws. The second is to address the question of how similar shared language needs to be in order to be considered the same idea. The third is to differentiate between conferrals of authority relevant to a study of the progress of policy ideas from other conferrals of authority. Most bills contain an authorization of appropriations section, or provisions authorizing commissions, mandating reports, making adjustments for inflation, providing protections to whistleblowers, and so on. These "boilerplate" conferrals of authority seem less relevant in that they tend to be peripheral to the main thrust of a policy proposal.

To summarize, existing approaches to tracing ideas or making connections among bills based on their substance are either too limiting (as in the case of bill summaries) or unreliable (as in the case of related bills). We propose a systematic approach that focuses on whether bills share common language. We define a policy idea in narrow statutory terms—as a conferral of authority. This distinguishes the policy idea from other shared legislative language, except that we also need to filter common conferrals of authority, or boilerplate provisions. The next step is to investigate whether it is possible to systematically differentiate shared policy ideas from other shared language between bills.

A Text Reuse Approach to Tracing Policy Ideas

To address these challenges, we turn to computer science methods developed to trace "text reuse" in documents (Brin, Davis and García-Molina 1995; Büchler et al. 2010). The appropriate unit of analysis for studying policy ideas in legislation is the bill section, as sections have long been an important break point in legislation: "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, 8). Further, the rules of construction for laws in the U.S. Code dictate that each section "shall contain, as nearly as may be possible, a single proposition of enactment."[8] We are interested in identifying when two sections of two different bills propose the same policy idea.

Many machine learning algorithms perform quite well in assessing document similarity using a "bag of words" approach (Grimmer and Stewart 2013). However, text reuse research finds that additional information about word or character sequence is generally helpful. Instead of simply asking whether two documents share words, word ordering also matters. Plagiarism software is perhaps the most familiar application (Hoad and Zobel 2003), but text reuse methods are employed broadly—in information retrieval to identify duplicate search queries; in communications research to study the diffusion of memes; in digital humanities research to trace the historical influence of important books, and even to compare musical scores (Downie and Nelson 2000; Henzinger 2006; Leskovec, Backstrom, and Kleinberg 2009).

In general, incorporating more information about sequence implies more computational effort, so that processing time becomes an issue where large numbers of comparisons are involved (as is the case for comparing every section of thousands of lengthy bills). One of the most efficient text reuse algorithms simply calculates the proportion of character pairs (bigrams) shared by two documents (Dice 1945). Other "n-gram" approaches judge similarity based upon more extended character or word sequences. With respect to the latter, there are two main options where bill sections are concerned. The first

[8] Title 1, Chapter 2, Section 104.

calculates the "global" or overall similarity of documents (Needleman and Wunsch 1970). The second "local" alignment option finds and scores shared subsequences of text within documents. Because the latter does not penalize cases where two bill sections share a substantial amount of text while differing in other important respects, we opt for a local alignment approach.

The Smith-Waterman local alignment algorithm (SWAlign) used here was specifically developed for genetic sequencing applications (Smith and Waterman 1981). Essentially, the method uses a dynamic programming approach to calculate a score for all alignments above a baseline matching n-gram (in our case, the baseline is a 10-word sequence). The algorithm score goes up whenever the next character building out from the original 10-gram is a match, and it goes down when it is a mismatch. The alignment terminates when the mismatches become too numerous. This tolerance for some difference is beneficial for a study of policy ideas, given that minor changes in language are expected as ideas migrate. The appendix describes dynamic programming in more detail and how tolerances are determined.

Table 1 provides an example of a Smith-Waterman alignment. The text on the left comes from a section of S 1244, a Senate bill introduced on June 11, 2009, that never made it out of committee. The text on the right comes from the PPACA as enacted. They are not identical but clearly propose the same policy idea. This particular example also illustrates the advantages of a local alignment approach for a study of policy ideas. The sections are not that similar overall. The alignments themselves span just 59% and 55% of the respective bill sections. A local alignment approach will capture shared policy ideas in sections that are viewed as very dissimilar from a global alignment perspective.

Dealing with Big Data

The data include the complete texts of every version of every bill introduced in the 111th Congress (28,891 versions of 11,081 bills). After downloading the data, we wrote a Python script to parse each bill by section and exclude other common text features such as titles and tables of content.[9] The end product is a database of 119,704 unique bill sections, for a total of 7.2 billion unique pairwise section comparisons.

As discussed, n-gram approaches such as the Smith-Waterman algorithm are computationally expensive. This can be a problem for projects involving lots of comparisons of relatively lengthy texts. Early experiments indicated that calculating similarity scores for the 111th Congress using a robust off-the-shelf global alignment plagiarism package (WCopyFind)[10] would have required more than 2,000 hours of processing time on a single-instance Amazon EC-2 micro server. This was unacceptable given that we would probably need to run our analyses multiple times. Fortunately, this is a common problem in machine learning research. We employ an approach that reduces memory requirements by converting text strings to shortened references, or hashes, and that reduces processing time by initially filtering section pairs below a minimum level of similarity (Huston, Moffat, and Croft 2011). The appendix provides more details about this process. The result was to reduce the corpus for which alignments were calculated from 7.2 billion pairings to approximately 1.6 million. The computing time required was reduced to just a couple of hours.

Building the PPACA Train/Test Validation Corpus

The 1.6 million section pairings yielded approximately 4 million alignments. The question of interest is whether the SWAlign scores for these alignments can predict which are shared policy ideas. To answer this question, we turn to human annotators to develop a "gold standard" data set. We randomly selected 3,400 of the top 50% of the alignments where one of the sections came from the enrolled version of the PPACA.[11] Annotators (two of the coauthors and a graduate research assistant) judged whether these alignments included a shared policy idea. The annotation process did not begin with a clear set of instructions. We employed what Saldana (2012) describes as first-and second-cycle coding methods. We started with the goal of differentiating alignments that contained a shared policy idea from other alignments. Our rule was that a policy idea needed to be comprehensible to the annotator. Inevitably, some cases of obscure policy changes will be overlooked, such as when a law deletes unspecified 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

[9] The Government Printing Office posts published bills in plain text format from 1989 to the present. We exclude the 29 Public Prints (http://www.gpo.gov/fdsys/browse/collection.action?collectionCode=BILLS). Although XML versions are made available through the House of Representatives for the 111th Congress only, many bills are missing, including the enrolled version of the PPACA.

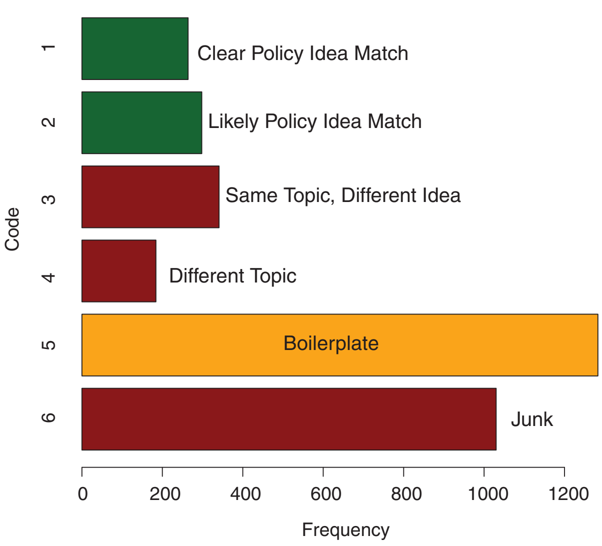
[10] http://plagiarism.bloomfieldmedia.com.

[11] This top 50% includes alignments that span at least 7% of one or both sections. The alignments in the sample range from 42 to 24,790 matching characters, with a median of 288.

TABLE 1 A Local Alignment Example

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	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

FIGURE 2 Histogram of Human-Labeled Alignments by Category



to facilitate the and all that follows"). Applying this rule raised questions about particular cases, which helped us to further refine our coding protocol.

The end result of the first coding cycle was six categories of alignments (see Figure 2). Category 1 includes cases where the annotator had no doubts about whether the alignment included a shared policy idea. Category 2 was for cases where there was some doubt. Categories 3 and 4 include cases where the aligned texts contained policy ideas that either addressed the same general topic but proposed different conferrals of authority (Category 3) or proposed similar conferrals of authority but for different topics (Category 4).[12] Category 5 was for alignments that addressed common boilerplate conferrals of authority, such as those establishing commissions or requiring reports. A final category contains the remaining cases that did not fall into the other categories. These tended to be shorter, often incomprehensible alignments.

Shared policy ideas (Categories 1 and 2) made up about 16% of the sample. Alignments that contained different policy ideas made up another 15% of the sample. The largest category is boilerplate, although Category 6 would undoubtedly have been the largest had we not sampled from the top 50% of all alignments. Interannotator agreement (three raters) with respect to the presence of a shared policy idea (Category 1 or 2 versus other) was above 90%.

Predicting Shared Policy Ideas

Can SWAlign scores distinguish the cases of shared policy ideas? We first train a supervised machine learning algorithm (support vector machine, or SVM) to predict and exclude boilerplate alignments (Joachims 2002). Next, we divide the remaining sample into cases of shared policy ideas (Categories 1 and 2) versus the rest before applying a logistic regression model where the only independent

[12] An example of the former would be education subsidies for nurses versus dentists, whereas the latter might include education subsidies for nurses and engineers.

TABLE 2 Results from Thousandfold Cross-Validation (2,900 Train, 500 Test)

	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 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]

variable is the pairing's SWAlign score. Table 2 reports the thousandfold cross-validation results (2,900 train, 500 test). Overall accuracy is 92% (accuracy will be higher across the entire corpus because the sample is biased in favor of higher alignment scores). A threshold SWAlign score of 1046 does an excellent job of predicting "true" cases of policy ideas. Recall is 97.3%, which means that Type I errors (false negative) are relatively rare. However, precision is 65%, indicating that Type II (false positive) errors are more common.

A review of the falsely positive cases revealed that most were boilerplate that the initial SVM learner failed to predict.[13] To improve the training set for boilerplate, we clustered all of the alignments for the 111th Congress (on the correct assumption that alignments across many bills are often boilerplate). We then combed the top clusters (those including 50 or more sections) and tagged the ones that were examples of boilerplate. We then coded an additional 2,000 alignments above the SWAlign threshold (1046) prediction for a shared policy idea for whether they were examples of boilerplate.

The findings presented in the remainder of this article are based on a replicable three-step method: (1) Retain section pairs with SWAlign scores above 1046; (2) retain only cases that include "shall," "may," "must," or "is amended"[14] and (3) exclude predicted boilerplate. For the PPACA analysis, we only consider pairs from the 111th Congress that include sections of HR 3590 as enrolled and a version of another bill introduced before the Senate passed the final version of HR 3590 on December 24, 2009. This yields 1,207 shared policy ideas. For the broader analysis of the 111th Congress, we limit our attention to pairs that include an enrolled bill and the introduced version of another bill (introduced prior to the enrollment date of the enrolled bill). We also restrict the scope of the policy areas examined for reasons to be discussed. This produces 2,474 shared policy ideas for that analysis.

Tracing the Policy Development of the PPACA

The limited legislative history of HR 3590 calls for an alternative approach that can trace policy ideas found in the PPACA to prior bills. Because much of the House and Senate's health care reform efforts centered on earlier markup bills, we expect a substantial number of connections between bills and provisions of the PPACA. Given that HR 3590 as amended was essentially the Senate's version of health care reform, we further expect the law to align most strongly with the Senate markup bills (S 1679 and S 1976). In addition, the history of the PPACA may include other bills that shaped the PPACA directly, or indirectly through the earlier inclusion of ideas in the markup bills.

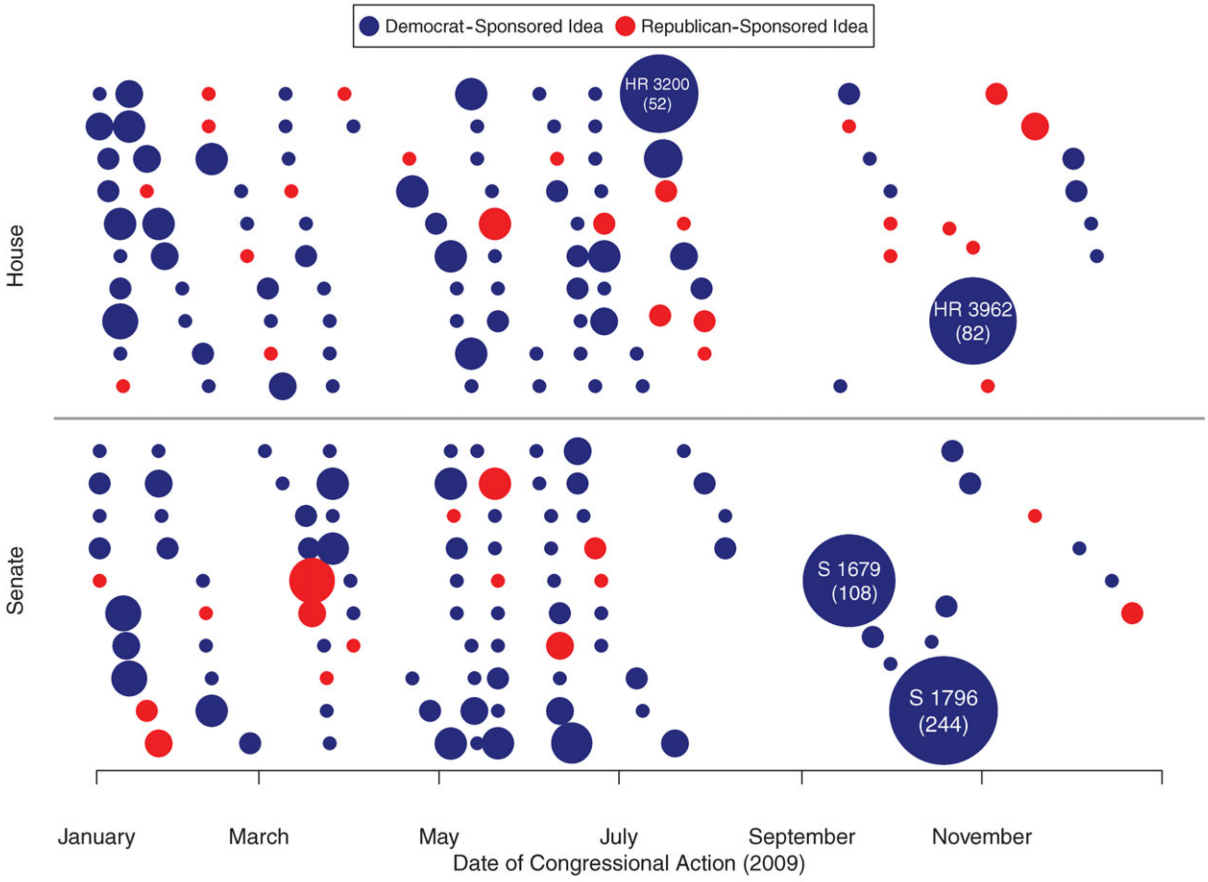
Each dot in Figure 3 is a bill introduced before the Senate passed what turned out to be the final version of the PPACA on December 24, 2009.[15] Blue indicates a bill sponsored by a Democrat. A dot's size corresponds to the number (natural log) of shared policy ideas between that bill and the PPACA. Of 432 substantive sections of the PPACA, 312 align with sections of 204 bills introduced earlier. Only four of these bills became law on their own. The four largest blue dots are the four markup vehicles. As expected, the Senate bills have the most in common with the law. The figure additionally indicates that 124 sections of the PPACA align with sections of other bills published before the first markup bill was published (on September 17).

[13] The small number of Type I false-negative errors were Category 2:- alignment where the annotators were less confident about whether it was a shared policy idea.

[14] Of the shared policy ideas, 99.7% in the sample of 3,400 included one of these legal terms associated with conferrals of authority, compared to only 70% of the other cases.

[15] The vertical dimension is purely for spacing purposes.

FIGURE 3 **Bills Sharing Policy Ideas with the PPACA (by Date of Introduction)**



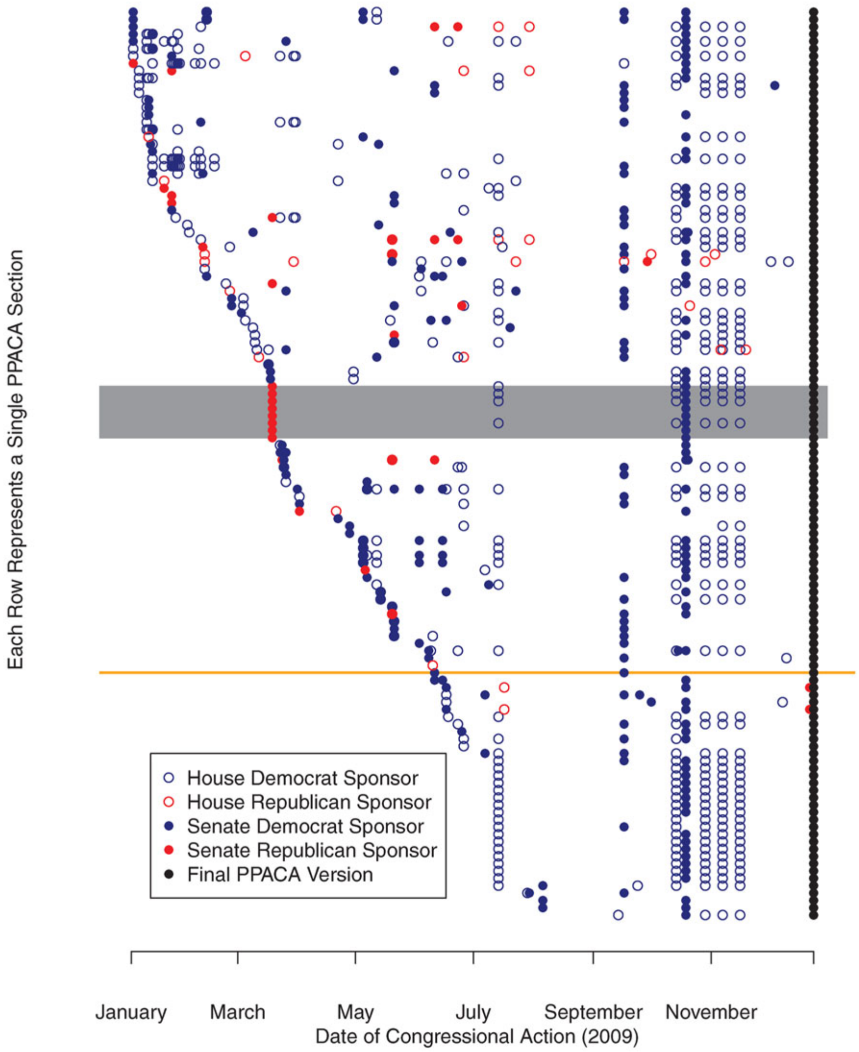
Earlier, it was noted that Obamacare was the product of two laws. Filibuster concerns led the House to pass the Senate's bill (HR 3590) without changes. The House then employed a budget reconciliation bill (HR 4872) to make additional changes without incurring a filibuster. Given that HR 3590 was essentially the Senate's version of health care reform (as we have seen it shared the most in common with the committee markup bill S 1976), we would expect HR 4872 to more strongly align with House bills. Ninety percent (96/105) of the alignments are with other House bills (not shown). The greatest number are with HR 3221, the Student Aid and Fiscal Responsibility Act of 2009, and the two markup bills, HR 3200 and HR 3962.

Another perspective is to trace the origins of particular policy ideas in the PPACA. Each row in Figure 4 is a PPACA section, and each column is a date prior to Senate passage of the act, starting with the first day of the congress (January 6, 2009). The colored dots (Senate) and circles (House) indicate sections of bills proposing a policy idea later found in that section of the PPACA. The upper rows include the policy ideas that can be traced furthest back in time. For example, the first row includes language authorizing grants to promote essential services for postpartum depression. Its earliest alignment is with HR 20, the Melanie Blocker Stokes MOTHERS Act, which was introduced on the first day of the congress. Three other bills introduced on that day also contain policy ideas later found in the PPACA: the Mental Health Parity Act (S 77) sponsored by John Kerry (D-MA), the Prevention First Act (S 21) sponsored by Harry Reid (D-NV), and the Small Business Empowerment Act (S 93) sponsored by Sherrod Brown (D-OH).

The latter may not sound like a health care bill, but it "directs the Secretary of Health and Human Services to establish a national program to make quality, affordable health insurance available to small employers and self-employed individuals in a manner that will spread risk on a national basis, modeled on the federal employees health benefit program." The PPACA section that aligns with this bill defines a small employer.

Further down, the shaded area of the figure encompasses sections of Title VI, Subtitle B, of the PPACA.

FIGURE 4 **Sections of Other Bills Sharing Policy Ideas with PPACA Sections**



Here, the earliest alignments are with the Nursing Home Transparency and Improvement Act (S 647) sponsored by Charles Grassley (R-IA). Grassley's work on this issue can be traced back to when he was chair of the Special Committee on Aging from 1997 to 2000.[16] Toward the bottom of the figure, the thin gold line highlights the workplace breastfeeding language discussed earlier. The earliest alignments in this case are with two identical bills, S 1244 sponsored by Jeff Merkley (D-OR) and HR 2819 sponsored by Carolyn Maloney (D-NY). Representative Maloney's efforts in the area can be traced back several congresses on THOMAS, and her website confirms that this is a case of a borrowed policy idea: "I was so proud to partner with Senator Jeff Merkley (D-OR) to pass into law a provision of our bill, the Breastfeeding Promotion Act (H.R. 2819, S. 1744), in comprehensive health care reform legislation signed by President Obama on March 23, 2010."[17]

Significantly, none of the bills discussed above are on the Congressional Research Service's list of related bills.[18] The point is not to single out CRS for criticism. As they note, the complexity of relationships among bills makes for a difficult task. The point is to underscore the potential of text reuse methods as a means for discovering relationships difficult to detect using other means. Five of

[16] In a 2007 hearing on the topic, the current chair (Herb Kohl, D-WI) acknowledged Grassley's long-standing interest and called him as the first witness (http://www.gpo.gov/fdsys/pkg/CHRG-110shrg41836/html/CHRG-110shrg41836.htm).

[17] https://maloney.house.gov/issue/breastfeeding.

[18] thomas.loc.gov.

the 13 bills that CRS does relate to the PPACA were not sources of policy ideas found in the law. Two proposed home mortgage subsidies for service members (HR 3780, S 1728); two were introduced after HR 3590 became law (HR 4872, S 2864); and one does not share significant substantive content with any section of the PPACA (S 1790). The other eight bills also show up on our list, but our list includes many more bills.

Inclusiveness in Lawmaking

How might a focus on the progress of policy ideas contribute to legislative research? One point of departure is the long-standing focus on bill success in legislative effectiveness and productivity research (Ainsworth and Hanson 1996; Anderson, Box-Steffensmeier, and Sinclair-Chapman 2003; Berry, Burden, and Howell 2010; Hasecke and Mycoff 2007; Krehbiel 1998; Krutz 2005; Mayhew 1991; Schiller 1995; Volden, Wiseman, and Wittmer 2013). Members of Congress object that bill success is a misleading measure of their effectiveness (Farenthold 2014). A policy ideas perspective suggests that bill success both overstates and understates legislative effectiveness. It overstates effectiveness in the sense that the law's sponsor receives all of the credit for what a bill contains. Often this is highly inappropriate. The version of HR 3590 introduced by the chair of the House Ways and Means Committee, Charles Rangel (D-NY), subsidized mortgages for military personnel. Rangel played little role as HR 3590 evolved into the PPACA in the Senate, yet he receives the credit. More generally, the extraordinary success of committee leaders and majority party members probably reflects their disproportionate ability to monopolize valued credit-claiming opportunities as much as it does their policy influence (Adler and Wilkerson 2012).

Bill success understates legislator effectiveness by providing no credit to the lawmakers, such as Merkley, Maloney, and Grassley, who successfully advocated for policy ideas that became law as provisions in other bills. In the 111th Congress, minority party members sponsored 8.7% of the non-minor bills that became law.[19] Every Republican lawmaker opposed the PPACA, yet Table 3 indicates that many of its provisions align with bills introduced by Republicans earlier in the process. The top half of the table considers the bills that contain shared policy ideas. The bottom half considers aligned sections (there can be more than one per bill, after excluding the four Democratic-sponsored markup vehicles). By either measure, a substantial number of policy ideas in the PPACA align with ideas proposed in Republican-sponsored bills.

TABLE 3 "Inclusiveness" by Chamber and Party Status

	House	Senate
Sponsors of All Bills Aligning with the PPACA		
Minority	25.5%	22.4%
Majority	74.5%	77.6%
N	106	98
Sponsors of Aligned Sections, No Markup Bills		
Minority	16%	25.1%
Majority	84%	74.9%
N	250	183

FIGURE 5 Distribution of Policy Idea Sponsors

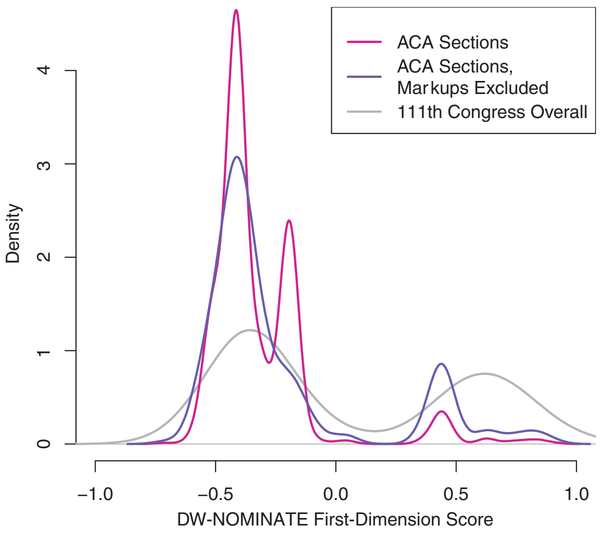


Figure 5 compares DW-NOMINATE density distributions for the 111th Congress with those for the sponsors of policy ideas (including and excluding the major markup bills). The blue line indicates that the law was shaped by a more diverse distribution of lawmakers than voting patterns on the PPACA would lead us to expect.

Scaling Up to the 111th Congress

Having examined how a focus on policy ideas might inform understandings of the legislative history of a major law, we now begin to explore the challenges and

[19] Minor bills name buildings, transfer small properties, provide relief for individuals, and authorize the minting of coins and medals.

opportunities associated with using text reuse methods to study policymaking patterns more generally. We see two main concerns. The first is whether SWAlign scores are equally valid predictors of shared policy ideas across different issues. The second is whether we can be as confident about whether a bill was the source of a policy idea.

Issue Variations and Validity

Bill drafting conventions differ across issue areas. These differences mean that a SWAlign threshold that predicts shared policy ideas in one issue area may be less accurate in others. These areas tend to involve lots of bills that typically address relatively minor issues. In the area of trade, hundreds of virtually identical temporary duty suspension bills are introduced each congress. The only difference between these bills may be a couple of characters describing the thickness of a textile fiber. The differences are somewhat greater for bills proposing land transfers between the federal and local governments, naming federal buildings, minting commemorative coins, and providing relief for private individuals. Appropriations bills also tend to conform to a standard format. For these issue areas, an algorithm with less tolerance for difference is needed. One of the advantages of the Smith-Waterman algorithm is that the penalties for character mismatches and gaps can be adjusted.

Law Histories and Inference

The PPACA was fairly unique in that HR 3590 had no legislative history. The bill addressed a different issue (mortgage subsidies) until the Senate inserted its version of health care reform as a floor amendment. The House then accepted that version without changes. It is impossible to prove that another bill was the inspiration for an idea found in the PPACA, but this absence of history made it easier to argue that bills published earlier helped to shape its content.

Most laws have histories in that a version is published months or even years before enactment. This raises an additional challenge for inference because it is possible that the bill that eventually became law may have influenced the substance of other bills. In other words, the chain of causality may be reversed. As we saw with the PPACA, companion bills are sometimes introduced simultaneously. On hot or recurring issues, members of both parties may sponsor bills that are similar (even identical) in many respects.

The results reported here for the 111th Congress address the first but not the second of these concerns. The latter raises a new set of methodological challenges. It requires an approach that controls for the substance of the law as introduced. The focus of analysis would be on how a bill that later became law changed after introduction (or its earliest version), and whether those changes can be linked to other bills proposed in the interim or before.

For the 111th Congress, 44,000 alignments exceed our SWAlign 1046 threshold for a shared policy idea. Limiting attention to alignments between introduced versions of bills and other bills that became law after those bills were introduced reduces this number to 3,860. We then exclude alignments involving the problematic issue areas discussed above, as well as boilerplate language.[20] The final data set includes 2,474 shared policy ideas (involving 779 introduced bills and 136 laws).

The 906-page PPACA accounts for 493 of these cases.[21] The next law that shares the most in common with other bills is the 849-page Dodd-Frank Wall Street Reform and Consumer Protection Act (266 shared policy ideas), followed by two large defense authorization laws (HR 2647, HR 6523). With respect to Dodd-Frank, our alignments catch all but one of the CRS related bills (S 3217) and many others besides. The anomaly is explained by the censored nature of our analysis. There was no "as introduced" version of S 3217. The first published version was the bill as placed on the Senate calendar (after the Senate Banking, Housing, and Urban Affairs Committee reported a "clean" bill).[22]

Minority Republicans sponsored 8.7% of the nonminor bills that became law in the 111th Congress. Here, Republicans sponsored 20.8% of the House bills that include policy ideas found in later laws and 22.1% of the Senate bills. There are also substantial variations across laws. Excluding the Senate markup bill (S 3217), 30% of the alignments between Senate bills and Dodd-Frank are with Republican-sponsored bills, compared to 15.9% for the House. In contrast, for the omnibus Military Appropriations Authorization (HR 2647), 28.8% of the House alignments are with bills introduced by Republicans, compared to just 8.1% for Senate Republicans. The laws with the greatest number of Republican-supported

[20] We excluded all bills in Policy Agendas Project major topic 21 (Public Lands) and subtopics 709 (also public lands related), 2000 (Appropriations), 2008 (Government Property Management), 2006 (Commemorative Coins and Medals), and 1807 (Tariffs) (Baumgartner and Jones 1993).

[21] This is a smaller number than in the earlier analysis because here we only consider the introduced versions of other bills.

[22] http://www.gpo.gov/fdsys/pkg/BILLS-111s3217pcs/html/BILLS-111s3217pcs.htm.

ideas besides the PPACA (111) and Dodd-Frank (51) are the Small Business Jobs Act of 2010 (34) and the Coast Guard Authorization Act of 2010 (26).

Discussion

By the time the average bill becomes a law, it is substantially longer than when it was introduced. Scholars appreciate that bills evolve, yet systematic studies of how laws develop are rare. One implication is that many policy successes go unnoticed by scholars and the general public. This article confirms that it is possible to systematically document the "tapestry of histories" that characterize many laws (Cannan 2013).

We define a policy idea as a conferral of statutory authority and draw on computer science methods to trace policy ideas in legislation. Similarity scores based on the Smith-Waterman local alignment algorithm accurately predict shared ideas first identified by human annotators. These methods provided new insights into the legislative history of the massive Patient Protection and Affordable Care Act. The final law shared ideas with 232 bills introduced earlier. We were also able to show how many ideas in the law and the main markup bills could be traced to provisions of earlier bills, and that many of these antecedent provisions were sponsored by Republican lawmakers (who ultimately voted against the law).

Our initial implementation indicated that false positives are a concern. An improved algorithm for excluding boilerplate language will address many of them, but there are undoubtedly other undetected cases where small text differences are meaningful. False positives are an even greater concern in more general applications where bill drafting conventions differ and where the bill that became law may have influenced the substance of other bills at an earlier stage. And, of course, text reuse methods cannot establish causality. Our claims of policy influence in specific cases were ultimately based on assumptions and contextual evidence.

Much more remains to be done both in terms of method and analysis; we believe that this investigation clearly demonstrates the potential of reuse methods for advancing legislative research related to the progress of policy ideas. One important finding appears to be that lawmaking is a more inclusive process when judged in terms of policy ideas. That good ideas matter often seems lost in all of the research attention devoted to demonstrating partisan polarization and dysfunction. When Grassley announced that he would not support the PPACA, Democratic leaders did not respond by stripping his language from the bill. What Grassley proposed had been vetted over several congresses, and it made sense. He probably would not have been able to pass a separate bill in a Democratically controlled congress, but his ideas did take root as provisions of a Democratically sponsored bill. A preliminary analysis of the 111th Congress found similar patterns, with thought-provoking differences across issues, chambers, and parties. Of course, motivations other than good ideas, such as logrolling and standard operating procedures (as in the case of miscellaneous trade bills), can also explain the "uptake" of policy ideas (Sulkin 2005).

Text reuse methods can also be used to investigate processes that come before lawmaking, for example, by connecting bill language to "model" legislative language proposed by interest groups, and after, by examining the regulatory process. The Administrative Procedures Act requires agencies to solicit public comments as they develop regulations. Text reuse methods can be used to study how regulations change between the initial and final versions, and to relate those changes to public input. Which comments (submitted by which actors) have the most influence on the development of regulations? Which agencies tend to be the most responsive?

Where lawmaking is concerned, many additional questions can be investigated. How often do ideas become law as provisions of other bills? How might assessments of member effectiveness differ when policy idea successes are combined with bill successes? Which stages of the legislative process produce the greatest amounts of policy change? In which issue areas are idea entrepreneurs most likely to succeed? How do credit-claiming considerations impact the sharing of policy ideas? Has idea sharing changed over time? Is it more or less likely under divided governments?

Appendix: Computing Local Alignments

Reducing Candidate Section Pairs

We first build an inverted index of every repeated word n-gram in the corpus of 119,704 sections. In the first pass, n-grams of 10 words or more are hashed into a fixed number of bins.[23] In the second pass, the n-grams that hash to bins with just one occupant (indicating a unique n-gram) are discarded.[24] Next we create a list of sections

[23] Replicating this process using a more inclusive five-word-gram threshold had no impact on the results of our validation experiment.

[24] Hash collisions (instances where a single shortened reference maps to more than one text string) can allow a small number

containing each distinct 10-gram, excluding n-grams too frequent to be discriminative. In our default setup, we exclude n-grams that occur more than 100 times. We output all combinations of section pairs in each list, excluding section pairs from (different versions of) the same bill. Lastly, we retain only those section pairs that have at least five 10-grams in common.

On a five-year-old cluster of commodity servers, indexing the 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 eightfold parallelism; and performing Smith-Waterman alignment on these pairings took 28 minutes with 50-fold parallelism.

Calculating Smith-Waterman Local Alignment Scores

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:

$$\begin{aligned} H(0,0) &= 0 \\ H(i,0) &= E(i,0) = W_g + i \cdot W_c \\ H(0,j) &= F(0,j) = W_g + j \cdot W_c \end{aligned}$$

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 Smith-Waterman alignment score reflects the extent to which the characters in the sequences align and the overall length of the sequence.[25]

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 Smith-Waterman alignment scores.

References

Ainsworth, Scott, and Douglas Hanson. 1996. "Bill Sponsorship and Legislative Success among Freshmen Senators, 1954–1986." *Social Science Journal* 33(2): 211–21.

Anderson, William J., Janet M. Box-Steffensmeier, and Valeria Sinclair-Chapman. 2003. "The Keys to Legislative Success in the U.S. House of Representatives." *Legislative Studies Quarterly* 28(3): 357–86.

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

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

of singleton n-grams to make it past the first stage. These are subsequently filtered as the index is written.

[25] See also http://www.cs.kent.edu/ssteinfa/files/PDSEC08˙handouts.pdf.

Berry, Christopher R., Barry C. Burden, and William G. Howell. 2010. "After Enactment: The Lives and Deaths of Federal Programs." *American Journal of Political Science* 54(1): 1–17.

Brin, Sergey, James Davis, and Héctor García-Molina. 1995. "Copy Detection Mechanisms for Digital Documents." *SIGMOD Rec.* 24(2): 398–409.

Büchler, Marco, Annette Geßner, Gerhard Heyer, and 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*. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-637.html.

Burstein, Paul, Shawn Bauldry and Paul Froese. 2005. "Bill Sponsorship and Congressional Support for Policy Proposals, from Introduction to Enactment or Disappearance." *Political Research Quarterly* 6(1): 295–302.

Cannan, John. 2013. "A Legislative History of the Affordable Care Act: How Legislative Procedure Shapes Legislative History." *Law Library Journal* 105(2): 131–73.

Dice, Lee R. 1945. "Measures of the Amount of Ecologic Association between Species." *Ecology* 26(3): 297–302.

Downie, Stephen, and Michael Nelson. 2000. "Evaluation of a Simple and Effective Music Information Retrieval Method." In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 24–28 July 2000, Athens University, Athens, Greece. New York: ACM, 73–80.

Farenthold, David A. 2014. "Andrews Proposed 646 Bills, Passed 0: Worst Record of Past 20 Years." Washington Post, February 4. http://www.washingtonpost.com/politics/andrews-proposed-646-bills-passed-0-worst-record-of-past-20-years/2014/02/04/3240bb4e-8dbb-11e3-833c-33098f9e5267_story.html.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3): 267–97.

Hasecke, Edward B., and 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): 607–17.

Henzinger, Monika. 2006. "Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms." In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 284–91.

Hoad, Timothy C., and Justin Zobel. 2003. "Methods for Identifying Versioned and Plagiarized Documents." *Journal of the American Society for Information Science and Technology* 54(3): 203–15.

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*. New York: ACM, 127–36.

Joachims, Thorsten. 2002. *Learning to Classify Text Using Support Vector Machines*. New York: Kluwer/Springer.

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

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

Krutz, Glen S. 2005. "Issues and Institutions: Winnowing in the U.S. Congress." *American Journal of Political Science* 49(2): 313–26.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. 2009. "Meme-Tracking and the Dynamics of the News Cycle." In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 497–506.

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

Needleman, Saul B., and 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): 443–53.

Saldana, Johnny. 2012. *The Coding Manual for Qualitative Researchers*. Thousand Oaks, CA: Sage.

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

Sinclair, Barbara. 2011. *Unorthodox Lawmaking: New Legislative Processes in the U.S. Congress*. Washington, DC: Congressional Quarterly.

Smith, T. F. and M. S. Waterman. 1981. "Identification of Common Molecular Subsequences." *Journal of Molecular Biology* 147(1): 195–97.

Sulkin, Tracy. 2005. *Issue Politics in Congress*. London: Cambridge University Press.

Volden, Craig, Alan E. Wiseman, and Dana E. Wittmer. 2013. "When Are Women More Effective Lawmakers Than Men?" *American Journal of Political Science* 57(2): 326–41.

Walker, Jack L. 1977. "Setting the Agenda in the U.S. Senate: A Theory of Problem Selection." *British Journal of Political Science* 7(4): 423–45.