THREE ESSAYS ON LEGISLATIVE TEXT ANALYSIS

A Dissertation in
by
Matthew J. Denny

© 2019 Matthew J. Denny

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2019
The dissertation of Matthew J. Denny was reviewed and approved* by the following:

Bruce A. Desmarais  
Associate Processor of Political Science  
Dissertation Adviser, Chair of Committee

Burt Monroe  
Professor of Political Science

Michael Nelson  
Associate Processor of Political Science

David Hunter  
Professor of Statistics

John Wilkerson  
Professor of Political Science  
Special Member, University of Washington

Glenn Palmer  
Professor of Political Science  
Graduate Program Director

*Signatures are on file in the Graduate School.
Abstract

Over the past decade, Congress scholars have increasingly benefitted from the confluence of open government efforts to make vast amounts of government documents available online, and the development of computational resources powerful enough to analyze them at scale. In this dissertation, I introduce three complementary computation methods for analyzing legislative text, and apply them to a newly collected corpus of U.S Congressional bills introduced between 1993 and 2016. I do so in service of substantive contributions to the study of bureaucratic politics, lawmaking, and productivity in the U.S. Congress. In the first chapter, I show how partisan views on the scope of government permeate the placement the strategic use of constraints on the bureaucracy. In the second chapter, I attempt to characterize one dimension separating serious policymaking efforts from position taking legislation in the inclusion of legal details in legislation. And in the third chapter, which is joint work with Andreu Casas and John Wilkerson, we show how accounting for hitchhiker bills in the U.S. Congress reveals a more inclusive and productive lawmaking process. This dissertation also introduces a new corpus of congressional bill provisions, and an open sources statistical software package implementing the methods it introduces.
# Table of Contents

List of Figures 

List of Tables 

Acknowledgments 

Chapter 1
   Introduction 1

Chapter 2
   More than Control: Partisan Differences in the Use of Statutory Constraints on the Bureaucracy 4
   2.1 Introduction .............................................. 4
   2.2 Revisiting the Ally Principle ................................ 6
   2.3 Measuring Statutory Constraints on the Bureaucracy .......... 8
   2.4 Data and Preprocessing ..................................... 13
      2.4.1 Selecting Bills for Inclusion .......................... 13
      2.4.2 Splitting Bills Into Sections .......................... 15
      2.4.3 Substantive Bill Sections .............................. 16
      2.4.4 Implementation and Validation ......................... 18
   2.5 Analysis .................................................... 21
   2.6 Discussion .................................................. 28

Chapter 3
   Title of the Second Chapter 32
   3.1 Introduction .................................................. 32
   3.2 Theory and Hypotheses ...................................... 36
   3.3 Data and Methods ............................................. 44
      3.3.1 Identifying Legal Detail in Legislation .................. 47
      3.3.2 Using Contingency Tables to Classify Boilerplate Terms 48

iv
Appendix F
Efficient Optimization: Decomposing Mutual Information 111

Appendix G
Pre-processing 113

Appendix H
Constructing Statistical Features 117

Appendix I
Active Learning with a Massive Ensemble 120

Appendix J
Logistic Regression Models 123

Appendix K
Hitchhikers Bills for two Target Law Examples 126

Bibliography 128
List of Figures

2.1 S.1 - Keystone XL Pipeline Approval Act proportion of bill text devoted to each purpose. Note that roughly 5% of the text (one section) is devoted to the stated intention of the bill. ......................................................... 12

2.2 Difference in average substantive bill section word count based on presidential partisanship for bill sections introduced by Democrats and Republicans (1993 - 2016). A positive differential indicates that members of the party introduced longer bill sections under a President from the opposite party (consistent with the Ally Principle). ......................................................... 23

2.3 Moving average substantive bill section word count for Democrats and Republicans, 1993-2016. The stacked red (Republican) and blue (Democratic) bars along the bottom of the plot indicate partisan control of the Presidency, Senate, and House. ......................................................... 24

2.4 Histogram of number of substantive bill sections by word count, with a log$_{10}$ scale x-axis. ......................................................... 25

3.1 Distributions of the average proportion of boilerplate terms in substantive provisions of the versions as introduced of bills that eventually became law, and those that did not. ......................................................... 58

3.2 The proportion of introduced version of bills that eventually became law (or did not) containing at least one of each of seven types of non-substantive provisions. ......................................................... 59
3.3 Marginal effects (with 95% confidence intervals) of covariates on the relative likelihood of a bill being reported out of committee. Variables of interest are highlighted in gray. ........................................ 60

3.4 Marginal effects (with 95% confidence intervals) of covariates on the relative likelihood of a bill becoming law when it is first introduced. Variables of interest are highlighted in gray. ........................................ 61

3.5 Marginal effects (with 95% confidence intervals) of covariates on the relative likelihood of a bill becoming law, conditional on it having been reported out of committee. Variables of interest are highlighted in gray. ............................. 64

4.1 A comparison of two versions of a section of the Dodd-Frank Wall Street Reform and Consumer Protection Act. ........................................ 80

4.2 Counts of laws versus Hitchhiker bills (103rd-113th Congresses). ....... 82

4.3 How far do hitchhiker bills advance on their own? ............................. 83

4.4 Marginal effects of sponsor and bill characteristics on law versus hitchhiker success. .................................................. 85

4.5 Percentage of legislators sponsoring at least one law, or at least one law or hitchhiker. .................................................. 87

4.6 Comparison between the sponsor’s enactments and the Legislative Effectiveness Scores (LES) of (1). ........................................ 92

4.7 Where hitchhiker bills get picked up during the legislative process. ....... 93

I.1 Bill insertion predictions for an ensemble of 99 models ...................... 121

J.1 Key coefficients of interest when estimating a separate model for each Congress. .................................................. 123
List of Tables

2.1 Descriptive statistics for bill sections by type. *# Sections* records the number of bill sections of each type, and *# Bills* records the number of unique bills associated with that section type. *Substantive?* records where each section type was considered substantive for the purposes of my analysis. .................................................. 17

2.2 Descriptive statistics for substantive bill sections by issue area. The last two columns display the number of bill sections, and the number of unique bills associated with a given issue area, respectively. ................. 19

2.3 Average substantive bill section length for all bills introduced by Democratic and Republican legislators under Democratic and Republican Presidents, respectively (1993-2016). Substantive bill section counts in each category are displayed in parentheses under the average word count. .... 22

2.4 Regression results for all bills, and bills with at least five cosponsors. Dependent variable is the log of word counts for substantive sections from bills introduced between 1993-2016. Issue area fixed effects were mostly statistically significantly different from zero ($\alpha = 0.01$), and are omitted from these results for clarity. They are included in C. Parameter estimate standard errors are presented in parentheses under the estimate. Note that the (2) correction was used to calculate the standard errors for the interaction terms. ................................. 31
3.1 Counts of provisions by type. **# of Provisions** indicates the number of provisions of a given type, while **# of Bills** records the number of unique bills containing a provision of that type. **Substantive?** indicates whether each provision type was considered to be a substantive policy provision. Provision types highlighted in blue are the non-substantive provisions I focus on for the purpose of my analysis.

3.2 Each row in this example contingency table records the count of each unique term aggregated across all bills in a particular category (e.g. Democrat sponsored bills, about nuclear energy policy).

3.3 An example joint distribution over parties and terms.

3.4 Example mutual information calculations for two joint distributions.

3.5 Example mutual information scores for a distribution with and without the first column included.

3.6 Example terms with a negative ACMI.

3.7 Confusion matrix for human coding validation results. Total count of all terms in category is provided in parentheses.

3.8 Correlations among bill section legal and technical term counts using different cut-offs for identifying these terms.

C.1 Parameter estimates for Democrats and Republicans. Base category is Agriculture legislation. Parameter estimate standard errors are presented in parentheses under the estimate.

G.1 HR-146 bill insertion example. Matches highlighted in red.

G.2 HR-146 bill insertion example after pre-processing. Matches highlighted in red.

I.1 Summary of the Active Learning Process.
J.1 Model data: descriptive statistics table ........................................ 124

J.2 Results for two logistic regression models predicting whether a bill becomes a stand alone law (LAW) and, conditional on that not happening, whether it becomes law as a hitchhiker (HITCHHIKER). We include Congress (103 to 113th) and topic (Policy Agendas major topic) fixed-effects, although for simplicity we do not include the fixed-effect coefficients in the table. .......................................................... 125
Acknowledgments

I would like to begin by thanking Bruce Desmarais, Burt Monroe, Michael Nelson, John Wilkerson, and David Hunter for serving on my committee and supporting me through the ups and downs of writing my dissertation and going on the job market. I would also like to thank Hanna Wallach and Brendan O’Connor for everything they taught me about (political) text analysis. I would also like to thank all of the anonymous reviewers, conference panel discussants and friends and faculty who took the time to read and give feedback on parts of my dissertation research. And any thank you’s would be incomplete without a huge thanks to Kristy Boob for helping me navigate the Penn State bureaucracy with a great deal of kindness, patience, and warmth.

For their feedback on the third chapter, I would like to thank the anonymous reviewers at APSR and AJPS, as well as the editorial staff at AJPS that have been so patient in dealing with formatting issues. I would also like to thank Lexi Greenberg for her research assistance, Barry Pump for his procedural insights, and Amber Boydstun for suggesting the “hitchhiker” term we use in the paper. Last but not least, I would like to thank John Wilkerson and Andreu Casas for being amazing co-authors and sticking with this paper through a long publication process.

On a personal note, I would like to thank Bruce for always believing in me, and for helping me through all of the personal and professional challenges I faced while in grad school: with kindness, and without judgement. I would also like to thank Burt for for his flexibility and willingness to support me, both financially and academically in the wide range of projects I undertook while at Penn State. Along with those already mentioned, I owe a huge debt of gratitude to: Gerald Epstein, Dan Lass, Eve Voegel, Johnathan Wynn, John Spraggon, Angela de Oliveira, Nancy Folbre, Kevin Crocker, Tatishe Nteta, and all of the other faculty that supported and encouraged me to pursue a PhD in (anything).

This research is based upon work supported by the National Science Foundation under Grant No. 1224173 and IGERT Grant DGE-1144860, and by the Moore-Sloan Data Science
Environment. 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. It is also worth noting that many of the good parts of this dissertation would not have been possible without the help of many others, and that all errors and omissions are my own.

Finally, I want to thank my (step)parents Lorie, John, Alice and Walter, and my Grandfather John Foster for always believing in and supporting me. I could not have made it this far without you.
Chapter 1  |  Introduction

This dissertation introduces three related and complementary methods for legislative text analysis. I then apply these methods to three difficult and open questions in the study of the United States Congress. I make substantive contributions to the study of congressional constraints on bureaucratic policymaking, and to the study of lawmaking and legislative productivity. This dissertation also represents my broader philosophical approach to political science. Substantively, it fits with my goals of improving the measurement of latent political variables, and capturing the intents and strategic maneuvering of politicians through the legislation they write. Methodologically, it represents a “technology transfer” from political science and linguistics to computer science. I believe that social science and linguistic theory represent a substantial and relatively untapped resource for improving performance in many machine learning tasks, and that is exactly what I seek to do in this dissertation.

To build on this point, the methods I develop and implement in this dissertation are deliberately understandable, and make use of insights about politics to do measurement, instead of letting a black-box statistical model decide how to measure my variables of interest. All three chapters involve a human in the loop in the measurement process, highlighting the value that social scientists and domain experts can bring to heavily computational research tasks. And to the degree that algorithms do the heavy lifting, they are intended to try and capture human intuition about, for example, what a bill section is, what kinds of words are legalese, or when a bill has been inserted in the middle of another bill. To that end these methods introduce concepts and software that are portable to other domains, while
also being highly tailored to studying legislation. I believe this approach is an important complement to some of the ever more complex and uninterpretable (if accurate) machine learning models being applied in the social sciences.

Substantively, my motivation for selecting the topics for these dissertation chapters boils down to a relatively simple idea: that legislators are trying to tell us all sorts of things in the legislation they introduce. I make the case in this dissertation that if we listen (or in this case read) carefully, we can learn a lot of subtle details about the lawmaking process and congressional politics more broadly from the words legislators write. In particular, the first two chapters highlight how political consideration have percolated into a part of the lawmaking process we might expect to be less “political”—the inclusion of legal and technical language to clarify and support the substantive policy ideas included in that legislation. The third chapter was born in a slightly different way, through an incredibly fruitful collaboration with John Wilkerson and Andreu Casas where we set out to find the “hitchhiker” bills we knew ought to be there, but no one had systematically identified up to that point. What all of these chapters share in common is that they seek to uncover hidden political processes in Congress, something that I intend to make a focus of my research agenda going forward.

In the first chapter, I ask whether Republican lawmakers strategically time their efforts to constrain the bureaucracy, and offer a refinement of previous methods using the length of a bill as a measure of the constraints it places on the bureaucracy. Perhaps the most substantial contribution of this chapter is the computational methods for segmenting bills into sections, and the resulting dataset of bill sections I produced to answer my research question. This dataset has opened up numerous additional research projects by accurately breaking legislation into its constituent policy proposals, which serve as much more comparable units than bills themselves. This computational system relied on hundreds of hours of refinement through tracking down numerous corner cases in legislation, and highlights the benefits of investing time and social science expertise in producing accurate and useful data. Theoretically, this chapter builds on a long-standing hunch that partisan politics have permeated every aspect of the lawmaking process in Congress, by arguing that Republicans focus their efforts to constrain bureaucratic policy making to times where a sympathetic
copartisan President is likely to agree with and allow their efforts to pass a presidential veto. My analysis supports this hypothesis, while suggesting that Democrats follow the more accepted “Ally Principle”, whereby they seek to place more constraints on bureaucrats under an opposite party President.

In my second chapter, I argue that legislators leave clues about where a piece of legislation falls on the spectrum between serious policymaking and position taking in the amount of legal and technical language they include in legislation. My theory is that legislators signal the “quality” of a piece of legislation based on the amount of legal resources they devote to drafting it, and higher quality, more serious legislation is therefore characterized by increased legal and technical language content. I also introduce a novel statistical approach to identifying this legal and technical language, and use it to measure the amount of legal detail included in over two decades of U.S. Congressional legislation. I present evidence I believe demonstrates that this language does carry a signal about bill quality, while also making the case for continued attention to the idea that some bills are intended to be serious policymaking efforts, while others are not.

Finally, my third chapter, which is joint work with Andreu Casas and John Wilkerson, examines the inclusion of hitchhiker bills (bills that do not become law themselves, but later are included as provisions of other bills that do become law) in U.S. Congressional legislation. We develop an approach that blends a computational linguistic measurement method with active learning to classify these hitchhiker bills with very high accuracy. We find that when we account for these hitchhiker bills, Congress is actually about twice as productive as we might think based purely on recording the number of stand-alone bills that become law. Furthermore, we find that the lawmaking process is significantly more inclusive when we account for hitchhiker bills.

The rest of this dissertation is broken up into three substantive chapters along with their supporting materials, and a short concluding chapter at the end. Additionally, many of the computational tools developed as part of this dissertation can be accessed through an open source R package available here: https://github.com/matthewjdenny/SpeedReader.
Chapter 2  
More than Control: Partisan Differences in the Use of Statutory Constraints on the Bureau-
cracy

2.1 Introduction

The “Ally Principle” is established finding that legislators who are co-partisan with the
President will delegate more, and place fewer constraints on the bureaucracy than legislators
from the opposite party (see, for example: 3; 4; 5; 6; 7). The theoretical argument is that
when legislators have policy preferences that are closer to those of bureaucratic officials,
they will be better able to trust those officials to implement policies in a manner consis-
tent with their goals. This trust allows legislators to leave more implementation details
up to bureaucratic agencies, and reduces their demand for constraints on bureaucratic
policy-making. The key link with presidential partisanship is that the President appoints
bureaucratic agency heads, so being co-partisan with the President should imply greater
policy preference similarity to those officials.

The Ally Principle has a long history of theoretical and empirical support.¹ Scholars

¹For a review of this literature, see: (8; 5; 6).
have found evidence in support of the Ally Principle in the U.S. states, at the federal level, and even cross-nationally (9; 8, see, especially:). The core logic behind the Ally Principle also makes intuitive sense. Democratic legislators during the Trump administration have strong reasons to believe that the Environmental Protection Agency (directed by Scott Pruitt) is less likely to implement environmental regulations in the way they would have preferred than under the Obama administration, for example.

At the same time, there are also examples where members of Congress seem to team up with a co-partisan President to place greater restrictions on bureaucratic officials. To follow the example of the EPA under President Trump, the types of legislation being advanced by Republican legislators appear to be more consistent with a strategy of making it more difficult for EPA officials to do their jobs (through placing numerous constraints on those officials), than simply dissolving the agency, or directly preventing it from regulating (10). One would be hard pressed to find a more conservative EPA director than Scott Pruitt, yet Republican legislators seem to be more focused on making it harder for his agency to function than on giving him free reign. Additionally, a recently published study actually finds evidence of a “reverse” Ally Principle among state bureaucratic agencies (7). This contradictory evidence poses a puzzle: does the Ally Principle only apply in some settings, is this contradictory evidence just an anomaly, or is policy preference alignment between bureaucrats and legislators just one piece of the puzzle in explaining legislators’ decisions to curb bureaucratic discretion?

I argue that statutory constraints (details written into legislation with the purpose of reducing the discretion of the implementing agency) are a policy tool, and that partisan views on the role of government are just as important as policy preference alignment in determining legislators’ use of statutory constraints. In particular, I develop theoretical expectations for how Democrats and Republicans will use statutory constraints differently under a more or less ideologically aligned President. To investigate this theory, I refined and extended the approach to measuring bureaucratic discretion in legislation developed by (8), by shifting from examining entire bills to considering each constituent policy proposal separately. In doing so, I was able to measure statutory constraints on the bureaucracy across all U.S. congressional legislation introduced between 1993 and 2016.
My findings suggest that statutory constraints may be used much more strategically and in a more targeted fashion than was previously thought. For example, I find that Republicans tend to write significantly longer substantive provisions in their legislation under a Republican President than a Democratic one (evidence consistent with a reverse Ally Principle), while Democrats behave in a manner that is much more consistent with the traditional Ally Principle. These new findings were made possible by the use of a new preprocessing pipeline for legislative texts, which allows me to make more accurate and much larger scale measurements of statutory constraint than previous studies.

2.2 Revisiting the Ally Principle

There is a rich literature seeking to model the relationship between Congress (both as an institution and from the perspective of individual legislators) and the bureaucracy tasked with implementing the policies members of Congress introduce. Theoretical work has often conceptualized this relationship as a principal-agent problem (11; 12). Legislators must delegate some policy-making authority to bureaucratic officials because they lack the time (13; 14), expertise (15), and/or political incentives (16) to develop precise solutions to all policy problems (17; 3; 8). However, agencies may not have incentives to enact policies in a way that those legislators would have preferred (18; 19). As a consequence, legislators have developed a number of mechanisms to constrain the actions of bureaucratic officials (11). These mechanisms include congressional oversight (20), limiting statutory discretion (9; 8), budgetary restrictions (21), and the design of administrative procedures (22; 23).

In general, most theories of bureaucratic control frame the relationship between bureaucratic officials and legislators in terms of a divergence in policy preferences (see, for example: 24; 3; 5; 11; 12). From this perspective, legislators will be more or less likely to achieve their desired policy outcomes based on how closely their preferences align with those of the relevant bureaucratic officials, and the mechanisms available to enforce compliance. All legislators want to achieve policy goals, so the theoretical focus is placed less on differences in those policy goals and more on congruence between the goals of legislators and bureaucrats.
Principal-agent theory suggests we should see some evidence of the Ally Principle, whereby members of Congress who are more ideologically aligned (co-partisan) with the President will seek to place fewer constraints on the bureaucracy. The argument goes that the President appoints agency heads, so ideologically aligned (co-partisan) legislators can more easily trust that bureaucratic agencies will implement policies in a way they would have chosen (3; 4; 5). This allows these members of Congress to write less specific, constraining legislation, and to delegate more implementation details to “friendly” bureaucratic officials. The Ally Principle tells us that, for example, Republican lawmakers will seek to place more constraints on the bureaucracy under a Democratic President than a Republican one.

While the Ally Principle has previously found empirical support, more recent work calls this finding into question (7). Palus and Yackee draw on social identity theory and argue that partisan alignment may actually decrease the perceived policy discretion by bureaucratic officials as they seek to align themselves with the in-group. There are also notable instances where Congress seems to completely violate the predictions of the Ally Principle. For example, President George W. Bush and Republican lawmakers placed significant constraints on the Department of Education with the passage of the *No Child Left Behind Act*, even though there was a Bush appointee heading the agency (25). More recently, Republican members of Congress and President Trump have worked together in an attempt to place greater constraints on the Environmental Protection Agency, even though one would be hard pressed to find an EPA director (Scott Pruitt) more aligned with their goals (10). While these legislative actions seem to go against the Ally Principle, they were not out of step with the goals of Republican lawmakers at the time.

One could argue that Republican lawmakers wanted to place greater restrictions on the EPA or Department of Education as part of a vision for reducing the regulatory scope of those agencies. As Antonin Scalia put it:

> the basic goal of the Republican Party is not to govern, but to prevent the Democrats from doing so... Distrustful of government in general, and executive government in particular, they are not only less eager than their political opponents to grasp the levers of government power but are also inclined to view all impediments to the exercise of that power as a victory for their
cause. (26, p. 13, emphasis added)

If a goal of Republican legislators is to impede the functioning of the government, then we might reasonably expect them to place greater statutory constraints on the bureaucracy than Democrats, all else equal. Yet doing so is costly, as it requires more time and effort in crafting legislation, and is more likely to be met with a veto when Democrats control the presidency. A reasonable strategy for Republican legislators to adopt in this context is simply to focus more on placing statutory constrains on the bureaucracy when they have a sympathetic (Republican) President.

Thus, I expect to see a reverse Ally Principle for Republican legislators, whereby they introduce more constraining legislation under a co-partisan (Republican) President than a Democratic President.

**Hypothesis 1** Republican legislators will introduce legislation that places greater constraints on the bureaucracy under a co-partisan President.

Importantly, this expectation does not necessarily contradict our expectations under the traditional Ally Principle. A large body of literature has provided evidence that legislators from both parties will place less trust in the bureaucracy when its agency heads were appointed by a President from the opposite party. However, larger political goals (impeding the functioning of the government), coupled with the practical issues of achieving those goals, may simply prove to be more important. The key to assessing this new theoretical prediction will be accurately measuring statutory constraints in a large sample of legislation, which is the focus of the next section.

### 2.3 Measuring Statutory Constraints on the Bureaucracy

There are two prominent approaches scholars have taken to assessing statutory constraints on the bureaucracy. The first is to use human coding of legislative texts, as (9) do for several hundred “major bills” introduced between the 1950’s and 1990’s\(^2\). The

\(^2\)Due to time constraints, (9) actually code Congressional Quarterly’s year end summaries of statutes instead of the bill text itself.
authors develop a rigorous human coding form that separately measures both instances of explicit delegation of authority and various categories of statutory constraints placed on bureaucratic officials. These constraints fall into one of 14 categories under their coding scheme (e.g. spending limits, reporting requirements appeals procedures, etc.). From the author’s perspective, a constraint is an additional requirement included in a bill that is designed to ensure that bureaucrats comply with the intent of Congress.

On the other end of the spectrum, (8) use word counts in legislation as a measure of statutory constraints. The intuition here is that including more policy details requires more words, and more detailed legislation leaves less room for discretion in implementing the policy. To quote the authors:

... we argue for a simple measure of discretion in legislation. That measure is the length of legislation. We argue that with two statutes that address the same issue, the longer one typically places greater limits on the actions of other actors, because it is filled with policy-specific details that constrain what these actors can do (8, p. 45).

Huber and Shipan go on to argue that while procedural requirements (time and spending limits, rule making requirements, etc.) play a role in constraining bureaucrats (9), the amount of detail used in laying out policy specifics plays a more important role.

This word-count approach has also been employed and validated to study statutory constraints on the courts at both the state (27) and federal level (28). For example, (28) found that legislation with higher word counts contained more references to existing laws and court decisions, as well as more details on how they should be interpreted. In another vein of research, (29) used the length in California state legislation as a measure of the “complexity” of that legislation, and found that the length of a bill was strongly related to human evaluations of complexity. Taken together, these studies provide further validation of the word-count or document-length approach to measuring statutory constraints in legislation.

To make the idea that longer bills reduce bureaucratic discretion more concrete, consider the following two passages taken from the Patient Protection and Affordable Care Act. Both of these passages direct the Secretary of Health and Human Services to conduct
a study, and report the findings to Congress within two years. The first passage requires a study on expanding the healthcare acquired conditions policy, and leaves little room for interpretation:

The Secretary of Health and Human Services shall conduct a study on expanding the healthcare acquired conditions policy under subsection (d)(4)(D) of section 1886 of the Social Security Act (42 U.S.C. 1395ww) to payments made to other facilities under the Medicare program under title XVIII of the Social Security Act, including such payments made to inpatient rehabilitation facilities, long-term care hospitals (as described in subsection(d)(1)(B)(iv) of such section), hospital outpatient departments, and other hospitals excluded from the inpatient prospective payment system under such section, skilled nursing facilities, ambulatory surgical centers, and health clinics. Such study shall include an analysis of how such policies could impact quality of patient care, patient safety, and spending under the Medicare program. (Section 3008 (b) (1))

The second passage simply directs the secretary to study the benefits of screening for postpartum conditions, but gives no further detail.

The Secretary shall conduct a study on the benefits of screening for postpartum conditions. (Section 512, part (c) (2) (A))

The first passage seems to be clearly more restrictive than the second, as it includes a lot more specifics as to what is to be studied, and how. Therefore, the Department of Health and Human Services would likely have less discretion in how it went about implementing the study described in the first passage.

For this study, I decided to follow the document length approach developed by Huber and Shipan, based on their careful and extensive validation efforts, and its general applicability. This method is also highly scalable and simple to calculate, while the human coding approach proposed by Epstein and O’Halloran is impractical to apply to a large corpus. However, Huber and Shipan explicitly note that word counts are unlikely to produce valid comparisons across bills about different issues. This is because some policy areas may require more language to simply describe the proposed policy, such as a complex
tax law or health insurance regulation. Thus applying simple word counts to the full text of all congressional bills would be theoretically invalid from the start. An even more troubling potential issue with using bill length as a measure of statutory constraint is that this approach will provide inaccurate measurements for omnibus legislation, or legislation that simply covers several closely related issues (30; 31). This is problematic because we may confuse a long bill that deals with many disparate issues in minimal detail with one that deals with a single issue, but in great detail. For these reasons, Huber and Shipan restricted their original empirical analyses to a very narrow band of healthcare legislation.

To begin to address these potential issues, I split bills into their constituent sections and used these as my unit of analysis. As (30) point out, bill sections are a more appropriate unit of analysis than the entire bill if we are interested in the amount of language used to describe each policy proposal. To illustrate this point, consider the first bill introduced in Senate during the 114th session of Congress – the “Keystone XL Pipeline Approval Act”. From the title, we might assume that upon reading the bill, we will find it is mostly about allowing an oil pipeline to be built through the central United States. However, after reading it, I found that the portion of the text related to approving the pipeline was actually quite small (approximately 5%)\(^3\), and a number of other substantive provisions were included (see Figure 2.1). If I were to based my measurement of the amount of policy detail in the bill on the length of the entire bill, I might conclude that it greatly restricted the discretion of bureaucrats with regard to the pipeline approval process (because it is long); But the part of the bill approving the pipeline is actually quite short and relatively vague.

Moreover, as (32) point out, some bill sections have almost nothing to do with actual policy. Sections such as the front and back-matter, table of contents, and “effective dates” often contain no policy content at all. At the same time, U.S. Congressional legislation follows a strict set of formatting guidelines (see, for example: 33) that include standardized bill section naming and delineation conventions, making it possible to filter out bill sections containing no policy content with high accuracy.

\(^3\)The text classification was performed by reading the entire bill, and then counting the number of lines in the original text on the congress.gov website that related to each topic.
<table>
<thead>
<tr>
<th>Lines</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>Approve Keystone XL pipeline</td>
</tr>
<tr>
<td>88</td>
<td>Energy Retrofit for Schools</td>
</tr>
<tr>
<td>34</td>
<td>“Climate change is real and not a hoax”</td>
</tr>
<tr>
<td>489</td>
<td>Energy Efficiency Improvements</td>
</tr>
</tbody>
</table>

Figure 2.1: **S.1 - Keystone XL Pipeline Approval Act** proportion of bill text devoted to each purpose. Note that roughly 5% of the text (one section) is devoted to the stated intention of the bill.

If I were able to accurately split up bills into sections and identify the substantive ones (as discussed in the next section), I could use the word count in each substantive bill section as a measure of the degree of statutory constraint in that section. The longer the bill section, the greater the degree of statutory constraint. In this way I could take to heart the work of Huber and Shipan (that longer bills tend to leave less room for discretion in implementation), but apply it more broadly (because individual bill sections are not subject to the same omnibus issues as whole bills). To summarize, my unit of analysis is a substantive bill section (a bill section containing at least part of a policy proposal), and my measure of statutory constraint (or conversely lack of bureaucratic discretion in implementation) is the number of words in that substantive bill section, with longer bill sections being more constraining.
This measurement approach keeps the theoretically well-grounded, heavily validated formulation developed by Huber and Shipan, but refines it to be more precise, and more easily comparable across units. The primary measurement challenge with this approach is not in the actual counting of words (this is easily accomplished using off-the-shelf tools for text processing), but in: (a) identifying the correct bills to include in my analysis; (b) correctly splitting up each bill into its constituent sections; and (c) determining which sections are substantive and which are not. Below, I discuss my data collection efforts, and my solution to each of these challenges, as well as my validation efforts.

2.4 Data and Preprocessing

The dataset I selected for this study was a corpus of all 102,763 Congressional bills (House and Senate) introduced in the 103rd–114th sessions of Congress (January 3, 1993 – January 3, 2017). Part of this corpus (103rd–113th Congresses) was originally collected from the congress.gov website by (34), and bills from the 114th Congress were also scraped from the same website for consistency. This corpus extends back as far as data are electronically available, and as far forward as bill metadata including hand-coded topic labels are available as part of the Congressional Bills Project dataset (35). It is ideal for this study because it provides the longest time period over which to assess my theory, and a large enough sample size to make precise inferences.

2.4.1 Selecting Bills for Inclusion

This corpus covers the full presidencies of Clinton, Bush, and Obama, although bill metadata were not available for bills introduced after March, 2016. The raw dataset consisted of 102,763 unique bills. This included the text of all bills that were originally introduced in the House (IH) and Senate (IS), and for which the Congressional Bills Project contained a metadata entry. While many of these bills also had text available for a later version of the bill (such as the version that was reported to the floor, all the way through the version that became law), I only used the text from the version as introduced in this analysis. The reason for doing so was that it is unreasonable to attribute the text of a bill to the bill sponsor after it goes through the markup process, and I am primarily interested
in relating characteristics of the bill sponsor to the length of bill sections. From this raw corpus, I began to remove bill sections that met a number of exclusion criteria (discussed below), as well as bill sections I determined were non-substantive.

As mentioned above, the Congressional Bills Project maintains a database of metadata on all bills included in this study. These metadata include a “Major Topic Label” for a bill, which falls into one of approximately 20 different categories (health, defense, environment, etc.). I began by removing all bills for which this label was missing (about 2,500 bills from 2016). As I am interested in differences between the two major parties, I also removed all bills with sponsors who did not caucus with Democrats or Republicans. Furthermore, I followed (32) in removing all private bills, as these tend to deal with the relief of a particular individual or group, and are not a good comparison to regular bills. Additionally, I excluded the small number of bills with the “Culture” topic from my analysis, as these bills tended to be shorter, and deal with minor issues. The results of my analysis were not sensitive to the exclusion of these topics. I removed 3,563 bills in this way, with 99,194 remaining.

As numerous previous studies have noted, not all bills are created equal in terms of their importance. Failing to remove these minor bills could potentially confound any inferences I made about variations in bill section length, as these tend to be much shorter than regular bills. The Congressional Bills Project provides a binary important/not important label for all bills, so I filtered out all bills that were marked as not important. This resulted in 10,359 bills being discarded (spread relatively evenly across all sessions of Congress), with 88,835 remaining. Additionally, I also removed all temporary import tariff suspension bills not covered under the non-important bill tag, as the sections in these bills are short, highly standardized, and do not constitute major policy proposals. This resulted in 147 bills being discarded, with 88,688 remaining. Finally, I removed 93 bills for which the sponsor did not have a first dimension DW NOMINATE ideology score in the Congressional Bills Project metadata (this was the case for some temporary appointments). This resulted in a total of 88,595 bills remaining in the dataset.

4The data and associated codebooks are available at: www.congressionalbills.org
2.4.2 Splitting Bills Into Sections

Having selected a corpus of bills for analysis, the next challenge was to correctly split these bills into sections. The Office of the Legislative Counsel provides precise guidelines for bill formatting (33) which make it theoretically possible to use a rule-based system for splitting bills into sections. However, a combination of spelling and formatting mistakes by legislative staffers, failures to follow formatting guidelines, and difficult to disambiguate section delineators made this a challenging task. In general, I made use of a series of regular expressions for identifying section boundaries, that were then fed through several additional rule-based layers to correct for potential errors in the process. Below, I discuss the different types of sections and document delineators, and provide an overview of my approach to identifying them. The full details of this approach are documented in the accompanying software.

The input to the bill “sectionizer” was the plain text of the piece of legislation as published on the congress.gov website. The text was formatted to a standard line width, with section headers and other document delineators always starting a new line. In addition to the actual meat of the bill, there are also a number of non-substantive sections in these bills, including: Front Matter, Back Matter, Short Title, Table of Contents, Document Structure Delimiters (Title/Subtitle, Part/Subpart, etc.), Findings, Sense of Congress/Senate/House, Effective Date, Technical Corrections, Conforming Amendments, Definitions, Authorization of Appropriations, Purpose(s) sections. For a more detailed description of these sections, see A.

In practice, I developed a number of regular expressions (textual rules) for identifying the boundaries between sections, and their type, as well as other document delineators based on my reading of a sample of legislation and the Office of the Legislative Counsel’s formatting guidelines (33). These rules were then refined through an iterative process of preprocessing and hand coding of a sample of approximately 500-1,000 bill sections to find any remaining errors. In addition to hand coding a random sample, I also selected particularly short and long examples of sections of each type, and checked these for errors at each iteration. This process was repeated approximately twenty times by the author, until no further errors in bill segmentation or section classification could be detected through my validation efforts.
These validation efforts and results are described in more detail below.

### 2.4.3 Substantive Bill Sections

Having split each bill into its constituent sections, and classified each section using the taxonomy described above, I then proceeded to remove all bill sections from my corpus that I deemed to be non-substantive (as described above). On top of the non-substantive sections described in the previous section, I also removed a small number of sections I identified through my validation efforts. These included a total of 41 “Table of titles” (instead of Table of Contents), “Table of contents amendments”, “enforcement date” (similar to effective date), and back matter (listing the signatories to a bill) sections. The choice to remove these sections did not affect my results. Table 2.1 provides an overview of these section types, and whether or not they were included in my analysis. Out of 653,426 total sections in the 88,595 bills remaining in my dataset, 382,111 were deemed substantive (associated with 88,132 bills) and included in my analysis. These substantive sections were broken down into two types, with the first being regular bill sections as described above, and the second being Titles from a small subset of bills which used titles as the primary document delimiter (instead of sections).

Finally, in addition to only keeping bill sections identified as substantive (by their type), I also adopted a bill section word count threshold for inclusion as a potential substantive section in my analysis. To illustrate this choice, consider the following section which was tagged as substantive according to my rule-based classifier:

SEC. 512. PERSONS ACCREDITED TO CONDUCT INSPECTIONS.

This section would not have been classified in any of the non-substantive categories described above, yet the change it proposes is very similar to those found in conforming amendments (which I consider non-substantive). After examining a large number of short sections that were tagged as substantive by my system, I determined that the overwhelming majority of these sections containing thirty or fewer words were non-substantive. Below
Table 2.1: Descriptive statistics for bill sections by type. **# Sections** records the number of bill sections of each type, and **# Bills** records the number of unique bills associated with that section type. **Substantive?** records where each section type was considered substantive for the purposes of my analysis.

<table>
<thead>
<tr>
<th>Section Type</th>
<th># Sections</th>
<th># Bills</th>
<th>Substantive?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorization of Appropriations</td>
<td>6,575</td>
<td>5,664</td>
<td>No</td>
</tr>
<tr>
<td>Conforming Amendments</td>
<td>3,199</td>
<td>2,356</td>
<td>No</td>
</tr>
<tr>
<td>Definitions</td>
<td>17,909</td>
<td>14,854</td>
<td>No</td>
</tr>
<tr>
<td>Effective Date</td>
<td>9,036</td>
<td>7,497</td>
<td>No</td>
</tr>
<tr>
<td>Findings</td>
<td>18,687</td>
<td>17,525</td>
<td>No</td>
</tr>
<tr>
<td>Front Matter</td>
<td>88,586</td>
<td>88,586</td>
<td>No</td>
</tr>
<tr>
<td>Miscellaneous Non-Substantive</td>
<td>41</td>
<td>40</td>
<td>No</td>
</tr>
<tr>
<td>Part</td>
<td>2,584</td>
<td>338</td>
<td>No</td>
</tr>
<tr>
<td>Purposes</td>
<td>4,084</td>
<td>3,602</td>
<td>No</td>
</tr>
<tr>
<td>Sense of XXX</td>
<td>2,577</td>
<td>2,046</td>
<td>No</td>
</tr>
<tr>
<td>Short Title</td>
<td>68,122</td>
<td>64,836</td>
<td>No</td>
</tr>
<tr>
<td>Substantive Section</td>
<td>381,713</td>
<td>88,126</td>
<td>Yes</td>
</tr>
<tr>
<td>Substantive Title</td>
<td>398</td>
<td>158</td>
<td>Yes</td>
</tr>
<tr>
<td>Subtitle</td>
<td>16,752</td>
<td>2,102</td>
<td>No</td>
</tr>
<tr>
<td>Technical Correction</td>
<td>658</td>
<td>532</td>
<td>No</td>
</tr>
<tr>
<td>Title</td>
<td>31,319</td>
<td>7,390</td>
<td>No</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>1,186</td>
<td>1,183</td>
<td>No</td>
</tr>
</tbody>
</table>

In this threshold, the sections were almost exclusively short amendments or definitions, while over this threshold, I started to find short, but meaningful policy details. Therefore, I decided to remove all bill sections containing thirty or fewer words from my analysis. This resulted in the removal of 16,016 bill sections, with 366,095 (from 87,240 bills) remaining in the final dataset. The results of my analysis are not sensitive to changes in this word count threshold from 0 to 40.

Descriptive statistics for this final version of the dataset are presented in Table 2.2. Counts
are recorded for (using Congressional Bills Project “major topic” labels). The **Average Words per Section** column records the average number of words in sections whose parent bills were assigned to a given major topic label, while the **# Sections** and **# Bills** columns record the total number of sections and unique bills (respectively) associated with a given issue area. We can see that there is significant variation in average substantive bill section length across issue areas, with the average substantive section in a bill about foreign trade being the shortest (307 words), and the average substantive section in an education bill being the longest (552 words). Clearly, it will be important to control for bill issue area in my analysis.

### 2.4.4 Implementation and Validation

Before moving on to my analysis, it is important to discuss my efforts to validate my data preprocessing. As (36) note, human participation is critical to any text preprocessing or analysis task. While statistical and rule based approaches to segmenting bills into sections (for example) can be highly accurate in general, these approaches will inevitably miss unusual cases such as spelling and formatting errors, or irregularly formatted sections. Therefore, I adopted an iterative approach to identifying bill section boundaries and classifying them into the various types described in Table 2.1. Under this approach, I would split bills into sections and classify their types, hand validate my results to look for errors, and repeat.\(^5\) This process was repeated until I could not detect any errors in the validation step, and took approximately six months.

Once my validation efforts could no longer detect any errors in the process of splitting

\(^5\)For more details on this approach, see B
<table>
<thead>
<tr>
<th>Issue Area</th>
<th>Average Words per Section</th>
<th># Sections</th>
<th># Bills</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
<td>400.82</td>
<td>10,481</td>
<td>2,326</td>
</tr>
<tr>
<td>Civil Rights</td>
<td>338.58</td>
<td>8,747</td>
<td>2,096</td>
</tr>
<tr>
<td>Defense</td>
<td>321.25</td>
<td>23,367</td>
<td>5,843</td>
</tr>
<tr>
<td>Domestic Commerce</td>
<td>387.81</td>
<td>28,839</td>
<td>6,644</td>
</tr>
<tr>
<td>Education</td>
<td>551.75</td>
<td>17,084</td>
<td>4,893</td>
</tr>
<tr>
<td>Energy</td>
<td>401.36</td>
<td>21,581</td>
<td>4,063</td>
</tr>
<tr>
<td>Environment</td>
<td>488.91</td>
<td>15,162</td>
<td>4,212</td>
</tr>
<tr>
<td>Foreign Trade</td>
<td>307.12</td>
<td>12,088</td>
<td>3,209</td>
</tr>
<tr>
<td>Government Operations</td>
<td>373.52</td>
<td>25,761</td>
<td>6,581</td>
</tr>
<tr>
<td>Health</td>
<td>517.38</td>
<td>49,337</td>
<td>11,685</td>
</tr>
<tr>
<td>Housing</td>
<td>487.32</td>
<td>8,099</td>
<td>1,851</td>
</tr>
<tr>
<td>Immigration</td>
<td>373.15</td>
<td>8,983</td>
<td>1,676</td>
</tr>
<tr>
<td>International Affairs</td>
<td>305.38</td>
<td>12,943</td>
<td>2,711</td>
</tr>
<tr>
<td>Labor</td>
<td>438.67</td>
<td>15,848</td>
<td>3,692</td>
</tr>
<tr>
<td>Law and Crime</td>
<td>371.54</td>
<td>27,677</td>
<td>6,016</td>
</tr>
<tr>
<td>Macroeconomics</td>
<td>446.99</td>
<td>17,354</td>
<td>4,022</td>
</tr>
<tr>
<td>Public Lands</td>
<td>352.59</td>
<td>25,894</td>
<td>7,792</td>
</tr>
<tr>
<td>Social Welfare</td>
<td>478.69</td>
<td>12,489</td>
<td>2,357</td>
</tr>
<tr>
<td>Technology</td>
<td>356.33</td>
<td>7,300</td>
<td>1,928</td>
</tr>
<tr>
<td>Transportation</td>
<td>359.56</td>
<td>17,061</td>
<td>3,643</td>
</tr>
<tr>
<td>Combined</td>
<td>410.55</td>
<td>366,095</td>
<td>87,240</td>
</tr>
</tbody>
</table>

Table 2.2: Descriptive statistics for substantive bill sections by issue area. The last two columns display the number of bill sections, and the number of unique bills associated with a given issue area, respectively.

bills into sections and identifying substantive sections, I performed one final round of validation. I selected a random sample of 100 bills containing 701 total sections (including non-substantive sections) and hand coded each section for whether its boundaries were correctly identified, and whether it was assigned to the correct class (according to the taxonomy in Table 2.1). This coding task was very well defined (section boundaries are clear to humans, as are the classes), and was completed by the author over the course of
approximately four hours. The results of this coding exercise revealed no errors in splitting bills into sections, or in classifying them by type.

My broader experience working with these data is that the true error rate is probably closer to 0.1%. There are numerous spelling errors and non-standard formatting choices in this corpus, and it is likely that some number of these have slipped by my validation efforts. However, I have reason to believe that these errors do not substantively change the conclusions of my analysis. During the last several iterations of preprocessing refinement, I conducted the same analysis presented in this paper, but using version of the data containing various number of errors that were later corrected. The results at each stage were very similar to the results presented here, even when not excluding short bill sections, for example. Therefore, I feel confident that the results presented in the next section are unlikely to be an artifact of yet to be discovered errors in preprocessing.

I this study, I chose to follow (8) in treating policy details as a proxy for statutory constraints on the bureaucracy, and word counts as an operationalization of policy details. Therefore, the dependent variable in my analysis is the count of words in each substantive bill section. To arrive at these counts, I made use of the “phrasemachine” software developed by (34). While the primary goal of this software is to extract semantically coherent phrases from an input document, it also provides the user with the tokenized (split into individual words) text of the document. Each token is then assigned a tag for its part of speech (noun, verb, adjective, etc.). Importantly, phrasemachine also includes a “junk” tag for tokens that are not standard words (parentheses, numbers, punctuation, etc.). I removed all terms containing these junk tags before calculating word counts for each substantive bill section.
What was left over were only tokens that a human would consider real words, providing an ideal implementation of the word counts described by Huber and Shipan. Importantly, the analyses presented below are not sensitive to this choice to exclude punctuation, numbers, parentheses, etc. as tokens for the purposes of forming word counts.

2.5 Analysis

In Section 2.2, I laid out my theoretical expectations for a reverse Ally Principle for Republican legislators as they seek to collaborate with a co-partisan president to make it more difficult for the bureaucracy to operate. As for Democratic legislators, I expect to find evidence of the standard Ally Principle. To reiterate the logic of the Ally Principle, if legislators have more similar policy preferences to bureaucrats, then they should place more trust in those officials to implement policies in the way they would have intended. This should lead them to write shorter, less detailed, and thus less constraining legislation. As the President appoints agency heads, co-partisanship with the President should therefore imply greater policy preference similarity. To begin assessing my theoretical predictions, I calculated the mean number of words in all substantive bill sections introduced by Democrats and Republicans under Democratic and Republican Presidents, respectively. This is the simplest operationalization of the Ally Principle, and if it holds, we should expect that Democrats (Republicans) will introduce longer substantive bill sections under a Republican (Democratic) president than a Democratic (Republican) one, on average.

The results are presented in Table 2.3 and as we can see, substantive bill sections from Democratic legislation are approximately seven words longer on average under a Republi-
Table 2.3: Average substantive bill section length for all bills introduced by Democratic and Republican legislators under Democratic and Republican Presidents, respectively (1993-2016). Substantive bill section counts in each category are displayed in parentheses under the average word count.

<table>
<thead>
<tr>
<th></th>
<th>Democratic</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic Legislators</td>
<td>419.08</td>
<td>426.09</td>
</tr>
<tr>
<td></td>
<td>(128,718)</td>
<td>(70,694)</td>
</tr>
<tr>
<td>Republican Legislators</td>
<td>387.01</td>
<td>413.23</td>
</tr>
<tr>
<td></td>
<td>(100,769)</td>
<td>(65,914)</td>
</tr>
</tbody>
</table>

can president, consistent with the Ally Principle. However, substantive bill sections from Republican legislation are approximately twenty six words shorter on average under a Democratic President, consistent with my prediction of a reverse Ally Principle for Republicans. Figure 2.2 depicts the 95% confidence intervals for these differences in mean substantive bill section length under a opposite party President, for each party. A positive differential in this figure is consistent with the Ally Principle, while a negative differential is consistent with a reverse Ally Principle. The results in this figure indicate a non-statistically significant positive differential for Democrats, and a statistically significant negative differential for Republicans ($\alpha = 0.01$).

While these aggregate level results provide preliminary evidence of partisan difference in the Ally Principle, it is likely that they are confounded by multiple sources of heterogeneity in the characteristics of the bill sections on which they are based. To illustrate this point, consider Figure 2.3, which plots the rolling average party mean substantive bill section word count from 1933-2106. The colored bars on the bottom of the plot illustrate partisan
Figure 2.2: Difference in average substantive bill section word count based on presidential partisanship for bill sections introduced by Democrats and Republicans (1993 - 2016). A positive differential indicates that members of the party introduced longer bill sections under a President from the opposite party (consistent with the Ally Principle).

control of the Presidency, Senate, and House respectively. One obvious difference between the trends for Democrats and Republicans is that the democratic moving average displays significantly more variation over time. In particular, the average bill length for democratic bills seems to increase with Democratic control of the Senate. Furthermore, for both parties, there seems to be an increasing time trend in the average substantive bill section length.

In order to control for potential sources of latent heterogeneity among bills, I estimated a hierarchical linear model with the log of substantive bill section word counts as the dependent variable. I chose to estimate a single model for both parties as I did not expect parameter estimates for my control variables to vary by party. As I expected, for example, significant dependence among observations within the same bill, I also included random intercepts for the bill, session of congress, and sponsor. I chose to use the log of substantive bill section word counts, as the distribution over substantive bill section word counts was heavy tailed (see Figure 2.4). As these data are (over-dispersed) counts, an alternative approach would have been to model them using negative-binomial regression. I prefer the log linear model for its interpretability (percent changes in the dependent variable), but my
Figure 2.3: Moving average substantive bill section word count for Democrats and Republicans, 1993-2016. The stacked red (Republican) and blue (Democratic) bars along the bottom of the plot indicate partisan control of the Presidency, Senate, and House.

findings are robust to using negative-binomial regression. The main quantity of interest in this model was the difference in the effect of the Ally Principle for Democrats and Republicans. To measure this quantity, I included both an indicator for the party of a bill’s sponsor and an indicator for each bill (section) as to whether its sponsor was copartisan with the President. This variable (“Ally Principle”) was equal to 0 if the sponsor and President were from the same party, and 1 if they were from opposite parties. Thus, a positive coefficient in my model can be interpreted as consistent with the traditional Ally Principle (introducing longer, more constraining legislation under a President from the opposite party). To operationalize the change in the effect of the ally principle with party, I included an interaction effect between the ally principle and sponsor party indicators. //
Additionally, I included a number of controls in my regression analysis. As illustrated in Table 2.2, there is significant variation in average substantive bill section length across issue areas. To control for this variation, I included bill issue area fixed effects in my regression models. From looking at Figure 2.3 it also seems clear that there is a time trend in the length of substantive bill sections. I therefore included a time trend for session of Congress starting at one for sections from bills introduced in the 103rd Congress, and increasing to twelve for sections from bills introduced in the 114th Congress. Figure 2.3 also seems to illustrate significant differences in average bill section length by partisan control of the chamber, particularly for Democrats. I therefore included a control for whether the sponsor of the parent bill of each substantive section was in the majority party in their chamber at the time of introduction. The intuition here is that legislators may be more motivated so spend more time an energy writing (longer) legislation when it has a better chance of at least advancing our of committee in their own chamber. I also found that Senate bills tend to be longer than House bills, so I included an indicator for whether a bill was introduced
in the Senate. Additionally, I included a control for the total number of (substantive and non-substantive) sections in a bill. My reasoning was that some legislators might prefer to write fewer, longer sections into their legislation, while others might prefer to write a greater number of shorter sections into their legislation.

In addition to these factors I also included a number of controls for special types of bills (revenue and appropriations bills), as well as factors associated with the probability of a bill becoming law (see, for example: 32). For example, due to the additional technical requirements associated with revenue generating, and appropriations legislation, I included controls for whether a bill was a revenue or appropriations bill. Furthermore, if some bills are more likely to advance for procedural or institutional reasons (32), it is possible their authors will recognize this and put more effort into drafting that legislation, resulting in longer substantive sections. Therefore, I control for whether a bill was a leadership bill (the first 10 bills introduced in the House in each session of Congress), or whether it was a reauthorization bill, as both types of legislation have been shown to be more likely to become law (37). Finally, and for the same reasons, I include controls for whether a bill’s sponsor was either the chair or the ranking member of the committee to which the bill was introduced.

Regression results for both Democrats and Republicans are presented in Table 2.4, with issue area fixed effects omitted from the table for readability. Looking at the regression results, we see that the parameter estimate for the effect of the unconditional ally principle is positive but not statistically significantly different from zero ($\alpha = 0.01$). However, we also see a large negative and statistically significant parameter estimate for party-ally principle interaction effect ($\alpha = 0.01$), which is consistent with my expectation of a reverse Ally
Principle for Republicans.

As the dependent variable is on a log scale and the independent variable is on a nominal scale, we can interpret these parameter estimates as the percentage change in average substantive bill section length for that party if we were to substitute an opposite party president for a co-partisan president, all else equal. If we hold all else constant and replace President Clinton with President George W. Bush, the model would predict that a Republican legislator would introduce approximate 6.3% longer legislation (about 108 words longer on average when aggregated to the bill level). This equates to several sentences per section, and potentially dozens of pages per bill.

One potential alternative explanation for Republican legislators introducing shorter legislation under a Democratic President is that they simply introduce more position taking bills that are never intended to become law under a Democratic President. Thus what we might interpret as Republicans introducing longer legislation under a copartisan President could just be attributable to them introducing relatively fewer position taking bills because they have higher expectations about being able to enact their preferred policies. To assess this alternative explanation, I specified the same regression model but this time only for bills with above the median number of cosponsors (five or more). The intuition here is that a bill collecting a greater number of cosponsors is an indicator that it is more serious and more likely to advance, and thus less likely to be a pure position taking bill. As we see in the second column of Table 2.4, the parameter estimates are qualitatively similar even for this subset of bills. These results are also robust to increasing the threshold to 20 and even 40 cosponsors.
2.6 Discussion

The relationship between Congress and the bureaucracy has received decades of attention from political scientists, and with good reason: it is vital to understanding how policies enacted by Congress will actually be implemented. Many scholars view this relationship as one of principals and agents, with bureaucrats providing information and expertise (38), and members of Congress seeking to induce bureaucrats to implement policies in a desired way. From this perspective, divergence in policy preferences between bureaucrats and legislators is the driving force behind legislators’ decisions about the statutory constraints they place on the bureaucracy. This literature has captured an important dimension of strategic interaction between Congress and the bureaucracy, but has paid less attention to partisan political factors potentially influencing these decisions.

In this study, I extended the theory of demand for statutory constraints on the bureaucracy, to include partisan political goals, such as reducing the scope of government. In particular, I posited a reverse Ally Principle for Republican legislators, as they seek to introduce more constraining legislation under a co-partisan President as a way to hinder the functioning of the bureaucracy. To assess the predictions of my new theory, I implemented and validated a new preprocessing pipeline for legislative texts, which I then used as input to a text-based measure of statutory constraints at the substantive-bill-section level. Finally, I assessed my theoretical predictions using a regression framework, and found support for my hypothesis of a reverse Ally Principle for Republicans.

This finding is important because it goes against decades of research on the Ally Principle
in Congress. While I find evidence in support of the traditional Ally Principle of Democrats, my results suggest one strong similarity between both parties: members of both parties introduced legislation containing longer substantive bill sections under George W. Bush than either Clinton or Obama, all else equal. While it would be ideal to extend my dataset to include more Republican Presidents, future research should also explore additional potential explanations for this regularity. Perhaps if Republican legislators are more likely to pursue an obstructionist legislative agenda under a Democratic President, members of both parties may simply expend less energy drafting legislation if they feel it is likely to simply be filibustered or stonewalled in the House or Senate. Going further with this finding, it would be interesting to discover any regularities in the types of policy details that are being added to Republican legislation under a Republican President.

More generally, my results suggest that a plurality of factors determine legislators’ use of statutory constraints. I have provided evidence that partisan policy goals may play a role in determining the use of statutory constraints, but other institutional factors may also play an important role. For example, future work should explore other incentives (such as the likelihood of a bill advancing) for legislators to write legislation containing more or less statutory constraints. Finally, partisan differences in the use of statutory constraints may vary across issue areas. Future work could develop agency or issue-specific predictions about the use of statutory constraints by each party and empirically examine the Ally Principle at a more fine-grained level.

There are also several aspects of this study that could be strengthened in future work. I only consider one half of the relationship between bureaucrats and members of Congress
(constraints), and not instances of explicit delegation of authority. Furthermore, I only consider legislation introduced during a 24 year period (1993-2016). It would be valuable to extend the study back further to include more Republican Presidents, and forward to the Trump administration. It would also be valuable to try more nuanced methods of assigning credit (and thus partisanship) to bill sections (perhaps account for initial cosponsors, or hitchhiker legislation). Finally, it would be valuable to buttress the claims I make with congressional staffer and agency official interviews and case studies, to better understand how the macro level trends I identify relate to the experiences of these bureaucrats.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>All Bills</th>
<th>5+ Cosponsors</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(Section Word Count)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sponsor Republican</td>
<td>-0.036</td>
<td>-0.030</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>Ally Principle</td>
<td>0.017</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>Republican × Ally Principle</td>
<td>-0.080*</td>
<td>-0.089*</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>Sponsor in Majority</td>
<td>0.037*</td>
<td>0.041*</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Senate Bill</td>
<td>0.087*</td>
<td>0.082*</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Number of Sections</td>
<td>-0.001*</td>
<td>-0.001*</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>Congress (ordinal)</td>
<td>0.014*</td>
<td>0.010*</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Revenue Bill</td>
<td>0.160*</td>
<td>0.152*</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Appropriations Bill</td>
<td>-0.172*</td>
<td>-0.236*</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>Reauthorization Bill</td>
<td>-0.026</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>Leadership Bill</td>
<td>0.165</td>
<td>0.136</td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td>(0.088)</td>
</tr>
<tr>
<td>Sponsor Committee Chair</td>
<td>0.013</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Sponsor Committee Ranking Member</td>
<td>0.023</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>Constant</td>
<td>5.285*</td>
<td>5.361*</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.028)</td>
</tr>
</tbody>
</table>

Observations: 366,095 170,135
AIC: 1,048,576 492,947

Note: *p<0.01

Table 2.4: Regression results for all bills, and bills with at least five cosponsors. Dependent variable is the log of word counts for substantive sections from bills introduced between 1993-2016. Issue area fixed effects were mostly statistically significantly different from zero (α = 0.01), and are omitted from these results for clarity. They are included in C. Parameter estimate standard errors are presented in parentheses under the estimate. Note that the (2) correction was used to calculate the standard errors for the interaction terms.
Chapter 3  
Title of the Second Chapter

3.1 Introduction

The overwhelming majority of bills introduced in the U.S. Congress never get much consideration from the chamber. Of the approximately 86,000 non-minor, non-appropriations bills introduced in a House or Senate committee between 1993 and 2016, less than 10% made it out of committee, and only about 3% became law. As a result, the conventional wisdom suggests that Congress is broadly unproductive (39).

Scholars have long recognized these challenges and have devoted significant attention to understanding why some bills succeed and others fail. Some have argued that constraints on the legislative agenda mean that certain bills, such as reauthorizations or leadership bills will be much more likely to become law because they are institutionally advantaged (e.g. 40). Others have shown that support from outside interest groups can help push a bill through the legislative process, while opposition can keep a policy idea from making it out of committee (e.g. 41; 42; 43). Yet another strain of research has demonstrated
the importance of characteristics of the coalition of legislators involved in drafting and advancing a bill, such as securing bipartisan support, to its chances of advancing (e.g. 44; 45; 46; 47; 48; 49). Finally, a large body of research has explored the relationship between the size of a bill’s co-sponsoring coalition, and its chances for success (see, for example: 50; 51; 52; 53; 54; 55).

Yet there is another perspective that suggests that many legislators simply are not focused on the legislation they introduce actually becoming law, and are instead primarily interested in legislation as a position taking opportunity (44). This is exemplified by the now famous quote from of Senator Carl Hayden’s advice to new members of Congress:

There are two kinds of Congressmen—show horses and work horses. If you want to get your name in the papers, be a show horse. If you want to gain the respect of your colleagues, keep quiet and be a work horse (44, p. 94).

Extending this perspective beyond legislators to bills themselves, we can also find examples of purely position-taking legislation being introduced by legislators who also introduce legislation that is intended to become law. This poses a fundamental problem for all studies of lawmaking: how are we supposed to understand the determinants of legislative success when a (potentially) large number of bills are not intended to succeed from the get-go?

In this study, I begin to lay the theoretical and methodological foundation for understanding the seriousness of the lawmaking effort encapsulated in a bill by examining the text of the bill itself. I follow a simple intuition, that legislators will devote more time, effort, and staffing resources towards drafting bills that are intended to become law, and relatively less resources towards drafting legislation that is intended primarily as a position taking
instrument. I operationalize this intuition as the amount of legal and technical boilerplate language (relatively standardized legal jargon meant to “glue” together and clarify the substantive policy ideas in a bill) included in the substantive provisions of a bill—what I refer to as the amount of *legal detail* in those provisions, and develop a new statistical method to identify this language.

To bolster this analysis, I also explore the use of non-substantive provisions in legislation, such as *definitions, findings,* and *technical corrections*. These provisions provide context for interpreting and implementing the substantive policies contained in a piece of legislation. I argue that these two types of legislative text (legal details and non-substantive provisions) are important because they indicate that legal and staffing resources were devoted to drafting the bill—and by extension that it was intended as a more serious law-making effort. As a result, *I expect that the initial versions of bills containing more of both types of legal and technical content will be more likely to advance out of committee, and ultimately to become law.*

Measuring the amount of legal and technical language and non-substantive provisions in a bill is a challenging endeavor. Existing empirical investigations of legislative text that sought to (remove) legal and technical language and non-substantive provisions have relied on hand curated lists of terms and ad-hoc rules for identifying this language. Such an approach is insufficient for my purposes because I seek to precisely identify the universe of legal and technical language and non-substantive provisions in a bill, and not just those terms that interfere with an analysis of substantive language. To this end, I introduce new approaches to identifying both types of language in legislation. Specifically, I extend a
concept from information theory called *mutual information*, and use it to find the statistical signature of legal language in legislation. To identify non-substantive provisions in bills, I use the legislation parser from (56), and develop a typology of non-substantive sections. Both methods are also subjected to human validation efforts, which indicate that the computational methods I employ bear similarity to human judgements.

To assess my theory that the inclusion of legal details in legislation signals the seriousness of the lawmaking effort contained in that piece of legislation, I analyze a corpus of approximately 86,000 non-minor, non-appropriations bills introduced in the U.S. Congress between 1993 and 2016. I find that the amount of legal detail included in the substantive provisions of a piece of legislation is a strong positive predictor of a bill seeing any action beyond committee (the first major hurdle a bill must clear in the U.S. Congress), and of it ultimately becoming law. I also find that the initial versions of bills that advance out of committee (and those that eventually become law) tend to contain a greater number of several types of non-substantive provisions than those that do not, on average. Importantly, I find that the main effect of both types of legal and technical language is on the probability a bill advances out of committee, indicating that bill “quality” is most important as a signal early in the legislative process. Taken together, I believe my findings provide evidence that bill quality, operationalized as the amount of legal and technical detail included in legislation, is an important factor associated with bill advancement. By extension, I argue that the amount of legal and technical detail included in legislation is therefore a reasonable proxy for the seriousness of the lawmaking effort represented by a bill.

This study has broad methodological and theoretical implications. The method (and
accompanying software) I develop for identifying legal and technical jargon in legislation is broadly applicable to settings where scholars have document-level metadata, and want to identify domain stopwords\(^1\) in their corpus. For example, scholars often apply topic models (see, for example, \(36; 57\)) to try and understand the topical themes in a collection of political texts, but domain stopwords can make the output uninterpretable. Theoretically, my approach to characterizing the amount of legal and technical detail included in legislation represents a first step towards accounting for variation in bill quality for scholars studying legislative success. My findings also build on the growing body of work around the textual signature of unorthodox lawmaking (see, for example, \(58; 30; 32\)), and productivity of Congress \((39; 59; 60)\). I do so by highlighting another dimension (legal and technical content) on which successful lawmaking efforts differ from unsuccessful ones, emphasizing the need to rethink traditional measures of legislative productivity and effectiveness.

### 3.2 Theory and Hypotheses

“I love these members, they get up and say, ‘Read the bill,’ ” said Rep. John Conyers. “What good is reading the bill if it’s a thousand pages and you don’t have two days and two lawyers to find out what it means after you read the bill?” – John Conyers (D-MI, qtd. in \((61)\)).

Rep. John Conyers (D-MI) was famously quoted as saying that most members of Congress do not (carefully) read the bills that they vote on. And, if they did, he claimed, most legislators would be unlikely to understand the language without the help of a team of lawyers.\(^2\) Most serious legislative proposals are challenging to read because they are

\(^1\)Very common words such as “the” and “if” that tend to obscure patterns in substantively interesting words.

\(^2\)https://sunlightfoundation.com/2009/07/27/rep-conyers-dont-read-the-
composed of complicated, technically written provisions that seek to clarify and codify the intentions of the authors. While this technical language makes bills difficult to parse, it is incredibly important. After all, bureaucrats, judges, and other government officials tasked with implementing the legislation must be able to determine precisely what they are required to do. Thus, far from purely satisfying technocratic legislative drafting standards, this language can be critical to ensure that a law is implemented as its authors intended and that it can survive legal challenges.

Put another way, words matter, even if they seem unrelated to the actual substance of a policy proposal. For example, a key provision of the 900-page Affordable Care Act was challenged throughout the federal judicial hierarchy, including the Supreme Court, over a four-word drafting error.\(^3\) At the same, bills that do not contain any of the requisite legal and technical content to integrate with existing laws are generally dismissed as position taking stunts. A recent example was a one-sentence bill introduced by Rep. Matt Gaetz (R-FL) that would have abolished the EPA, effective December 31, 2018.\(^4\) This bill was generally derided in the press for ignoring all of the complex legal issues that would arise from abolishing the EPA, making it completely impractical for Congress to enact, even for those legislators who generally agreed with its sentiment. In a similar vein, (62) shows that many legislators introduce overly broad legislation to curb the federal judiciary with little intention that these bills become law.

---


Surprisingly, despite the fact that the text of legislation is what must be interpreted and implemented as a bill becomes law (63), scholars have devoted relatively little attention to understanding lawmaking in terms of the bill text itself. While some have focused on the content of legislation in a particular issue area (e.g. 8) and others have taken a bird’s eye view of the topical diversity of the legislative agenda (e.g. 64; 39; 59), our understanding of lawmaking is still mostly divorced from the words that eventually become law. Furthermore, where scholars have used text as data methods to study lawmaking, they tend to remove legal and technical language as a preprocessing step, treating the very language that bureaucrats and judges will actually seek to interpret as a hindrance to their analyses (65; 30; 32).

While the legal and technical minutiae of legislation are certainly not headline-grabbing, legislative drafting experts recognize how critical they are to a bill actually encapsulating the policy ideas it contains. This is because such language is required for a bill to interface with the existing body of law, and to meet the formatting requirements for legislation set out by the Office of the Legislative Counsel (63). As Strokoff puts it: “Attorneys are charged with taking the idea of any Member or committee of the House of Representatives requesting the services of the (Office of the Legislative Counsel) and transforming it into legislative language or, as one of my clients used to say, ‘the magic words.’” If the inclusion of this legal detail in legislation is necessary for a bill to become law, then the quality of these details should be important indicator as to whether a bill was intended advance through the legislative process (and whether it actually will).

Put another way, it seems reasonable that the inclusion of legal detail in a bill may serve
as an indicator of bill “quality”. Because including this detail requires the staffing and legal resources of the authoring coalition, I expect that legislators will devote more of these resources towards bills they are more heavily invested in becoming law. For example, the authors of the “Healthy Forests Restoration Act of 2003” (a bill that became law in 2003), spent almost four pages defining what constitutes a community that is “at risk” for wildfires, for the purposes of funding under the bill. This attention to detail beginning with the introduced version of the bill could be seen as an indication that it was a “serious” lawmaking effort, and worthy of further consideration beyond committee. Contrast this with the one-sentence position taking bill (to terminate the EPA) introduced by Rep. Matt Gaetz (R-FL) in early 2018, that he likely could have written by himself, in ten minutes, with minimal aide from his legal staffers. What becomes clear by reading the text of as-introduced versions of successful bills is that they tend to contain lots of legal detail, supporting the idea that legal detail is an indicator of quality.

Another possible explanation for this devotion of resources towards including legal detail in the early versions of successful bills is that the bill’s authors may believe the bill is likely to advance for institutional or procedural reasons. For example, when a bill is sponsored by a committee chair, party leadership, or is a reauthorization bill, members of Congress may deem it effective to devote their limited staffing resources towards that bill. One prominent example of such a bill was the USA PATRIOT Act. The introduced version of this bill spanned 341 pages and contained deliberately broad, yet complex legal language granting security agencies wide ranging powers to pursue individuals they deemed to be threats to

---

the United States. This represented a significant outlay of staff resources, but the bill’s authors had strong reasons to believe that it would become law, due to the recent 9/11 terrorist attacks driving demand for a response from Congress.

While it is difficult to empirically distinguish between a quality-signaling, and a beliefs-based explanation for the inclusion of legal detail in successful legislation, the end results should be the same. In either case, we should expect that the early versions of these more serious lawmaking efforts will contain more legal detail than the early versions of bills primarily intended as position taking vehicles, for example. Regardless of the direction of causality (legal detail → success, beliefs about success → legal detail), I still expect that legal detail in the text of a bill should be an important predictor of whether it will advance out of committee (see serious consideration in Congress), and whether it will ultimately become law. In formulating my hypothesis, I focus more specifically on the substantive policy provisions of serious (successful) lawmaking efforts (those provisions that actually spell out what the bill does, as opposed to justifying why the bill is needed, for example) as containing more legal detail than the substantive policy provisions of less serious bills.

**Hypothesis 2** Bills containing more legal detail in their initial versions will be more likely to advance out of committee, and to eventually become law.

In addition to the use of legal detail to clarify the meaning of substantive policy proposals, a bill can also include a number of non-substantive provisions that clarify the intent and implementation of a policy. Most successful bills contain a number of these non-substantive provisions, which typically serve to either clarify the meaning of terms or concepts in the substantive provisions of the bill, or provide some justification for why the policies included in the bill are a good idea. To give a sense of the types of provisions contained in
legislation, Table 3.1 provides descriptive statistics of the types of sections found in the bills I analyze in this study. These counts are drawn from the replication data provided by (56), who develops an algorithm for segmenting U.S. Congressional bills into their constituent sections.

<table>
<thead>
<tr>
<th>Type</th>
<th># of Provisions</th>
<th># of Bills</th>
<th>Substantive?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorization of Appropriations</td>
<td>6,363</td>
<td>5,485</td>
<td>No</td>
</tr>
<tr>
<td>Conforming Amendments</td>
<td>3,166</td>
<td>2,326</td>
<td>No</td>
</tr>
<tr>
<td>Definitions</td>
<td>17,283</td>
<td>14,331</td>
<td>No</td>
</tr>
<tr>
<td>Effective Date</td>
<td>8,860</td>
<td>7,349</td>
<td>No</td>
</tr>
<tr>
<td>Findings</td>
<td>18,349</td>
<td>17,234</td>
<td>No</td>
</tr>
<tr>
<td>Front Matter</td>
<td>85,941</td>
<td>85,941</td>
<td>No</td>
</tr>
<tr>
<td>Purposes</td>
<td>4,000</td>
<td>3,533</td>
<td>No</td>
</tr>
<tr>
<td>Sense of House/Senate/Congress</td>
<td>2,499</td>
<td>1,974</td>
<td>No</td>
</tr>
<tr>
<td>Short Title</td>
<td>66,571</td>
<td>63,445</td>
<td>No</td>
</tr>
<tr>
<td>Substantive Provision</td>
<td>372,729</td>
<td>85,945</td>
<td>Yes</td>
</tr>
<tr>
<td>Substantive Title</td>
<td>104</td>
<td>93</td>
<td>Yes</td>
</tr>
<tr>
<td>Technical Correction</td>
<td>618</td>
<td>497</td>
<td>No</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>1,118</td>
<td>1,117</td>
<td>No</td>
</tr>
<tr>
<td>Combined</td>
<td>587,601</td>
<td>85,949</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Counts of provisions by type. **# of Provisions** indicates the number of provisions of a given type, while **# of Bills** records the number of unique bills containing a provision of that type. **Substantive?** indicates whether each provision type was considered to be a substantive policy provision. Provision types highlighted in blue are the non-substantive provisions I focus on for the purpose of my analysis.

While the bulk of the provision in these bills are substantive provisions or titles (e.g. “TITLE III: Improving The Nutritional Value of School Lunches ...”), the bills I examine in this study also contain a wide variety of non-substantive provisions. In particular, I focus on the inclusion of seven of these types of non-substantive provisions as potential indicators
of bill quality. These provisions are highlighted in blue in Table 3.1, and fall into two broad groups: provisions that clarify or specify what other provisions in the bill (or existing laws) do, and provisions that clarify the views of a bills’ authors. Before discussing these provisions, note that I do not consider four additional types of non-substantive provisions (Front Matter, Short Titles, Tables of Contents, Effective Dates), because they are typically short, and highly standardized across bills.

Authorizations of Appropriations, Conforming Amendments, Definitions, and Technical Corrections all serve to clarify the interpretation of other provisions in the bill (or existing laws). For example, Authorizations of Appropriations specify the dollar amounts to be allocated to pay for programs described in the substantive provisions of a bill, while Definitions sections typically clarify the meaning of terms that are used across numerous substantive provisions in a bill. Similarly, Conforming Amendments, and Technical Corrections alter everything from formatting and spelling errors in other bills or laws, to changing the content of their substantive provisions. What these types of provisions share in common is that they interface with other substantive provisions (both in the same bill, and in other bills), and serve to clarify and or specify the meaning of these provisions.

Findings, Purposes, and Sense of House/Senate/Congress provisions are non-binding, and do not directly interface with the substantive provisions of a bill. Yet I argue that they still have an important political purpose in justifying the policy proposals included in a

---

8Note that (56) classifies unusually long Authorization of Appropriations sections as substantive provisions because they typically mix a specification of how much is to be spent with a description of how it can be spent.

9Note that substantive provisions may also precisely define terms within the provision, but that these definitions are typically less germane to the rest of the substantive provisions in the bill.
bill, and expressing the views of those involved in supporting the bill. *Findings* provisions typically report the conclusions of academic, government, or industry research on the issues addressed in a bill. For example, a bill seeking to address opioid addition might include a findings section that describes the scope of the public health problem, and what researchers have determined to be effective solutions. Similarly, *Purposes* provisions typically spell out in plain English what a bill’s authors intend to accomplish with that legislation. Finally, *Sense of House/Senate/Congress* offer a platform for Members of Congress to register their opinions on issues (usually) related to the policy content of a bill. In particular, these provisions can be added strategically to a bill as a (non-binding) form of political compromise.

A prime example of the use of *Sense of House/Senate/Congress* provisions as a mechanism for (soft) political compromise was the Keystone XL Pipeline Approval Act (114-S-1), that was introduced by Sen. John Hoeven (R-ND) in 2015. This controversial bill was eventually vetoed by President Obama, but would have allowed for the construction of a pipeline connecting Canada’s tar sands with the Gulf of Mexico. The bill was supported by all Republicans in the Senate, and a number of Democrats in oil-producing states, but was controversial due to the environmental impact of burning fossil fuels, and the impact of the pipeline on the communities it was supposed to pass through. As an apparent compromise, the bill contains two non-binding *Sense of the Senate* provisions, one of which “Expresses the sense of the Senate that climate change is real and not a hoax”, and the other of which encourages Congress to pass an excise tax on oil derived from tar sands. Thus, adding provisions of this type may be seen as a form of political compromise, and should be positively associated with bill advancement.

---

10 https://www.congress.gov/bill/114th-congress/senate-bill/1
To summarize, my argument is that the inclusion of these non-substantive provisions can serve as a mechanism to clarify and constrain the scope of policy proposals, and as a non-binding mechanism for political compromise. I therefore expect that the inclusion of any of the seven types of provisions discussed above is another indicator of the seriousness of a lawmaking effort, and the likelihood a bill will advance out of committee and eventually become law. Here again, it could be that the bill’s authors include these non-substantive provisions to head off opposition, or because they believe that a bill is likely to advance for institutional or procedural reasons, and therefore put more effort into ensuring its smooth passage.

**Hypothesis 3** Bills containing more non-substantive provisions in their initial versions will be more likely to advance out of committee, and to eventually become law.

### 3.3 Data and Methods

In order to assess the hypotheses laid out in the previous section, I needed to: (1) develop a measure of the amount of legal detail included in the substantive provisions of a bill; and (2) to count the number of non-substantive sections in that bill. Fortunately, I was able to rely on the labels assigned by (56) as a basis for (2), so my main methodological challenge was to measure the amount of legal detail included in the substantive provisions of a bill. Before I discuss my measurement approach, I begin by discussing the corpus of congressional bills I use, and how I preprocessed them.

The bill text data for this study were collected by (34; 56) and include all versions (intro-
duced, reported, etc.) of all bills introduced during the 103rd-114th sessions of Congress (1993-2016). These text data were also linked to bill-level metadata compiled by the Congressional Bills Project (35). Using the algorithm developed by (56), each bill was also broken up into its constituent provisions (sections) and each section was tagged for its type (front matter, substantive provision, definitions, purposes, etc.), so that I could determine which types of non-substantive sections each bill contained. The full dataset includes almost a million individual sections from 117,910 versions of 92,660 unique bills, containing over 300 million words.

Before analyzing these data, I decided to exclude four types of bills from my analysis. The first type were “minor bills”, as coded by the Congressional Bills Project (35). These bills typically deal with issues like renaming post offices or creating commemorative coins, and are different enough from the majority of policy proposals that I feel they should not be considered together. Additionally, because these bills are less controversial, they tend to become law at a higher rate than non-minor bills. I also exclude all private bills, because they also tend to deal with more minor issues (such as the relief of an individual), and are not a good comparison to legislation like the Affordable Care Act. Next, I follow (32) in excluding all bills referred to the House or Senate Appropriations Committees from my analysis because some of these bills are essentially must-pass legislation (to keep the government running), and therefore are unlikely to be subject to the same dynamics in terms of their chances for success. Finally, I exclude bills that were not originally introduced in either a House or Senate committee (for example, Joint Resolutions), as I seek to compare bills with similar starting points. Therefore, out of an initial corpus of 92,660 bills, I am left with 85,949 after applying these filters.
After applying the filters discussed above, I was left with the full text of 85,949 bills that had been split into their respective provisions by (56). My next step was to preprocessing my data into a document-term matrix (counts of each unique term in each provision), because the method I developed (described below) for identifying the legal and technical language in substantive provisions operates on this matrix. The standard approach to forming a document-term matrix is to use individual words (unigrams) as the terms (36). However, unigrams can often have an ambiguous meaning in political texts. For example, the term “section” could refer to “section 22(a)” of a bill, or “section 8 housing”.

To address this issue, I chose to use phrases instead of unigrams as the terms in my document term-matrix. A phrase is a coherent multi-word expression such as “wildlife preservation”, or “prohibit gun sales” that provides additional context for each individual word by keeping them in sequence. In addition, much of the legal and technical language I found while reading legislation tended to come in the form of multiword expressions (e.g. “establish a committee”, “subparagraph (b)”). I extracted phrases from the text of each bill section using the phrasemachine R package (34). These phrases were limited to a maximum length of 3 constituent words, as this length produced highly interpretable phrases while limiting the number of unique terms to several million.\(^\text{11}\) This process resulted in a document term matrix containing more than 200 million terms, with about 5.8 million unique terms.

\(^{11}\) Phrases were extracted using the “PhrasesNoCoord” grammar included in phrasemachine. I also limited the n-gram size to three in order to avoid too much double counting of terms in the vocabulary (e.g. “red car”, “shiny red car”, “new shiny red car”).
3.3.1 Identifying Legal Detail in Legislation

At a high level, we can think of bills as being composed of two types of language: policy language (what the bill does), and legal details (the legal and technical “glue” that holds the policy language together and clarifies its meaning).\(^{12}\) To a human coder, legal details are relatively easily distinguishable from policy language. But human coding is not readily scalable to a corpus containing tens of thousands of documents. I take a computational approach to identifying legal details in legislative texts based on their statistical signature. The method I develop runs in only a few minutes on a standard laptop, allowing me to quickly and accurately classify these terms.

This approach is based off of the following intuition. In the U.S. Congress, we expect that Democrats and Republicans will have different policy goals, so policy language should be used differently by members of each party. What I expect to remain relatively constant (between parties, within an issue area), is the use legal and technical language (legal details). If a member of Congress wants to place a number of restrictions on how money may be spent by an agency, or lay out a detailed rule-making process, or spell out the specifics of a grant program, their staff, or lawyers working for the Office of the Legislative Counsel will use standard legal and technical language to do so (63). This uniformity is reinforced by the standard training most law students receive in crafting the legal language that goes into legislation (see, for example 67).

To formulate my claim more precisely, I expect that: on average, in a given issue area, Democrats and Republicans will use legal and technical language in a similar way. Put

\(^{12}\)See (66) for a more in-depth discussion.
another way, we would say that if we are given information about the use of a legal or technical word or phrase across documents, it will not help us distinguish which party introduced it. For example, if we are given information about the number of times the phrase “described in section” occurs in each document in a corpus of Congressional bills, we are no closer to knowing what a bill that has a high or low count of this phrase is about, or which party introduced it. Building off of this intuition, I take a statistical approach to identifying which terms give us no meaningful information about the partisanship of a bill’s authors.

3.3.2 Using Contingency Tables to Classify Boilerplate Terms

The most common representation of text data in social science applications is as a document-term matrix, where each row represents a document, and each column represents a unique term in the vocabulary (36). Entries in this matrix then record the count of term \( j \) in document \( i \). The rows of this matrix (documents) can then be collapsed over various combinations of metadata attributes to form a contingency table. Thus, I started with the document term matrix described above, and used it (along with the Congressional Bills Project metadata) to create a contingency table where each row recorded the counts of terms in bills sponsored by members of a given party, in a given minor issue area (using the Policy Agendas Project minor topic codes). More specifically, in this study I formed a contingency table recording the number of times each term was used in Democrat and Republican sponsored bills in each of about 240 minor issue areas (e.g. “Health: Drug Industry” or “Environment: Drinking Water”) as coded by the Congressional Bills Project. Thus the full contingency table contained approximately 480 rows (240 issues \( \times \) 2 parties), and about 5 million columns (the number of unique terms in the corpus). A small example
<table>
<thead>
<tr>
<th>Category</th>
<th>“striking paragraph”</th>
<th>“opioid addiction”</th>
<th>“nuclear power”</th>
<th>“Affordable Care Act”</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat, Nuclear Energy</td>
<td>1,023</td>
<td>0</td>
<td>123</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Republican, Nuclear Energy</td>
<td>826</td>
<td>0</td>
<td>415</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Democrat, Health Insurance</td>
<td>956</td>
<td>124</td>
<td>0</td>
<td>36</td>
<td>...</td>
</tr>
<tr>
<td>Republican, Health Insurance</td>
<td>1,452</td>
<td>141</td>
<td>0</td>
<td>285</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 3.2: Each row in this example contingency table records the count of each unique term aggregated across all bills in a particular category (e.g. Democrat sponsored bills, about nuclear energy policy).

Intuitively, some columns (vocabulary terms) of this contingency table will give us more information about which row we are in. In the example in Table 3.2, the phrase “Affordable Care Act” is used much more frequently in Republican sponsored healthcare legislation than in Democrat sponsored legislation on the same issue. Thus, if we came across the term “Affordable Care Act” in a bill section, it would be reasonable to guess that the bill section came from a Republican sponsored bill about healthcare.

If we instead came across the phrase “striking paragraph” in a bill section, it would be much more difficult to make an educated guess about which category the bill section belonged in, because “striking paragraph” is used relatively similarly by both Democrats and Republicans. Therefore, to identify legal and technical terms, I needed to identify terms that were not associated with any particular party, within an issue area.
### Table 3.3: An example joint distribution over parties and terms.

<table>
<thead>
<tr>
<th>Category</th>
<th>“striking paragraph”</th>
<th>“opioid addiction”</th>
<th>“nuclear power”</th>
<th>“Affordable Care Act”</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat, Health Insurance</td>
<td>0.10</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>...</td>
</tr>
<tr>
<td>Republican, Health Insurance</td>
<td>0.15</td>
<td>0.01</td>
<td>0.00</td>
<td>0.03</td>
<td>...</td>
</tr>
</tbody>
</table>

#### 3.3.3 Term Contributions to Mutual Information

The approach I take to identifying legal and technical terms is based on the average contribution each term makes to a quantity called the mutual information \(\text{mutual information}\) (68) between terms and categories (Democrat and Republican bills, about a given issue). For a review of related approaches, see Appendix D. In practice, this involves considering the relative number of times each term is used by members of each party, in bills from each issue area. More concretely, I pulled out pairs of rows from the contingency table described in the previous section recording the count of terms in bills sponsored by Democrats and Republicans, respectively, on a given issue area. For example, on such pair of rows might record term counts from Democrat and Republican legislation about health insurance. Finally, each cell in the resulting \(2 \times \sim 5\) million contingency table was divided by its sum, forming a joint distribution over terms and categories:

Mutual information is a measure of the degree of statistical association between two discrete random variables. In other words, it is a number that characterizes how much information the columns of a joint distribution (vocabulary) give us about which row we are in (class), and vice versa. More formally, the mutual information of two discrete random variables \(C\)
and \( V \) is defined as:

\[
I(C; V) = \sum_{c \in C} \sum_{v \in V} p(c, v) \log \left( \frac{p(c, v)}{p(c) p(v)} \right)
\]  

(3.1)

In this equation, \( p(c, v) \) is the joint probability of observing \( c \) and \( v \), and \( p(c) \) and \( p(v) \) are the marginal probabilities of observing \( c \) and \( v \) respectively (69). Looking at the form of this equation, we see that the mutual information of a joint distribution will increase as we include more terms that distinguish between categories, and decrease as we include more terms that make it harder to distinguish between categories. This result is not intuitive, so I provide several illustrations below.

Table 3.4 provides examples of mutual information calculated on two toy joint distributions. Distribution 1 places uniform probability on all entries. If we calculate the mutual information of this distribution, we see that each term will involve multiplying by the log of 1 (which equals zero), so the mutual information will be zero. This makes intuitive sense, because if we come across a term in a document belonging to one of these categories, we can do no better than chance at guessing the category, because knowing the term was in the document gives us no new information. Now, if we consider Distribution 2, we see that the terms distinguish perfectly between categories. This is reflected in the higher mutual information for this joint distribution.

With this intuition in mind, we can now illustrate the core property of mutual information on which I rely to identify terms that make it harder to distinguish between categories. Distribution 1 in Table 3.5 includes three terms, two of which seem to distinguish between
Table 3.4: Example mutual information calculations for two joint distributions

<table>
<thead>
<tr>
<th></th>
<th>Distribution 1</th>
<th>Distribution 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“pursuant to section”</td>
<td>“fiscal year”</td>
</tr>
<tr>
<td>Democrat</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>Republican</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

\[ I(C;V) = 4 \times (0.25 \times \log(1)) = 0 \]

<table>
<thead>
<tr>
<th></th>
<th>Distribution 1</th>
<th>Distribution 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“repeal Obamacare”</td>
<td>“carbon tax”</td>
</tr>
<tr>
<td>Democrat</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td>Republican</td>
<td>0.50</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\[ I(C;V) = 2 \times (0.5 \times \log(2)) = 0.693 \]

Table 3.5: Example mutual information scores for a distribution with and without the first column included.

<table>
<thead>
<tr>
<th></th>
<th>Distribution 1</th>
<th>Distribution 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“section”</td>
<td>“birth control”</td>
</tr>
<tr>
<td>Democrat</td>
<td>0.36</td>
<td>0.08</td>
</tr>
<tr>
<td>Republican</td>
<td>0.20</td>
<td>0.02</td>
</tr>
</tbody>
</table>

\[ I(C;V) = 0.11 \quad \text{I}(C;V) = 0.428 \]

categories while one (“section”) does not. Distribution 2 in Table 3.5 was based on the same raw term counts, but with the “section” column removed. We can see that when we remove “section” from the vocabulary in this toy example, the mutual information of the joint distribution increases substantially. This again makes intuitive sense, because we have removed a lot of information (terms that occurred with high relative frequency) which made it harder to tell from a randomly selected term which category we were in. Thus we could say that in this toy example, the term “section” made a mutual information contribution of 0.11 - 0.428 = -0.317.

I use these mutual information contributions for each term as a way to quantify how much
information they give us about whether a Democrat or Republican wrote a particular bill (in a given issue area). In order to identify legal and technical terms I simply average the mutual information contributions of a term in each issue area by repeating the process of calculating its mutual information contribution for each issue area. I refer to this average over mutual information contributions as a terms’ Average Contribution to Mutual Information (ACMI). A formal definition of the ACMI for a given term is provided in Appendix E.

Finally, following from the intuition laid out earlier, if a term has a negative ACMI, I classify it as a legal/technical term.

**Definition 1** A legal/technical term is a term whose ACMI is less than zero.

To some readers, the ACMI approach described will seem quite similar to many other methods for calculating statistical associations on contingency tables. Why not use a $\chi^2$ test statistic, or any one of dozens of other measures of statistical association, instead of ACMI? The primary difference between ACMI and more standard test statistics is that ACMI is self-consciously relative. In other words, the “zero contribution threshold” changes for each corpus as it only depends on the corpus itself. This allows me to avoid having to select a statistical significance threshold for each dataset, and ensures that the threshold I do select (0) is always theoretically motivated.

### 3.4 Measurement and Validation

As discussed earlier, to form the contingency table necessary to identify legal and technical terms via my ACMI method, I used the partisanship of the sponsor of a bill, and the bills’ “minor topic label” (35) to form categories. There are approximately 240 minor topic
labels in the Congressional Bills Project metadata, covering relatively fine-grained topics such as “nuclear power” or “national parks” legislation. For each of these topics, I formed a contingency table consisting of counts of each term in substantive bill sections sponsored by members of each party. Having prepared the data for analysis, I then proceeded to calculate ACMI contributions for each term in the vocabulary.\textsuperscript{13} This process resulted in about 287,000 (3\%) terms with a negative ACMI. Table 3.6 contains some example terms with a negative ACMI.

\begin{table}[h]
\centering
\begin{tabular}{l}
\{pursuant to section\}, \{period beginning\}, \{provided in subsection\}, \{described in paragraph\}, \\ \{accordance with section\}, \{the secretary\}, \{for purposes\}, \{united states code\}, \{notwithstanding\}, \\ \{amendment made\}, \{act shall be\}, \{such regulations\}, \{effective date\}, \{shall take effect\}, \{date of enactment\}, \{striking subsection\}, \{inserting after paragraph\}
\end{tabular}
\caption{Table 3.6: Example terms with a negative ACMI.}
\end{table}

One well-known feature of information theoretic quantities like PMI and ACMI is that they tend to perform poorly for terms that appear infrequently in the corpus (70). Visual inspection of the terms with a negative ACMI revealed terms that only appeared in bills concentrated in five or fewer minor topics were overwhelmingly not legal or technical terms. Discarding these terms left approximately 15,000 terms remaining. Visual inspection of the remaining terms occasionally revealed phrases like “social security” or “healthcare premiums”, but generally indicated good performance in identifying legal and technical language.

In order to better assess the accuracy of my method, I performed a human coding validation task. The validation task consisted of hand coding all terms with a negative ACMI,\textsuperscript{13} Details of my implementation are provided in Appendix F).
that appeared in bills about at least 100 minor topics (about 7,800 terms). For each term, I marked it as a false positive if it gave any indication of being policy language. This criterion is likely to be overly strict, as some terms like “health” were used over a million times, and could have easily been written by a lawyer as part of a boilerplate passage. Additionally, I coded the top 1,000 most frequently appearing terms not coded as legal or technical for false negatives. The results of this coding exercise are provided in Table 3.7.

<table>
<thead>
<tr>
<th></th>
<th>Legal and Technical Terms</th>
<th>Substantive Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative ACMI (7,800 terms)</td>
<td>77.4%</td>
<td>22.6%</td>
</tr>
<tr>
<td></td>
<td>(157,629,373)</td>
<td>(23,165,756)</td>
</tr>
<tr>
<td>Positive ACMI (1,000 terms)</td>
<td>13.8%</td>
<td>86.2%</td>
</tr>
<tr>
<td></td>
<td>(751,747)</td>
<td>(15,991,770)</td>
</tr>
</tbody>
</table>

Table 3.7: Confusion matrix for human coding validation results. Total count of all terms in category is provided in parentheses.

My validation coding results indicated a false positive rate of approximately 23% and a false negative rate of 13.8%. However, it is difficult to characterize the quality of these results without assessing how they affect the final distribution over legal and technical terms across substantive provisions in legislation. I therefore created counts of legal and technical terms in each substantive provision using all terms I hand coded as legal or technical, all terms with a negative ACMI that appeared in at least 5 issue areas, and all terms with a negative ACMI (287,000 terms). The correlations among these counts are provided in Table 3.8.

The striking result from Table 3.8 is that the false positives largely do not matter for the overall distribution of legal and technical terms across substantive provisions. Including all terms with a negative ACMI essentially adds more to the count of legal and techni-
Table 3.8: Correlations among bill section legal and technical term counts using different cut-offs for identifying these terms.

Having successfully formed counts of legal and technical terms in all substantive provisions of all bills in my corpus, my final task was to prepare my data for analysis. As I wanted to assess the relationship between the inclusion of legal details and non-substantive sections and the chances of success for these bills, I needed to aggregate my measurements from the provision to the bill level. To do so, I calculated the average number of legal and technical terms in the substantive provisions of each bill, as well as the average proportion of terms in those substantive provisions that were legal or technical. I chose to operationalize the concept of legal details as the average proportion of legal and technical terms in the substantive provisions of a bill in order to account for the fact that bills about some issues might simply require more words to describe their policy proposals than others. To operationalize the use of non-substantive provisions in a bill, I created an indicator variable that recorded whether each of the seven types of non-substantive provisions I identified earlier in the paper were present in a bill. I chose to use an indicator instead of a count so
that these variables would not simply be conflated with bill length. Furthermore, despite noting similarities among some of these non-substantive provisions, I wanted to empirically explore the relative importance of the inclusion of each type of provision in a bill.

3.5 Analysis

In this section, I present the results of two regression analyses aimed at assessing the two hypotheses laid out earlier in the paper. To reiterate, my first hypothesis is that bills containing more legal detail in their initial versions will be more likely to advance out of committee, and to eventually become law. My second hypothesis is that bills containing more non-substantive provisions in their initial versions will be more likely to advance out of committee, and to eventually become law. To assess both of these hypotheses jointly, I estimated two logistic regression models at the bill-level, with an indicator for whether each bill advanced out of committee (became law) as the outcome. By selecting a regression modelling approach, I was also able to control for a number of standard predictors of bill success, and well as controlling for issue area, and session of Congress.

To assess Hypothesis 1, I included a variable in the model that recorded the average proportion of legal and technical language (legal detail) in the substantive provisions of a bill. Figure 3.1 depicts the distributions over the average proportion of legal and technical terms in the substantive provisions of the introduced versions of bills that eventually became law, and those that did not. To assess Hypothesis 2, I included dummy variables for the presence of each of the seven types of non-substantive sections identified earlier in the introduced version of a bill. To get a sense of how frequently these sections appear, Figure 3.2 displays
Figure 3.1: Distributions of the average proportion of boilerplate terms in substantive provisions of the versions as introduced of bills that eventually became law, and those that did not.

the proportion of successful and unsuccessful bills containing provisions of each type.

Following (32), I include a number of control variables in my regression model. These include fixed effects for the “major topic area” of a bill (approximately 20 policy areas from the Congressional Bills Project), and the session of Congress the bill was introduced (as a categorical variable). I also included controls for the gender and the ideological extremeness (absolute value of first dimension DW-NOMINATE score) of the sponsor, as well as whether the sponsor was a member of the majority party. Additionally, I included controls for whether the bill was a leadership bill (the first 10 numbered bills in the House each session), a revenue bill, a reauthorization bill, and an indicator for whether the bill was introduced in the Senate. I also included controls for the log of the number of cosponsors of the bill, and the log of the number of sections in the bill. Finally, I included controls for whether the sponsor was a member of the committee to which the bill was referred, or if
they were the chair or ranking member of either the committee or subcommittee of referral.

In particular, I expect the parameter estimates for these committee variables to be positive, as being a chair or ranking member of the referring committee is a significant institutional advantage in advancing legislation.

I began by estimating the model, and confirming that the model AIC is improved when including the non-substantive section dummies and proportion of legal detail variables. I then generated marginal effects on the relative likelihood that a bill advances out of committee or becomes law for all covariates (excluding the session of Congress and major topic fixed effects for readability). These marginal effects\(^{14}\) (with 95% confidence intervals)

\(^{14}\)Marginal effects were calculated by changing the value of a categorical covariate from its reference category, and from changing the value of nominal variables from their minimum to maximum values.
are depicted in Figure 3.3 for passage out of committee, and in Figure 3.4 for passage into law.

Figure 3.3: Marginal effects (with 95% confidence intervals) of covariates on the relative likelihood of a bill being reported out of committee. Variables of interest are highlight in gray.

As we can see from Figures 3.3 and 3.4 respectively, the marginal effects of the control variables tend to be in the expected direction. For example, a bill’s sponsor being a member, ranking member, or chair of the committee to which a bill was referred are all associated with an increased probability of that bill advancing out of committee and ultimately succeeding. Similarly, leadership and reauthorization bills are much more likely to be
successful, and the number of cosponsors is also positively associated with bill success. Interestingly, the effect for the log of the number of sections (provisions) in a bill is not significantly different from one, indicating that longer (potentially omnibus or otherwise collaborative) bills are no more likely to become law than shorter bills, controlling for the other variables in the model.

With regards to Hypothesis 1, Figures 3.3 and 3.4 indicate that increasing the average proportion of legal detail in the substantive sections of a bill from its minimum observed
value (about 6%) to its maximum observed value (99%) is associated with roughly a doubling in the relative likelihood of that bill advancing out of committee, and of eventually becoming law, respectively. These parameter estimates are consistent with my hypothesis that containing more legal detail should be more likely to advance out of committee and ultimately succeed in Congress.

As for Hypothesis 2, the results are more mixed. We see that the inclusion of Authorization of Appropriations and Technical Corrections provisions are associated with an increased relative likelihood that a bill is reported out of committee, and that it eventually becomes law. It makes intuitive sense that the inclusion of Technical Corrections provisions in a bill would be positively associated with its chances for success if these tend to clarify or correct relatively minor points in existing laws. Conversely, we see that bills containing Findings, and Purposes sections are relatively less likely to advance out of committee, or to become law. It may be the case that including Findings, and Purposes sections in a bill signals a position taking effort (because these provisions tend to be written in plain English), thus explaining their negative association with the relative probability of bill success. Overall, the results suggest that the inclusion of certain non-substantive provisions in legislation is predictive of bill advancement, but that these effects are not uniform. Future work should explore the role that these provisions play in greater detail.

While the regression results in Figures 3.3 and 3.4 tell a mostly consistent story about the association between increased legal and technical language content in legislation and bill advancement, they do not help us to understand where in the legislative process this language is most strongly associated with success. This is because the effects for bills that
become law could be predominantly driven by the effects on passage our of committee, or vice versa. To understand how legal language relates to bill advancement later in the legislative process, I replicated the analyses in Figures 3.3 and 3.4 but now only using the text of the version of bills that were reported out of committee (and thus only bills that were reported out of committee), and predicting whether those bills would become law. The marginal effects for this regression are reported in Figure 3.5. As we can see, most marginal effects are no longer different from zero, including those for legal details and the inclusion of most non-substantive provisions. This suggests that the primary impact of legal language comes in clearing the first hurdle of advancing out of committee. This is consistent with my theory that the inclusion of legal details serves as an early indicator for bill quality.

3.6 Discussion

Members of Congress introduce legislation for a variety of reasons beyond policymaking, and likely have varying expectations about the chances for success of each bill. Some bills may be almost certain to advance for institutional and procedural reasons, while others may be introduced as pure position taking stunts. While a large body of research has sought to understand why some bills succeed where others fail, scholars have lacked a way to characterize bill “quality”, and to account for it in their analyses of legislative success. In this study, I focus on the inclusion of legal and technical details in legislation, along with a number of non-substantive provisions as indicators that a bill is intended as a genuine lawmaking effort. I hypothesize that bills containing more legal detail in their substantive provisions will be more likely to advance out of committee, and ultimately to become
law. I argue that this is a result of legislators recognizing that bills containing more details represent more serious lawmaking efforts, and are thus worthy of further consideration.

To assess this hypothesis, I developed and employed text as data methods to identify and measure the use of legal and technical language, and non-substantive provisions in a large corpus of U.S. Congressional bills introduced between 1993 and 2016. I then estimated a logistic regression model predicting bill success, and controlling for a wide variety of factors associated with bill success that had been previously identified in the
literature. My results indicated support for my first hypothesis that the inclusion of a greater proportion of legal detail in legislation is associated with an increased likelihood of that bill advancing out of committee, and ultimately of becoming law. As for my second hypothesis that bills containing several different types of non-substantive provisions will also be more likely to become law, the results are more mixed. I find that the inclusion of several types of non-substantive provisions in a bill are positively associated with its chances of advancing, but that the inclusion of some non-substantive provisions may signal a position taking effort.

One of the main limitations of my approach is that I cannot parse out whether the inclusion of this legal and non-substantive content in legislation causes a bill to be successful, or whether it is consequence of the bill’s authors recognizing that it is likely to be successful for institution or procedural reasons, leading them to include this content in response. While my theory about the importance of this content does not rest on the direction of causality, it would be valuable to better understand the causal mechanism. Additionally, an intuitive consequence of my argument that genuine policymaking efforts will contain more legal detail is that pure position-taking bills should waste almost no time on the inclusion of such details. Yet, I find no sharp difference across successful and unsuccessful legislation, suggesting either that many position taking bills include at least some legal details, or that my measurement approach does not capture some important dimensions of bill quality. Furthermore, I do not account for “hitchhiker” bills (bills that are included in their entirety as provisions of other bills) or the reuse of specific provisions across multiple bills in my analysis, and this omission may lead me to underestimate the differences in the use of legal details across legislation.
My theory and findings also leave many open questions that should be explored in future research. For example, are most of the legal details included in legislation in the form of longer boilerplate passages, or do these details tend to be more dispersed through the text of its substantive provisions? And are there some staffers, outside interest groups, or individual legislators that are “better” at including these details in legislation than others? Furthermore, how do these legal details change as a bill moves through the legislative process? Are there some committees that have a particular propensity to add or remove these legal details as part of the committee markup process? Moving on to the inclusion of non-substantive provisions in legislation, more theoretical attention could be devoted to understanding the difference between these types of provisions in terms of their relationship to the success of a bill. For example, does the inclusion of Findings or Purposes provisions indicate that a piece of legislation was merely a platform for legislators to take positions?

My findings, along with the method I develop, also have important implications beyond the scope of this study. For example, the method I develop for identifying legal and technical language in legislation can be applied much more broadly to identify domain stopwords in political texts. While I am explicitly interested in studying how these terms are used, the method I develop can just as easily be applied to remove these terms before conducting analyses of substantive language. My approach is also easily scalable to millions of documents containing billions of words, and is implemented in an R package. Substantively, my findings have several important implications for the study of lawmaking. Most importantly, I argue that future studies of lawmaking need to account for bill quality in assessing legislative success, and my results indicate that legislators do take at least one dimension

---

15 https://github.com/matthewjdenny/SpeedReader
of bill quality (the inclusion of legal detail) into account when deciding which legislation deserves further consideration. Finally, in this study, I have sought to provide evidence that the actual *words* in legislation matter. To the degree that I have been successful in this endeavor, this should spur legislative scholars to look more deeply into the text of legislation in order to better understand the lawmaking process.
Chapter 4  
More Effective Than We Thought: Accounting for Legislative Hitchhikers Reveals a More Inclusive and Productive Lawmaking Process

4.1 Introduction

In 2014, a Washington Post article described the legislative record of retiring Representative Robert Andrews (D-NJ) as the worst in Congress: “Andrews proposed 646 bills, passed 0: worst record of past 20 years.”¹ In response, Andrews objected that journalists

were using the wrong metric: “I’m just a bill is not the way it works.”

Legislative scholars have also challenged this orthodox view of lawmaking: “The Schoolhouse Rock! cartoon version of the conventional legislative process is dead, if it was ever an accurate description in the first place” (71). Increasingly, a process of considering bills on an individual basis has been replaced by a leader-centered process of constructing larger omnibus bills that combine multiple policy proposals into one (72; 73; 31).

Andrews’ advice was to also count policy proposals that “germinate in a larger bill.” In this paper, we develop an approach for doing that - identifying bills that are enacted into law as provisions of other bills. We then consider the implications of accounting for these “hitchhiker” successes for legislative effectiveness research. The next section reviews the longstanding legislative effectiveness literature and its limitations. We then propose and implement a new text-based methodology for accurately identifying hitchhiker bills. Applying this methodology to two decades of lawmaking (1993-2014), we find that as many bills become law as hitchhikers as become law on their own.

We argue that agenda and procedural constraints are central to understanding why lawmakers pursue hitchhiker strategies. Legislators who sponsor bills that become law on their own are more likely to hold agenda setting positions that allow them to claim credit for bills that succeed for reasons other than their sponsorship (such as legislative reauthorizations). Aside from these agenda setting advantages, the sponsors of successful laws and successful hitchhikers have very similar attributes. We also find that procedural constraints lead the Senate to employ hitchhiker strategies more frequently than the House and that more hitchhikers are adopted under unified governments because those governments are more likely to engage in omnibus lawmaking.
4.2 Effectiveness Research and its Limits

Studies of legislative effectiveness fit into a broader literature examining legislative influence (see, for example: 74; 75; 45; 52; 76; 77; 78; 47; 79; 80; 81). They include some of the earliest quantitative analyses of legislative behavior. From then until now, scholars have focused on bill sponsorship success as the central indicator of effectiveness. In US Senators and their World, Donald Matthews observed: “To the extent that the concept as used on Capitol hill has any distinct meaning, effectiveness seems to mean the ability to get one’s bills passed” (44). Matthews found that senators who adhered to chamber “folkways,” such as specializing and spending less time giving floor speeches, were more likely to sponsor successful bills. A decade later, (82) asked whether members of the House who adhered to similar norms were also more legislatively successful (they weren’t). Subsequent research has continued to investigate bill sponsorship success patterns to better understand norms and coalition building (see, for example: 72; 83; 84; 48). An equally important body of research seeks to discover (in the words of 85) the “remarkable skills” of the lawmakers who are more successful in advancing their bills (86; 87; 88; 85; 89; 90; 1).

The methods employed in these studies have become considerably more sophisticated over time, but the central measure has changed very little. Effectiveness continues to be defined in terms of how far a sponsor’s bill progresses through the legislative process. Some define progress by whether a bill receives any committee consideration (72) whereas others define it by whether a bill passes the chamber. Some focus on “hit rates” — the percentage of a legislator’s bills that succeed (85) — whereas others focus on the progress of individual bills. The most recent research also offers the most thoughtful and sophisticated measure. (1) compute “Legislative Effectiveness Scores” (LES) by summing the number of bills a
member introduces, weighted by their progress and importance.

Bill success has also recently attracted the interest of scholars in other disciplines and even entrepreneurs. Rather than trying to understand why some lawmakers are more effective, the objective is to predict bill success as one might predict the winner of a sporting event or election (91; 92). Several commercial ventures are currently or soon will be offering bill success prediction services.²

We contend that an important limitation of these efforts is that bills are vehicles, not policies. The progress of a bill and a policy can be one and the same, but this is not always the case. The Affordable Care Act (HR 3590) started off as a seven page bill proposing a first time home buyer credit for service personnel. It became the Affordable Care Act when the Senate stripped that language and replaced it with a 900 page health care amendment.³ Current approaches give the original bill’s sponsor (Rep. Charles Rangel, D-NY) full credit for the Affordable Care Act, despite the fact that the final law was completely unrelated to the bill he introduced. As we will show, many other lawmakers deserved (but do not receive) credit for what is in the ACA.

Equally important, policies proposed in bills can progress when the bills themselves do not. The lawmaking process has fundamentally changed since Matthews equated bill passage and effectiveness. A process that used to be driven by largely autonomous committees recommending bills on an individual basis has been replaced, to an increasing extent, by leadership-driven negotiations. These negotiations often produce large “omnibus” bills that combine proposals originating in other bills (93; 73; 31). Recent research also finds that lawmakers view “must pass” legislation such as reauthorizations of expiring programs as exceptional opportunities to advance substantively related policy initiatives.

²See, for example: Skopos Labs, GovTrack, StateHill
We propose an approach to studying effectiveness that gets closer to what scholars (and citizens) ultimately care about — legislators’ ability to get their policy proposals enacted into law. One implication of more recent developments is that the legislative opportunity structure increasingly favors “hitchhiker” strategies. This suggests that legislative effectiveness research will benefit by crediting lawmakers not only for bills that become law on their own, but also for bills enacted into law as provisions of other bills. We find, for example, that the Affordable Care Act includes almost 50 “complete” hitchhiker bills (cases where the complete substance of a bill was enacted as a hitchhiker).

Accounting for hitchhiker bills constitutes an improvement over current approaches to measuring legislative effectiveness. In this paper we do not attempt to identify cases where only part of a bill became law as an insertion into another bill.\footnote{Prior research suggests that the attributes of successful sponsors of partial insertions will be similar to those reported here (30).} We also do not examine policy proposals that originate as amendments and we continue to inappropriately credit some sponsors for a bill’s progress (such as Rep. Charles Rangel in the case of the ACA). Despite these limitations, accounting for hitchhiker successes offers important opportunities to explore how laws are made, and to better understand the distribution and components of effectiveness in Congress.

### 4.3 Why Hitchhikers?

Why would a sponsor advance a bill as a hitchhiker when authoring a stand-alone law would seem to offer more visible credit claiming opportunities? The main reason is that legislators’ opportunities to advance stand-alone bills are limited. For the chamber,
hitchhiker strategies can be procedurally efficient and, in some cases, procedurally necessary. In this section, we propose three hypotheses about why lawmakers pursue hitchhiker strategies.

Before considering these hypotheses, it is also worth noting that legislators do claim credit for hitchhiker successes. Rep. Carolyn Maloney’s (D-NY) official website includes a “Laws Enacted” page. The majority of the enactments listed (40 out of 74) are either sponsored bills that were “included” in other laws, or laws (sponsored by others) that were “versions” of bills she had sponsored. Maloney also highlights hitchhikers in her direct communications with constituents. Her Spring 2010 Report to Manhattan newsletter specifically mentions provisions of the recently passed Affordable Care Act that are “based on” bills she sponsored.

We expect that many of the covariates reported to predict bill sponsorship success in prior effectiveness studies will also predict hitchhiker bill successes. However, we also expect two other political considerations — agenda control and procedural constraints — to explain why some lawmakers are more likely to sponsor successful laws, and why some are more likely to sponsor successful hitchhikers.

### 4.3.1 Agenda Control

Congressional agenda space is a scarce commodity. It has always been the case that only a small percentage of bills make it beyond introduction. Party polarization and legislators’ increased willingness to engage in obstruction seem to have made passing bills through the regular order increasingly difficult (31; 73). As a result, the number of laws enacted by Congress has declined significantly since the 1970s (96, p.145, Figure 7.1). The policies

---

that do become law also typically endure a lengthy incubation process (97).

Members of the majority party use their control over the agenda to monopolize these limited credit claiming opportunities (98). In the 113th Congress (the most recent of the Congresses we analyze in this study), about 30% of all non-minor laws were sponsored by just 63 House and Senate committee and subcommittee leaders (12% of all lawmakers). Majority party members (constituting 50-60% of the chamber) sponsored about 82% of all non-minor laws. Many of these successes have little to do with effectiveness. Agenda control provides majority party lawmakers with exceptional opportunities to put their names on bills that progress for other reasons. Majority pary leaders also have limited incentives to share the most visible credit claiming opportunities with members of the minority party, especially in the House. We expect to find that these partisan calculations are less applicable to (less visible) hitchhikers. Majority party leaders should be more willing to accept minority party hitchhikers that advance good public policy or increase support for other legislation (99; 73).

**Hypothesis 1 – Agenda Control:** Agenda control (serving as a committee or subcommittee chair or member of the majority party) will be a more important predictor of law success than hitchhiker success.

### 4.3.2 Procedural Constraints

The agenda control hypothesis above suggests that hitchhiker successes may be better indicators of true legislative effectiveness because many bills progress for reasons that have little to do with who sponsors them. In this section we hypothesize that procedural constraints also help to explain why some bills are more likely to advance as hitchhikers.

*Revenue bills.* The clearest example of a procedural constraint that incentivizes hitch-
hiking is the “origination” clause of Article I of the Constitution — all laws raising revenue must originate in the House (100). The House of Representatives vigorously guards this constitutional prerogative by “blue slipping” (rejecting) Senate bills with revenue implications. The practical result is that Senate proposals with revenue-related provisions can only advance as hitchhikers on House-originating laws.\(^6\) We treat all bills originally referred to the Senate Finance and House Ways and Means committees as revenue-related (because all tax bills must be referred to these committees).

**Hypothesis 2 – Revenue Bills:** Revenue-related Senate bills are less likely to become law on their own than House revenue-related bills, but they are not less likely to be enacted as hitchhikers.

*Amendments between chambers.* In both chambers of Congress, bills passed over from the other chamber are considered under different procedures than the chamber’s own bills (100). In the Senate, it can be easier to take up a House passed bill than a bill reported by a Senate committee. This is because House-passed bills are typically placed on the Senate’s Calendar of Business, bypassing the committee referral process. To bring up a Senate bill, the majority leader must negotiate a motion to proceed (which is subject to filibuster). In contrast, a referred House bill is already on the calendar, making it an attractive vehicle for Senate hitchhikers (101). This is why Senator Majority Leader (at the time) Harry Reid (D-NV) used H.R. 3590 as the vehicle for the Affordable Care Act (102).

Another reason to expect amendments between chambers to be important entry points for hitchhikers is the fact that the President can sign only one bill into law when the House and Senate pass separate bills on a policy. Rybicki notes that a common practice in such cases is for one chamber to take up the other chamber’s bill, “strike all after the enacting

\(^6\)In practice, this requirement also extends to appropriations bills, which we exclude from our analysis (100, p. 2).
clause” and insert its own proposal (100, p. 3). We should therefore expect the process of resolving differences to lead to many cross-chamber hitchhikers.

**Hypothesis 3 – Amendments Between Chambers:** Cross-chamber hitchhikers will be the most common type of hitchhiker. Senate bills are more likely to be enacted as hitchhikers than House bills.

### 4.4 Finding Hitchhikers: A Supervised, Active Learning Approach

In this section we describe how we use text reuse methods to identify hitchhiker bills. The general goal is to compare the text of every version of every bill that did not become law to the text of every law enacted in that Congress. If any version of a failed bill aligns with a law, we consider that bill to be a hitchhiker. We started with a corpus of 92,677 bills for the 103rd-113th Congresses (1993-2014) collected by (34). This corpus includes 4,176 bills and joint resolutions that became law and 111,758 versions of bills and resolutions that failed to become law. We excluded non-joint resolutions (because they cannot become law), appropriations bills (because they are quasi-compulsory (95)), and very minor private and duty suspension bills. After these exclusions, our primary analysis considers 84,913 bill versions. In much of our analysis, we also exclude minor legislation as defined by the Congressional Bills Project (examples include bills naming federal buildings or creating commemorative coins).

The standard supervised learning approach to matching bill content and law content is to manually label a large, random sample of bill-law pairings for whether the law contains the substance of the bill, train a classifier on part of this sample, and test its performance
on a held out set of labeled cases. Prediction accuracy is then assessed, and if it is high enough, the trained classifier is used to predict (label) bill-law pairs in the broader corpus.

The first problem with this standard approach for the current study is that hitchhikers are probably rare. If they are as rare as laws (about 3% of bills become law), we would have to visually examine and label about 10,000 bill-law pairs to obtain a sizable sample of true hitchhiker cases (3-400). One alternative solution from the machine learning literature is to use “active learning” to iteratively assemble a training sample of sufficient size (103). In the first iteration, a small number of likely hitchhiker cases is identified and labeled. This initial sample is then used to train a classifier to predict additional likely cases. These cases are then labeled and added to the training corpus and the process is repeated. Using this method, we were able to identify substantial number of true hitchhikers after labeling less than than 1,000 bill-law pairings (for a detailed explanation of the active learning method, see: I).

A second challenge, discovered during the labeling process, is that, even for true hitchhiker cases, the bill and law texts can be quite different. One common reason was that a bill often contains non-substantive front matter (such as the title and date of introduction) and even sections (e.g. Findings and Definitions) that are removed when its substance is incorporated into another law. To address this concern, we developed a pre-processing protocol that removed common non-substantive language from both the bill and law texts (see G for a full description of the pre-processing steps).

Even after this pre-processing, however, the substantive language of the law and hitchhiker bill could still differ due to relatively minor edits in the law language. We initially trained and tested several algorithms widely used in computational linguistics and infor-
All of them predicted the cleaner bill-law comparisons quite well, but none did a good job of predicting the somewhat messier cases that included reordered, deleted or inserted text or sentences. This common shortcoming inspired us to develop an entirely new approach. Below, we describe the basic intuition. A detailed description of the methodology can be found in H.

4.4.1 A New Sequence-Based Algorithm for Characterizing Document Similarity

Hitchhikers are similar to cases of plagiarism. They are characterized by lengthy sequences of matching text (between the bill and law), sometimes interspersed with shorter sequences of mismatched text. “Bag of words” approaches (e.g. Cosine similarity, Dice coefficients) do not value word sequence or proximity. Alignment algorithms do (e.g. Smith-Waterman), but they require that the researcher specify, in advance, the penalties for mismatches in scoring the similarity of two documents. These ex ante decisions can have important consequences.

Our approach accounts for word proximity without committing to a single parameterization (as Smith-Waterman requires). We propose a “sequence-based” algorithm that (like other alignment algorithms) uses only information about patterns of matching and non-matching text. It does not consider (for example) the frequency of co-occurring words as do many bag of words approaches. However, it differs from other alignment algorithms in important ways. To illustrate, below are two versions of the same section of the Dodd-Frank Wall Street Reform and Consumer Protection Act. The first (version A) is from the bill as introduced in the House:

\[ \text{diff}, \text{wdiff}, \text{Dice coefficients (104), Cosine similarity, and the Smith-Waterman algorithm (105)} \]
SEC. 1008. OVERSIGHT BY GAO.

(a) Authority to Audit.--The Comptroller General of the United States may audit the activities and financial transactions of--

(1) the Council; and

(2) any person or entity acting on behalf of or under the authority of the Council, to the extent such activities and financial transactions relate to such person’s or entity’s work for the Council.

The second (version B) is from the version signed into law by President Obama:

SEC. 122. GAO AUDIT OF COUNCIL.

(a) Authority To Audit.--The Comptroller General of the United States may audit the activities of--

(1) the Council; and

(2) any person or entity acting on behalf of or under the authority of the Council, to the extent that such activities relate to work for the Council by such person or entity.

These two versions clearly have the same intent, but they are not identical (e.g. the section titles are different). We first characterize each text as a set of overlapping or “shingled” n-grams. An n-gram is a contiguous sequence of n words. Overlap means that adjacent n-grams share words. Here we use 5-grams that overlap by n-1 words. In version A, two 5 grams that overlap by n-1 are “to work for the Council” and “work for the Council by.” We then compare each n-gram in version A to each of those in version B, recording whether there is a match as a vector entry. Figure 4.1 displays the results for this example comparison. Black rectangles indicate the version A 5-grams that have a match in version B, whereas grey rectangles indicate version A 5-grams that do not match any n-grams in version B. Thus, a sequence of black rectangles indicates a longer block of shared text, etc.

The key benefit of this approach is that this match/non-match information can be used to construct many sequence-based similarity statistics (e.g. longest matching sequence, average matching sequence length, number of unique matching blocks, etc.). These statistics can then be introduced as features of supervised learning models. These models can be
Figure 4.1: A comparison of two versions of a section of the Dodd-Frank Wall Street Reform and Consumer Protection Act.

IH version: 5-grams with a match in the PL version

PL version: 5-grams with a match in the IH version

Note: Black rectangles indicate where a 5-gram in the section of the introduced version of the bill exactly matches a 5-gram in version that became law, and vice-versa in the bottom plot.

trained to predict known hitchhiker cases, and the best of them can be selected and used to predict hitchhikers in the broader corpus.

We tested over 1,500 models using different combinations of 21 different statistics calculated on these sequences of matching and non-matching n-grams. We started with a small number of previously labeled examples (that included about 80 true hitchhikers) and used them to identify an initial set of high performing models. We then applied these models to the broader corpus (bill-law pairings of the 111th Congress) to predict additional hitchhiker cases. After manually labeling these newly identified cases and adding them to the training set, we repeated the process (until the best performing models stopped predicting new hitchhikers). We then used the majority vote of an ensemble of 22 high performing models to predict hitchhikers across 20 years of lawmaking. This approach proved to be much more accurate than earlier experiments with other algorithms.\(^8\)

\(^8\)Specifically, the majority vote of this ensemble had 95% precision (5% false positive rate) and 92% recall (8% false negative rate) based on 300-fold cross validation. The off the shelf algorithms had higher recall on average (99%), but much lower precision (75%). In this respect our approach is more likely to underestimate than overestimate the true number of hitchhikers.
4.5 Findings

We begin by examining hitchhiker patterns across eleven recent Congresses. We then test the hypotheses proposed earlier by comparing multivariate regression models predicting whether a bill becomes law on its own, and whether a bill is enacted as a hitchhiker. These models include indicators of standard explanations of legislative effectiveness as well as indicators of the agenda control and procedural constraints hypotheses presented earlier. We then explore how accounting for hitchhikers alters conclusions about legislative effectiveness in Congress. Finally, we shift our attention from effectiveness to exploring hitchhiker strategies more generally. What types of bills are particularly attractive vehicles for hitchhikers? Where in the lawmaking process do hitchhikers tend to be incorporated? Do broader political conditions help to explain more frequent use of hitchhikers?

4.5.1 Hitchhiker Bills in Congress, 1993-2014

Figure 4.2 confirms the importance of this type of unorthodox lawmaking. The figure compares the number of non-minor (left) and minor (right) bills that became law on their own and that became law as hitchhikers for each Congress.\footnote{As discussed earlier, we use the “Important Bill” filter of the Congressional Bills Project to distinguish minor bills.} For the 1993-2014 time period, our method indicates that more non-minor bills became law as hitchhikers (2,997) than became law on their own (2,905).\footnote{A list of all hitchhiker bills and their target laws will be made available with the replication materials for this paper. Two example target laws and their hitchhikers can be found in K. As noted earlier, Appropriations bills, private bills, and duty suspension/tariff bills are not included in these counts - see G.} Thus, focusing only on bills that become law on their own misses about half of all legislative enactments.

Interestingly, minor bills are much more likely to be enacted as stand alone laws than as
Figure 4.2: Counts of laws versus Hitchhiker bills (103rd-113th Congresses).

hitchhikers. We view this as consistent with the agenda control argument proposed earlier. Minor bills (e.g. naming federal buildings in the district) do not consume limited agenda space. They do not go through the markup process and typically pass under expedited procedures (Suspension of the Rules in the House and Unanimous Consent in the Senate). They are unrelated to the majority’s agenda. For all of these reasons, there is probably less need to pursue hitchhiker strategies in these cases.

4.5.2 Sponsor and Procedural Predictors of Bill Success

Does accounting for hitchhikers alter current understandings of who is effective in Congress? Prior studies measure effectiveness using either a single threshold of success (e.g. was the bill taken up in committee or passed by the chamber? 72; 86), or by weighting bills by how far they advance in the process (e.g. the LES scores of 1). We would not expect much difference if the bills that become law as hitchhikers also tend to advance most of the way through the process on their own. However, Figure 4.3 indicates that this is not usually the case. Most non-minor hitchhiker bills do not even make it out of committee.
on their own. This gives us reason to think that accounting for hitchhikers may lead to different conclusions about who is effective in Congress.

Figure 4.3: How far do hitchhiker bills advance on their own?

To test this expectation, we estimate two logistic regression models predicting whether a bill becomes law on its own and whether it becomes law as a hitchhiker.\textsuperscript{11} We test the same sponsor characteristics commonly found to be important in prior effectiveness research (such as seniority, ideology, gender, etc.). However, our committee-related variables differ from prior research. Whereas prior studies only ask whether the sponsor leads any committee, we ask whether they lead the committee responsible for the bill (or a subcommittee of that committee).\textsuperscript{12} We view these measures as better indicators of the effectiveness benefits of agenda control than more general committee leadership measures.

We also include several bill type and institution-related predictors. The first is whether a bill enjoys bicameral support. We measure this by whether the bill has an identical or nearly identical “companion” bill in the other chamber (106; 107).\textsuperscript{13} We also expect

\textsuperscript{11}Non-minor bills only. The second regression considers only bills that did not become law on their own; following a sequential logit logic. The results of a multinomial logistic regression model predicting a three-class outcome (a bill does not become law, becomes law as hitchhiker, or becomes law on its own) show very similar results.

\textsuperscript{12}For bills referred to multiple committees, this variable indicates if the sponsor led at least one of them.

\textsuperscript{13}Defined by whether the text of an introduced bill in the other chamber is at least 95% similar (after
certain types of bills to be more likely to advance regardless of sponsor. The first type are administration-initiated bills introduced “by request.” The second are legislative reauthorizations that reflect impending or past program expirations or “sunsets.”

Finally, we test two indicators of political conditions that may encourage hitchhiker strategies. The first is partisan gridlock. Lawmakers may turn to hitchhiker strategies as it becomes more difficult to pass laws in general. We use the gridlock interval (the ideological space between the members who represent the cloture and veto-override pivots, respectively) to control for this possibility. However it should be noted that prior empirical research does not generally find that larger gridlock intervals predict lower legislative productivity. The second political condition is unified government. Whereas partisan gridlock hypothesis is that legislators turn to hitchhiker strategies when the lawmaking process is not working, the expectation here is that actors in unified governments are better able to coordinate their lawmaking activities. More specifically, we expect to find that hitchhikers are more common under unified government because unified governments are more likely to engage in omnibus lawmaking.

Figure 4.4 presents the effects of the different independent variables as marginal probabilities of a bill becoming law on its own (LAWS), or as a hitchhiker (HITCHHIKER). Each set of results includes two scales because the marginal effects for two variables at

14Clause 7 of House Rule XXII prohibits the requesting party from being named, but House rules specify the types of bills that must be initiated by request. Most are trade or international agreements. Annual defense authorizations are also frequently introduced by request. We therefore designate, as administration bills, any “by request” bill that is primarily about defense, trade or international affairs.

15We search for bills that have “reauth” in their titles. This approach overlooks many cases (such as the reauthorization of the Elementary and Secondary Education Act in 2001 (“No Child Left Behind Act of 2001”). These omissions have the effect of making committee and subcommittee chairs (who typically sponsor them) appear more effective.

16The full results are presented in J. The estimates are based on min-max values because many of the independent variables are dummies where a one standard deviation change is meaningless.
Figure 4.4: Marginal effects of sponsor and bill characteristics on law *versus* hitchhiker success.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>LAW</th>
<th>HITCHHIKER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Committee Chair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subcommittee Chair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Committee Member (Majority)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Committee Rank Member</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subcommittee Rank Member</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Committee Member (Minority)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years in Congress</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extremism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bills Sponsored</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Co-sponsors (log)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unified Congress</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gridlock Interval</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reauthorization bill</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue Bill (Senate)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Companion Bill</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administration Bill</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Relative likelihood of a bill becoming a law on its own or as a hitchhiker

Note: *Each dot-triangle represents the average marginal effect of going from minimum to maximum value (for binary variables: the average marginal effect of being a Majority party member, Committee Chair, etc). The lines represent 95% confidence intervals. The top and bottom x-axes indicate how much the likelihood of a bill being enrolled increases (2 times, 3 times, etc.). All but the two bottom reported effects are on the top-x-axis scale. The triangles and shaded area indicate that the two last effects are on the bottom-x-axis scale.*

Overall, the models indicate that sponsors of successful hitchhikers possess characteristics that are very similar to successful law sponsors. As expected, however, committee
leaders and majority party members are much more likely to sponsor the bills that become law on their own. In addition, legislative reauthorizations are about 2.5 times more likely to be enacted into law than other bills, and administration bills about 15 times more likely.\textsuperscript{17} Bills that have companions in the other chamber (an indicator of bicameral support) are about 5 times more likely to become law. As expected, revenue-related bills that originate in the Senate, have virtually no chance of becoming law on their own. However, they are as likely as other bills to become law as hitchhikers.\textsuperscript{18}

The models also offer some evidence that the broader political context contributes to more hitchhiker lawmaking. As has been reported in prior research, we do not find that larger gridlock intervals predict lower overall productivity (110; 108; 109) or more hitchhiking activity.\textsuperscript{19} However, unified governments are both more productive and more likely to enact laws that include more hitchhikers. An important reason for this (not shown) is that unified governments are more likely to engage in omnibus lawmaking.\textsuperscript{20}

\subsection*{4.5.3 Consequences for Effectiveness}

Figure 4.5 examines how accounting for hitchhikers alters the proportion of lawmakers in each Congress that can claim at least one legislative success. In every category and

\footnotesize{\textsuperscript{17}}When these compulsory bill indicators are omitted from the law success model, the marginal effects of the agenda control variables (committee leader and majority party) are about 15\% larger. This confirms that the effectiveness of lawmakers in these positions is overstated in studies that do not control for compulsory legislation. The limitations of efforts to identify compulsory legislation further suggest that even our models exaggerate the relative effectiveness of these lawmakers.  

\footnotesize{\textsuperscript{18}}Bills referred to the Senate Finance Committee. The regression models themselves include a House and Senate revenue-related bills and an interaction with chamber. House revenue bills are somewhat less less likely than other bills to become law on their own.  

\footnotesize{\textsuperscript{19}}Here we use the \textit{Gridlock Interval} from (108). The results were the same for (60)’s measure.  

\footnotesize{\textsuperscript{20}}By using bill length to detect omnibus legislation, and by considering bills at the 99th length percentile as omnibus, we found that on average unified Congresses pass about 12 omnibus bills whereas non-unified Congresses only pass half as many.

86
in every Congress, hitchhikers add a substantial number of new legislators to the list of effective members. In proportional terms, the largest difference is for members of the minority party. Their list of effective lawmakers doubles from 16% to 32% over the time period.

Figure 4.5: Percentage of legislators sponsoring at least one law, or at least one law or hitchhiker.

Another perspective is to compare individual legislators using a measure of effectiveness that incorporates hitchhikers and one that does not. To do this we standardize Representatives’ Legislative Effectiveness Scores (1) for the 111th Congress and compare them to a standardized effectiveness score that is based on enactments (laws plus hitchhikers). We then examine differences between members’ scores on these two measures.

Figure 4.6 provides two views of the same results. The figure in the upper right shows the overall distribution of differences. A value of 0.0 indicates that a member was equally effective by both measures while a positive (negative) value indicates that the standardized LES score rates a member as more (less) effective than our enactment measure. The leftmost figure restricts attention to the cases of more extreme difference. Triangles indicate

21 We divide each member’s LES by the maximum LES, and each member’s enactments by the maximum number of enactments.
committee leaders whereas dots indicate rank-and-file members. The number on the left indicates the adjusted LES score for that member while the numbers of the right indicate the number of laws and hitchhikers (in parentheses) sponsored by that member.

Consistent with earlier findings, the LES score tends to rate rank and file lawmakers as less effective (those in the upper left of the figure are all rank and file members). For example, none of the bills Rep. John Salazar (D-CO) sponsored became law on their own during the 111th Congress, but five of his laws were enacted as hitchhikers. One of these bills (H.R. 71) established the Sangre de Cristo National Heritage Area in Colorado as a provision of H.R. 146. Another (H.R. 346) provided grants for physicians in rural areas to improve their professional training and was enacted as a provision of the Affordable Care Act.

In contrast, the legislators rated as more effective by LES (lower right) are disproportionately committee leaders. The most extreme case is David Obey (at the time, chair of the House Committee on Appropriations). All of Obey’s successful bills were appropriations bills. We exclude appropriations from our analysis because they are clearest examples of the kind of compulsory legislation that conflates effectiveness with agenda control. The second most extreme case is Sander Levin (D-MI), who took over as chair of the House Ways and Means Committee in 2010.

4.5.4 Where Are Hitchhikers Added?

Two final hypotheses to be tested are whether hitchhikers are frequently inserted while one chamber is considering a bill passed by the other, and whether Senate bills are more likely to become law as hitchhikers on House bills. These expectations are based on the fact that the origination clause requires that bills with revenue-related provisions originate in the
House, and the fact that it can be easier to take up a House-passed bill in the Senate than a Senate bill recently reported from committee. Figure 4.7 indicates although hitchhikers get added at every stage of the lawmaking process, the most common stage is when one chamber is amending a bill passed over by the other chamber.\textsuperscript{22} Perhaps most striking is that, in the vast majority of cases, the vehicle for Senate as well as House hitchhikers is a House bill (upper figures). In fact, more Senate bills became law as hitchhikers on House laws (1,118) than were enacted on their own (1,037). The largest proportion are revenue bills. In terms of topic, about half of these hitchhikers address the same major topic as the primary topic of the law (black shading), while about half address other topics (grey shading).\textsuperscript{23}

### 4.6 Discussion

In this paper we reexamine a longstanding subject of legislative studies. In 1960, Donald Matthews observed that “[t]o the extent that the concept as used on Capitol hill has any distinct meaning, effectiveness seems to mean the ability to get one’s bills passed.” For more than 50 years scholars have defined legislator effectiveness by whether the bills they sponsor advance through the formal stages of the legislative process. We redefine getting “one’s bills passed” to include bills enacted into law as provisions of other bills. Hitchhiker bills are just one way that lawmakers are able to exercise policy influence. They are closer to the “ground truth” of effectiveness than approaches that focus on how far bills progress in the legislative process on their own. We have not examined partial bill hitchhikers or

\textsuperscript{22}To produce this figure we compared the hitchhiker to each successive version of the bill that became law. We assume that it was inserted at the first match.

\textsuperscript{23}Using the 20 major topic codes of the Policy Agendas Project.
successful amendments.\textsuperscript{24} We have also excluded a number of issue areas from our analysis where hitchhikers are known to be common, including appropriations (earmarks) and miscellaneous tariff legislation (111; 112). Nevertheless, accounting for these hitchhiker successes provides new insights into effectiveness and into the lawmaking process more generally. We find that the congressional opportunity structure is less hierarchical and less partisan. We also observe differences in bill and hitchhiker success across chambers that reflect important procedural differences.

We have also tried to highlight limitations of bill success as a measure of effectiveness. Many bills progress for reasons that have little to do with who sponsors them. This leads to overestimates of the effectiveness of legislators in agenda setting positions (especially committee leaders), although the precise effects are difficult to estimate. But perhaps the best reason to be concerned about bill success as a measure of effectiveness is the fact that most of the bills senators sponsor that become law do so as hitchhikers on laws that originate in the House. Clearly, current approaches overlook many Senate successes and may even lead to misleading conclusions about relative chamber influence.

There is much more about hitchhikers to explore. We have not examined the policy areas that attract the most hitchhikers, or the most off topic hitchhikers. Hitchhikers also offer opportunities to study bicameral negotiations more systematically. Whereas current research examines just one or a very small number of cases (see (113) for a summary), the text based methods introduced here provide opportunities to assess the relative influence of the House and Senate in these negotiations across many bills, issues, and partisan circumstances (e.g. unified versus divided government).

Another intriguing question yet to be examined is the extent to which House bills

\textsuperscript{24}(30) conduct a cursory examination of section insertions for the 111th Congress and find similar minority party success rates to those reported here.
enacted as hitchhikers are added in the Senate and vice versa. The 900 page Senate amendment to HR 3590 that was the Affordable Care Act demonstrates that this occurs. It includes a number of hitchhikers that align with House bills that did not become law on their own. Furthermore, which legislators are most effective at advancing their proposal in this non-conventional way and why?

Research on legislative productivity currently measures it in two ways - counts of laws and counts of “major laws” (see, for example: 114). Counting hitchhikers as enactments has a dramatic impact on the former: Congress is about twice as productive. But hitchhikers also offer new opportunities for systematically categorizing laws and examining legislative productivity by defining omnibus laws in terms of the number of hitchhikers they include, the diversity of their topics, as well as the amount of text attention each receives.

More broadly, the similarly algorithm introduced in this paper can be used to investigate how the substance of thousands of individual bills evolves as they move through the lawmaking process. One basic yet to be examined question is — how much do the bills that become law change from one stage of the lawmaking process to the next? Statistical features derived from the algorithm can also be used to study more specific questions such as: Are bill edits mostly additions of new text or deletions? Do they tend to be granular (indicating focused word-smithing) or coarse (indicating the introduction or deletion of new provisions? Are new additions typically on-topic or off-topic? Do editing patterns differ depending on stage of the process (committee vs. floor), chamber, topic, or political context? Can editing patterns predict cosponsorship or whether a bill will progress?
Figure 4.6: Comparison between the sponsor’s enactments and the Legislative Effectiveness Scores (LES) of (1).

Note: We normalize LES and our enactment-based measure by dividing them by their maximum value. Then we take the difference between them, creating a normalized \([-1, 1]\) index communicating how LES underestimate (negative values) or overestimate (positive values) effectiveness relative to our measure. The numbers on the left indicates the adjusted LES score for these members while the numbers of the right indicate the number of laws, and the number of laws plus hitchhikers (in parentheses) sponsored by that member.
Figure 4.7: Where hitchhikers bills get picked up during the legislative process.

Note: Dark gray indicates hitchhikers that address the same major topic as the law. Light gray indicates the distribution of other topics.
Chapter 5  Overall Conclusions and Future Work

In this dissertation, I introduced three methods for legislative text analysis, and applied them to open questions in the study of the U.S. Congress. In doing so, I made substantive contributions to the study of constraints on bureaucratic policymaking, lawmaking and legislative productivity. It also encapsulated my philosophical approach to the development of methods for studying politics. I believe in the value of intuitive and interpretable methods that allow political scientists apply their domain knowledge to improve the measurement of latent political variables, and each chapter introduces a method I feel is consistent with this approach. These chapters also take to heart my goal of using computational methods try and understand the signals legislators are sending us in the documents they and their staffers produce. In doing so, I explore aspects of the lawmaking process that have traditionally been “hidden” in plain sight from large scale analyses. This fits with my broader research agenda, where I intend to continue to drill into subtle political aspects of the legislative drafting process through the lens of bill text analysis.
In the first chapter, I hypothesize that Republican lawmakers strategically time their efforts to constrain the bureaucracy to encounter less push back—when a copartisan President is in the White House. My analysis supports this hypothesis, while suggesting that Democrats follow the opposite pattern, whereby they seek to place more constraints on bureaucrats under an opposite party President. This finding speaks to a broad literature on bureaucratic politics, and more specifically to the literature on bureaucratic control. I challenge a longstanding finding, the “Ally Principle” which states that legislators will place fewer constraints on ideologically aligned bureaucrats (whose agency heads are appointed by a copartisan President). I show that while this holds for Democrats, the opposite seems to be true for Republicans, at least in the U.S. Congress. In the future, I intend to conduct more validation of my measurement approach using human coders, and to refine my theoretical contribution in preparation for publication.

In my second chapter, I hypothesize that legislators signal where the legislation they introduce falls on the spectrum between serious policymaking and position taking efforts in the amount of legal and technical language they include in a given bill. My analysis suggests that this language does carry a signal about bill quality, as bills containing more legal details are more likely to advance out of committee, and to eventually become law. This chapter presents the most opportunity for future work, as I am stepping into a relatively open and less heavily explored area of study. While I am confident the my measurement model picks up legal and technical language, more work can be done to try and relate its use to the theoretical concept I am trying to measure. I am also confident that to tell the full story around the seriousness of lawmaking efforts, I will need to incorporate other outside
information such as data on press conferences, floor speeches, and other indicators in the text of a bill to form a more wholistic measure.

In my third chapter, which is joint work with Andreu Casas and John Wilkerson, we examine undertook the first large scale effort to identify hitchhiker bills in U.S. Congressional legislation. We found that when we accounted for these hitchhiker bills, Congress is significantly more productive than we might think based purely on recording the number of stand-alone bills that become law, and that the lawmaking process is also more inclusive when we account for hitchhiker bills. This paper is forthcoming in the American Journal of Political Science so our efforts in this research area have shifted to new collaborative papers. While we focussed exclusively on hitchhikers that were included in bills that successfully became law, our next paper examines the full universe of hitchhikers, and explores what makes some of them successful, and which bills are the most effective targets for these hitchhikers to catch a ride with. Future work will dive even deeper to examine partial hitchhikers (where only part of a bill is included in another bill), to further extend our understanding of the lawmaking process in Congress.

Finally, I want to touch on what writing this dissertation has left me with, what I have gained from the process. Most tangibly, it left me with a heavily refined dataset of U.S. Congressional bills at the section level (and the means to update that dataset). This continues to serve as the foundation for a wide variety of projects focussed around legislative text. As a side effect, the method I described for identifying bill sections ended up getting me my first industry job at Skopos Labs, and ultimately contributed to my getting hired by Facebook. The process of writing this dissertation also helped to shape my view of what
kind of social science I would like to do in my career, and really highlighted the value of my substantive and theoretical training in developing methods to study politics. This process also taught me a huge amount about the lawmaking process simply as a result of forcing me to read lots of bills. This experience has deepened my understanding of existing research and left me with decades worth of questions to ask and answer about lawmaking in the U.S. Congress. In conclusion, I feel that this process has truly put a cap on my graduate school experience, and challenged me to integrate everything I learned over twelve years of post-secondary education. Thank you for this opportunity.
Appendix A

Bill Section Types

• **Front Matter:** All bills begin with some form of standardized front matter. This will include the title and number of the bill, the date, and the sponsor, among other metadata. These sections always start the bill, and are ended by the start of the first proper section in a bill. These sections do not contain the actual details of a policy, and are therefore considered non-substantive. An example is provided below:

```
110th CONGRESS
2d Session

H. R. 5655
To amend the Internal Revenue Code of 1986 to expand and improve the
dependent care tax credit.
IN THE HOUSE OF REPRESENTATIVES
March 14, 2008
Mr. Weiner introduced the following bill, which was referred to the
Committee on Ways and Means
A BILL

To amend the Internal Revenue Code of 1986 to expand and improve the
dependent care tax credit.

Be it enacted by the Senate and House of Representatives of the
United States of America in Congress assembled,
```

• **Short Title:** The first proper section in a bill is often the “Short Title” section, and usually begins with “SECTION 1. SHORT TITLE.” or a similar header. This type of section also contains no policy detail. Occasionally, a bill will omit the short title section and “SECTION 1.” will actually contain policy details. However, the difference is easy to identify through examination of the title. An example is provided below:

```
98
```
SECTION 1. SHORT TITLE.
    This Act may be cited as the ‘‘Tribal Government Amendments to the Homeland Security Act’’.

- **Table of Contents:** 5-10% of bills contain a “Table of Contents” section, which is sometimes combined with the “Short Title” section. This section will simply list the titles of other sections, and does not contain any policy detail. A truncated example is provided below:

SECTION 1. SHORT TITLE AND TABLE OF CONTENTS.
(a) Short Title.--This Act may be cited as the ‘‘Electric Consumers Power to Choose Act of 1997.
(b) Table of Contents.--
Sec. 1. Short title and table of contents.
Sec. 2. Findings and purpose.
Sec. 3. Severability.
...

- **Document Structure:** In addition to the sections that discuss specific policy details, some bills (longer ones in particular) are broken up into TITLEs, Subtitles, PARTs, and Subparts. These typically split up a bill at a higher level, separating provisions dealing with watershed protection and forest fire prevention in an environmental appropriations bill, for example. TITLEs, Subtitles, PARTs and Subparts do not usually span more than a few lines, and almost never contain any policy content themselves, so they are generally considered non-substantive. However, in a small number of bills TITLEs are used instead of sections and in these cases are considered substantive. A few examples are provided below:

TITLE II--PUBLIC UTILITY HOLDING COMPANY ACT OF 1935
Subtitle A--Department of Agriculture

- **Sections:** The substantive “meat” of a bill is made up of sections. These are typically numbered an titled, such as “SEC. 3. STRENGTHENING ENFORCEMENT AGAINST INTIMIDATION OF WORKERS.”, and contain the overwhelming majority of substantive policy details. Sections range anywhere from a few sentences to over one hundred pages in length. In a typical bill, different aspects of the policy will be dealt with in different sections, providing a relatively comparable unit of analysis. However, there are also a number of special sections that do not typically contain policy details. These will typically start with a “SEC. XXX” header, but them contain a special title. An example of a substantive bill section is provided below:

---

1I chose not to treat Subparts as document delimiters in my final analysis, as I identified fewer than 50 valid cases across my entire corpus, they were all very short (one line or less), and are particularly difficult to distinguish from simple references to a subpart in the substantive text of the bill (e.g. “as in Subpart F”). My results are not sensitive to the choice to ignore subparts as document delimiters, and instead use the nearest PART or proper section header as a delimiter.
SEC. 5. CONSULTATION REGARDING PROTECTION OF CORAL REEF SPECIES.

The Secretary of State, in consultation with the Administrator of the United States Agency for International Development, the Secretary of the Interior, and the Secretary of Commerce, may initiate consultations with foreign governments which are engaged in, or whose citizens include persons engaged in, commercial operations which take coral reef species, for the purpose of--

– **Findings:** These sections spell out a series of facts and opinions, typically in support of policies detailed in other sections in the same bill, but do not spell out those specifics themselves, and are therefore considered non-substantive. A truncated example of a findings section is provided below:

SEC. 2. FINDINGS.
The Congress finds the following:

(1) The generation of solid and hazardous waste has grown to alarming proportions in the United States. Each person in the United States throws away 3.6 pounds of garbage every day--enough annually to fill a convoy of 10-ton garbage trucks 145,000 miles long, which is the equivalent of half-way to the moon or roughly 7 times around the equator.

... 

– **Sense of Congress/Senate/House:** These sections typically express the opinion (or supposed opinion) of a chamber of Congress on some issue, but do not constitute binding policy. These sections are also considered non-substantive. A truncated example of a sense of Congress section is provided below:

SEC. 3. SENSE OF CONGRESS.

It is the sense of Congress that--

(1) emerging biological technologies, while providing the promise for unprecedented improvements in health and the environment, may pose a potential for harm, both intentional and accidental or inadvertent.

...

– **Effective Date:** These (typically) very short sections spell out when a policy or regulation is to become effective, and tend to use highly standardized language. Here again, these sections do not actually spell out policy specifics, and are therefore considered non-substantive. An example of an effective date section is provided below:

SEC. 4. EFFECTIVE DATE.

The amendments made by this Act shall take effect 180 days after the date of enactment of this Act.

– **Technical Corrections, Conforming Amendments:** These sections spell out any number of minor technical corrections, spelling fixes, date changes, document structure changes, etc. to an existing bill or law. While they often reference policy specifics, they do not prescribe substantively important changes to the
referenced policy, and are therefore considered non-substantive. A truncated example of an conforming amendment is provided below:

SEC. 13. OTHER TECHNICAL AND CONFORMING AMENDMENTS RELATING TO THE MERGER OF THE BIF AND SAIF.

(a) Section 5136 of the Revised Statutes.--The paragraph designated the ''Eleventh' of section 5136 of the Revised Statutes of the United States (12 U.S.C. 24) is amended in the 5th sentence, by striking ''affected deposit insurance fund' and inserting ''Deposit Insurance Fund'.

... 

– **Definitions:** These sections spell out general definitions of terms to be used in the rest of a bill. While some definitions can be critical to how a policy is applied (e.g. What is an “enemy combatant”?), these types of definitions tend to be overwhelmingly included as paragraphs in a standard bill section, as opposed to being split out into a different section. For the purposes of my analysis, these sections are also considered non-substantive. A truncated example of a definitions section is provided below:

SEC. 2. DEFINITIONS.

In this Act:

(1) Account. correspondent account. payable-through account.--The terms ''account, ''correspondent account, and ''payable-through account have the meanings given those terms in section 5318A of title 31, United States Code.

... 

– **Authorization of Appropriations:** These sections detail how much funding is to be allocated to the program(s) described in the rest of the bill. Typically, they are quite short and simply list out a few dollar amounts. In rare cases, they will contain much more detailed instructions about how the money is to be spent, including specific instructions to the implementing agency. Manual examination of these sections revealed that sections of this type containing fewer than 100 words were overwhelmingly of the former type, and were thus considered non-substantive, while the longer sections were considered substantive. An example authorization of appropriations is provided below:

SEC. 5. AUTHORIZATION OF APPROPRIATIONS.

For the purpose of carrying out this Act, there are authorized to be appropriated to the Attorney General $500,000 for fiscal year 2008 and such sums as may be necessary for each of fiscal years 2009 and 2010.

– **Purpose(s):** These sections provide a (mostly) plain English description of the purpose(s) of a bill or part of a bill. While they address the substance of a policy, they do not formally prescribe policy, and are therefore considered similar to Findings sections. For the purposes of my analysis, these sections are considered non-substantive. An example purpose section is provided below:
SEC. 2. PURPOSE.

The purpose of this Act is to set a date certain for replacing the Internal Revenue Code of 1986 with a simple and fair alternative.
Appendix B

Identifying Bill Sections

The system I developed for identifying bill section boundaries (and thus splitting bills into sections) and then classifying their types (Table 2.1) relied on a series of regular expressions (115). A regular expression is a way to compactly describe a class of character strings that the user would like to match. For example, to determine whether a section was of the type “Sense of the Senate/House/Congress”, I specified the following regular expression:

SENSE (OF|OF THE) (CONGRESS|HOUSE|SENATE)

which matches all strings of the form “SENSE OF CONGRESS”, “SENSE OF THE CONGRESS”, “SENSE OF THE HOUSE”, and “SENSE OF THE SENATE” (all case sensitive). I then applied this to the first line of each section to identify “Sense of XXX” sections. I then hand-checked a sample of matches to look for any mismatches or errors and then corrected these issues. As an illustration, most sense of the Congress sections have “SENSE OF CONGRESS” in the title, but a small number have “SENSE OF THE CONGRESS” in the title. These sections were identified through my hand validation process and added to my preprocessing pipeline as a result.
This regular expression based approach was applied to each section class, and to identifying the section boundaries themselves. In general, it tends to yield very high accuracy (above 99%), but there are some cases that it cannot handle. For example some bill section titles are of the form:

SEC. 2306. ORGANIZATIONS ACCREDITED TO CONDUCT INSPECTIONS UNDER TITLE III OF THE CLEAN AIR ACT...

The beginning of this section does not cause any trouble for the sectionizer, but the second line also matches the regular expression for the beginning of a Title section. Therefore, a simple rule-based approach would start a new section at the second line of the title, even though it is just a continuation of the section title from the line above. My solution was to write a program to go back and check for very short (one line) sections followed by a title (for example), and to combine these post-hoc. The combination of regular expression and length based rules brought the accuracy of the system to almost perfect, after approximate twenty iterations of validation.
Appendix C
Issue Area Fixed Effects
<table>
<thead>
<tr>
<th>Category</th>
<th>All Bills</th>
<th>5+ Cosponsors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Civil Rights</td>
<td>-0.125*</td>
<td>-0.156*</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>Defense</td>
<td>-0.080*</td>
<td>-0.056</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>Domestic Commerce</td>
<td>-0.042</td>
<td>-0.047</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td>(0.026)</td>
</tr>
<tr>
<td>Education</td>
<td>0.145*</td>
<td>0.115*</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>Energy</td>
<td>-0.025</td>
<td>-0.023</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>Environment</td>
<td>0.074*</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>Foreign Trade</td>
<td>-0.425*</td>
<td>-0.182*</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>Government Operations</td>
<td>-0.115*</td>
<td>-0.129*</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>Health</td>
<td>0.166*</td>
<td>0.144*</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>Housing</td>
<td>0.069*</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Immigration</td>
<td>-0.145*</td>
<td>-0.161*</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>International Affairs</td>
<td>-0.146*</td>
<td>-0.147*</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>Labor</td>
<td>0.019</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>Law and Crime</td>
<td>-0.100*</td>
<td>-0.117*</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td>(0.026)</td>
</tr>
<tr>
<td>Macroeconomics</td>
<td>-0.064*</td>
<td>-0.090*</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>Public Lands</td>
<td>-0.110*</td>
<td>-0.085*</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>Social Welfare</td>
<td>0.035</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.031)</td>
</tr>
<tr>
<td>Technology</td>
<td>-0.059*</td>
<td>-0.095*</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Transportation</td>
<td>-0.151*</td>
<td>-0.103*</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.030)</td>
</tr>
</tbody>
</table>

*Note:*  
*p<0.01

Table C.1: Parameter estimates for Democrats and Republicans. Base category is Agriculture legislation. Parameter estimate standard errors are presented in parentheses under the estimate.
The task of determining term-category associations has a long history, and methods for determining these associations have been profitably applied across a number of fields. For example, pointwise mutual information, or PMI (116; 117; 69) is a commonly used statistic for looking at the relationships between terms and categories in a corpus of documents (118; 70). PMI is calculated on a joint distribution, which can be easily formed from a contingency table by dividing each cell in the contingency table by the sum of the table. PMI tells us how much information knowing the particular value of a categorical variable (e.g. the probability of a particular vocabulary term appearing in a given category) gives us about which category we are in, and vice versa. For given values of discrete random variables $c \in C$ and $v \in V$, the PMI of these two values is defined as:

$$PMI(c; v) = \log \left( \frac{p(c, v)}{p(c) p(v)} \right)$$  \hspace{1cm} (D.1)$$

PMI is often used to find term collocations, or terms that tend to appear together (119; 120; 121), but can also be used to find words that are highly associated with a given category label (70). PMI, along with mutual information (the expected value of PMI over the entire joint distribution), have also been used extensively for feature selection\(^1\) in machine learning applications such as (122; 123; 124; 125; 126) and (127; 128; 129).

The task of feature selection is deeply related to the problem of identifying legal and technical language in legislation, except that the focus is placed primarily on finding terms that strongly distinguish between categories (policy terms). In the most directly relevant work, (130) introduce a weighted average PMI based approach to finding the optimal set of text features to distinguish between a set of categories. However, this approach is not ideal in the current application. The reason is that this method looks for terms that distinguish between all categories simultaneously, but in my application, I want to find

\(^1\)Feature selection involves picking features that produce good classification accuracy.
words that distinguish between Democrat and Republican sponsored legislation on average (controlling for issue area). Thus, this approach will likely include some substantive policy terms such as “increase spending” with legal and technical terms if they are used similarly across issue areas.

Turning to the political science literature, the “Fightin’ Words” feature selection method of (131) represents another promising approach to identifying legal and technical terms in legislation. Monroe et al. introduce a method for finding terms that statistically distinguish between documents written by members of different parties, which could also be used to find legal and technical terms (those that do not distinguish between parties, on average). However, because the Fightin’ Words feature selection method was specifically designed to find terms that do distinguish between categories, it does not produce a natural cutoff for identifying terms that actively make it harder to distinguish between categories. I build off of the intuition laid out in (131) to develop a statistical approach that is specifically tailored to identifying terms that make it harder to distinguish between categories.
Appendix E
ACMI Contributions

The mutual information of two discrete random variables \( C \) and \( V \) is defined as:

\[
I(C; V) = \sum_{c \in C} \sum_{v \in V} p(c, v) \log \left( \frac{p(c, v)}{p(c) p(v)} \right)
\]  
(E.1)

In this application, we are working with a large number of joint distributions formed by normalizing the counts of terms in Democrat and Republican bills, in a particular issue area. Our goal is to determine how much each term contributes to the mutual information of these conditional joint distributions, on average. Thus, we end up calculating the mutual information contributions of each term in these \( 2 \times V \) conditional joint distributions, and then averaging these contributions across all conditional joint distributions to get that term’s ACMI. Below, I formalize the method for calculating ACMI.

For reasons of notational sanity, let \( T \) be the full contingency table described in Section 3.3.3, and let \( T_k \), \( (k \in K) \) be the \( k \)'th Democrat/Republican row pair (on a given issue area) in this contingency table. So for example, one row pair could be: Democrat bills about nuclear energy policy; and Republican bills about nuclear energy (two rows from the full contingency table). Furthermore, let \( J_k \) be the conditional joint distribution implied by dividing all of the cells in \( T_k \) by the sum of \( T_k \) (normalizing). Now, let \( J_k^{(-v)} \) be a new conditional joint distribution generated by removing the \( v \)'th column from \( J_k \) and then re-normalizing. This new conditional joint distribution would have \( V-1 \) columns, and still sum to 1. Additionally, let \( \Sigma(T_k) \) be the sum of term counts in the \( k \)'th Democrat/Republican row pair, and \( \Sigma(T) \) be the sum of term counts in the full contingency table. Finally, we will denote the mutual information of the \( k \)'th conditional joint distribution as \( I(J_k) \).

Adopting the notation described above, the mutual information contribution of term \( v \) in conditional joint distribution \( J_k \) is just \( I(J_k) - I \left( J_k^{(-v)} \right) \). To get the full ACMI for term \( v \), we then just take the weighted average over its mutual information contributions in each \( J_k \).
as follows:

\[
\text{ACMI}(v) = \sum_{k \in K} \frac{\sum_{j \in J} T_k}{\sum_{j \in J} T} \left[ I(J_k) - I(J_k^{(-v)}) \right]
\]  \hspace{1cm} (E.2)

Here, the weight we give to the term mutual information contributions we calculate for each conditional joint distribution is just the proportion of all terms in the full contingency table. We do this type of averaging because we want to rely more on what we learn from category pairs that are associated with lots of tokens (documents), as opposed to those that are associated with only a few.
Efficient Optimization: Decomposing Mutual Information

To speed up the ACMI calculations, I decomposed the mutual information contribution of each term in the vocabulary. We start with the general form of mutual information, where $c \in C$ (categories) and $v \in V$ (vocabulary terms) are discrete random variables with joint probability $p_{c,v}(c,v)$ and marginal probabilities $p_c(c)$ and $p_v(v)$, respectively:

$$I(C;V) = \sum_{c \in C} \sum_{v \in V} p_{c,v}(c,v) \log \left( \frac{p_{c,v}(c,v)}{p_c(c) p_v(v)} \right)$$  \hspace{1cm} \text{(F.1)}

In the present application, the joint distributions will have two rows (Democrat and Republican term counts) and a large number of columns (terms in the vocabulary). What we want to find is the difference in mutual information if we remove one column (indexed by $v$) from the matrix:

$$\Delta(I)_{-v} = I(C;V_{-v}) - I(C;V)$$  \hspace{1cm} \text{(F.2)}

If $\Delta(I)_{-v}$ is positive, then term $v$ makes a negative contribution to mutual information in this distribution, and if $\Delta(I)_{-v}$ is negative, then it makes a positive contribution to mutual information in this distribution. We begin with the direct effect $DE_v$ of removing $v$, which is the removal of the following two terms from the sum:

$$DE_v = p_{c,v}(c_1,v) \log \left( \frac{p_{c,v}(c_1,v)}{p_c(c_1) p_v(v)} \right) + p_{c,v}(c_2,v) \log \left( \frac{p_{c,v}(c_2,v)}{p_c(c_2) p_v(v)} \right)$$  \hspace{1cm} \text{(F.3)}

These values can be cached in a straightforward way by recording their sum in a vector and simply subtracting that sum from the total mutual information with that column included. The challenging calculation comes in through the indirect effects $IE_v$ on the other values in the sum. $p_{c,v}(c,v)$ will be affected through the denominator (we are removing some number
of words from the corpus), so the denominator in each case will need to be multiplied by

$$D = \frac{\sum (C; V)}{\sum (C; V) - \sum v}$$  \hspace{1cm} (F.4)

For those terms outside of the log, this effect is common across all terms, so we can multiply the whole sum by $D$. We see that $p_{c,v}(c,v)$ also enters inside of the log, where we need to do the same multiplication. Fortunately, we can separate the log of a product into the sum of logs as:

$$\log(xy) = \log(x) + \log(y)$$ \hspace{1cm} (F.5)

so we can just take the sum of $\log(D)$ over the non-zero terms in $I(C; V_v)$. Finally, for the $p_c(c_i)$ terms, we will need to perform a similar multiplication:

$$D_c = \frac{\sum (c_i; V)}{\sum (c_i; V) - (c_i; V_v)}$$  \hspace{1cm} (F.6)

Thus we just need to keep track of the number of non-zero entries in the first and second rows ($NZ_1$, $NZ_2$), and we can use these counts to make a similar log addition adjustment. The critical point here is that the only entries we have to care about are the non-zero entries. We can therefore take advantage of sparsity in the category term distributions (most terms have zero count in a given distribution). Now we can use this decomposition of the effect of removing term $v$ from the vocabulary to efficiently calculate $I(C; V_v)$, which then gives us $\Delta(I)_{-v}$ (our objective).

$$I(C; V_{-v}) = D \times \left[ I(C; V) - D + \sum_i \sum_{I(C; V_{-v}) \neq 0} \log(D_c) \right]$$ \hspace{1cm} (F.7)

Using an $\{i, j, v\}$ sparse matrix representation of the document term matrix (where $i$ denotes the row index, $j$ denotes the column index, and $v$ denotes the non-zero value at $i, j$) makes this computation very efficient, facilitating fast vocabulary partitioning. After implementing this method, I compared the results of this method to the naive approach and confirmed that it yields identical results.
Appendix G
Pre-processing

This appendix describes our approach to identifying hitchhiker bills. We propose an original active, supervised-learning methodology that is tailored to studying legislative editing processes. As noted in the discussion, this new method offers research opportunities beyond the identification of hitchhiker bills. Its distinguishing attribute is the ability to create a wide variety of statistical features from a single, comparatively fast, algorithm. Software implementing the algorithm we use in this paper will be made available on publication.

As noted in the main text, we decided to exclude certain types of bills from our analysis. These included: private bills, duty suspension/tariff bills, and continuing appropriations bills. The problem in each case is that bills are very similar in content (often differing by just a word or two), so it is almost impossible to determine if a bill is a hitchhiker in these domains, or which bill was the “original” version of a law. We also exclude larger appropriations legislation because successful appropriations bills are always sponsored by Appropriations Committee leaders.

Research demonstrates that pre-processing decisions can have important consequences for prediction (132). Our pre-processing steps are tailored to the task at hand. Early on we discovered that stand-alone bills often contain language that is not retained when its policy provisions are incorporated into a law. To improve the fidelity of our bill-law comparisons, we systematically remove certain non-substantive content from each text:

- Exclude *Private, Duty Suspension/Tariff, and Appropriations* bills from the analysis.
- Remove the procedural head and tail of the bill (head = bill number, date, sponsors, etc. & tail = date, place of signature, etc.)
- Remove *Table of Contents*
- Remove *Findings, Definitions, and Authorization of Appropriations* sections.
• Remove a very frequent sentence: “Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled” from the text.

• Remove common procedural words (the top 100 words across all of the bills). Above this threshold, the word-distribution was essentially flat.

• Transform all text to lowercase.

• Remove all punctuation and numbers.

• Remove standard “stop words” (“the”, “and”, “it”, “we”, etc.).

Figures G.1 and G.2 illustrate the value of pre-processing. In Figure G.1, the left side contains the complete text of a bill, The Southern Nevada Limited Transition Area Act (sponsored by Dan Heller (R-NV)) while the right includes a portion of a much larger law, the Omnibus Public Land Management Act of 2009 (sponsored by Rush Holt (D-NJ)). The red text highlights the parts of each bill that match language in the other.¹ There is a lot of common text, but there is also a lot of non-matching text. In addition, some of the matching text (such as the very first part of the bill) does not seem particularly relevant.

Figure G.2 presents the same comparison after the texts have been subjected to our pre-processing protocol. The texts are now nearly identical.

¹Here we use a repeated n-gram algorithm WCopyFind (133) to define matching text
Table G.1: HR-146 bill insertion example. Matches highlighted in red.

<table>
<thead>
<tr>
<th>HR-408-IH</th>
<th>HR-146-ENR</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Congressional Bills 111th Congress] [From the U.S. Government Printing Office] [H.R. 408 Introduced in House (IH)] 111th CONGRESS 1st Session H. R. 408 To direct the Secretary of the Interior to convey to the City of Henderson, Nevada, certain Federal land located in the City, and for other purposes. IN THE HOUSE OF REPRESENTATIVES January 9, 2009 Mr. Heller introduced the following bill: which was referred to the Committee on Natural Resources A BILL To direct the Secretary of the Interior to convey to the City of Henderson, Nevada, certain Federal land located in the City, and for other purposes. Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, SECTION 1. SHORT TITLE. This Act may be cited as the “Southern Nevada Limited Transition Area Act”.</td>
<td>SEC. 2602. SOUTHERN NEVADA LIMITED TRANSITION AREA CONVEYANCE. (a) Definitions.--In this section: (1) City.--The term “City” means the City of Henderson, Nevada. (2) Secretary.--The term “Secretary” means the Secretary of the Interior. (3) State.--The term “State” means the State of Nevada. (4) Transition area.--The term “Transition Area” means the approximately 502 acres of Federal land located in Henderson, Nevada, and identified as “Limited Transition Area” on the map entitled “Southern Nevada Limited Transition Area Act” and dated March 20, 2006. (b) Southern Nevada Limited Transition Area.-- (1) Conveyance.--Notwithstanding the Federal Land Policy and Management Act of 1976 (43 U.S.C. 1701 et seq.), on request of the City, the Secretary shall, without consideration and subject to all valid existing rights, convey to the City all right, title, and interest of the United States in and to the Transition Area. (2) Use of land for nonresidential development.-- (A) In general.--After the conveyance to the City under paragraph (1), the City may sell, lease, or otherwise convey any portion or portions of the Transition Area for purposes of nonresidential development. (B) Method of sale.-- (i) In general.--The sale, lease, or conveyance of land under subparagraph (A) shall be through a competitive bidding process. (ii) Fair market value.--Any land sold, leased, or otherwise conveyed under subparagraph (A) shall be for not less than fair market value.</td>
</tr>
<tr>
<td>SEC. 2. DEFINITIONS. In this Act: (1) City.--The term “City” means the City of Henderson, Nevada. (2) Secretary.--The term “Secretary” means the Secretary of the Interior. (3) State.--The term “State” means the State of Nevada. (4) Transition area.--The term “Transition Area” means the approximately 502 acres of Federal land located in Henderson, Nevada, and identified as “Limited Transition Area” on the map entitled “Southern Nevada Limited Transition Area Act” and dated March 20, 2006.</td>
<td></td>
</tr>
<tr>
<td>SEC. 3. SOUTHERN NEVADA LIMITED TRANSITION AREA. (a) Conveyance.--Notwithstanding the Federal Land Policy and Management Act of 1976 (43 U.S.C. 1701 et seq.), on request of the City, the Secretary shall, without consideration and subject to all valid existing rights, convey to the City all right, title, and interest of the United States in and to the Transition Area. (b) Use of Land for Nonresidential Development.-- (1) In general.--After the conveyance to the City under subsection (a), the City may sell, lease, or otherwise convey any portion or portions of the Transition Area for purposes of nonresidential development. (2) Method of sale.-- (A) In general.--The sale, lease, or conveyance of land under paragraph (1) shall be through a competitive bidding process. (B) Fair market value.--Any land sold, leased, or otherwise conveyed under paragraph (1) shall be for not less than fair market value.</td>
<td>(C) Compliance with charter.--Except as provided in subparagraphs (B) and (D), the City may sell, lease, or otherwise convey parcels within the Transition Area only in accordance with the procedures for conveyances established in the City Charter. (D) Disposition of proceeds.--The gross proceeds from the sale of land under subparagraph (A) shall be distributed in accordance with section 4(e) of the Southern Nevada Public Land Management Act of 1998 (112 Stat. 2345). (3) Use of land for recreation or other public purposes.--The City may elect to retain parcels in the Transition Area for public recreation or other public purposes consistent with the Act of June 14, 1926 (commonly known as the “Recreation and Public Purposes Act”) (43 U.S.C. 869 et seq.) by providing to the Secretary written notice of the election.</td>
</tr>
</tbody>
</table>
Table G.2: HR-146 bill insertion example after pre-processing. Matches highlighted in red.

<table>
<thead>
<tr>
<th>HR-408-IH</th>
<th>HR-146-ENR</th>
</tr>
</thead>
<tbody>
<tr>
<td>cited southern nevada limited transition area conveyance notwithstanding federal land policy management et seq request city without consideration subject valid existing rights convey city right interest united transition area use land nonresidential development general conveyance city city sell lease otherwise convey portion portions transition area purposes nonresidential development method sale general sale lease conveyance land competitive bidding process fair market value land sold leased otherwise conveyed less fair market value compliance charter except paragraphs city sell lease otherwise convey parcels within transition area accordance procedures conveyances established city charter disposition proceeds gross proceeds sale land distributed accordance southern nevada public land management stat use land recreation public purposes city elect retain parcels transition area public recreation public purposes consistent june commonly known recreation public purposes et seq providing written notice election noise compatibility requirements city plan manage transition area accordance united code relating airport noise compatibility planning regulations promulgated accordance agree land transition area sold leased otherwise conveyed city sale lease conveyance contain limitation require uses compatible airport noise compatibility planning reversion general parcel land transition area conveyed nonresidential development reserved recreation public purposes years enactment parcel land discretion revert united inconsistent use city uses parcel land within transition area manner inconsistent uses specified discretion parcel revert united make election city</td>
<td>march southern nevada limited transition area conveyance notwithstanding federal land policy management et seq request city without consideration subject valid existing rights convey city right interest united transition area use land nonresidential development general conveyance city city sell lease otherwise convey portion portions transition area purposes nonresidential development method sale general sale lease conveyance land competitive bidding process fair market value land sold leased otherwise conveyed less fair market value compliance charter except subparagraphs city sell lease otherwise convey parcels within transition area accordance procedures conveyances established city charter disposition proceeds gross proceeds sale land distributed accordance southern nevada public land management stat use land recreation public purposes city elect retain parcels transition area public recreation public purposes consistent june commonly known recreation public purposes et seq providing written notice election noise compatibility requirements city plan manage transition area accordance united code relating airport noise compatibility planning regulations promulgated accordance agree land transition area sold leased otherwise conveyed city sale lease conveyance contain limitation require uses compatible airport noise compatibility planning reversion general parcel land transition area conveyed nonresidential development reserved recreation public purposes years enactment parcel land discretion revert united inconsistent use city uses parcel land within transition area manner inconsistent uses specified discretion parcel revert united make election clause</td>
</tr>
</tbody>
</table>
Appendix H
Constructing Statistical Features

Unfortunately, not all cases of true hitchhikers are as clean as the example above. Laws incorporating language from other bills often delete, add or rearrange the original language. Thus an approach for distinguishing these messier hitchhiker cases from other cases of shared language was needed. We initially experimented with off the shelf similarity algorithms before developing the new approach that is described here.

We first tokenized the pre-processed text of each document in a way that preserved information about word ordering. We then represent each document as a set of overlapping n-grams. Here we opt for five grams (e.g. “any land sold under this”) and a one word overlap. The tradeoff that must be made in terms of n-gram length is that longer n-grams (e.g. 50-grams) provide a tougher standard for shared text but open the door to more false negative predictions. Imagine two long documents that are identical except for every 50th word. A 50-gram approach will find no matches. Shorter n-grams (e.g. unigrams) will find the same two documents to be highly similar, but they open the door to false positive predictions. Imagine two documents that include the exact same words, but completely reversed. A unigram approach will conclude that the two documents are identical. Our decision to use 5 grams represents a middle ground approach.

We next record whether each 5-gram in a document has a match in the other document as a vector to retain information about each n-gram’s location in the document. One limitation of simply asking if each n-gram has a match is that two matches are recorded when (for example) “increase funding for this program” occurs 2 times in the first document but only once in the second. On the other hand, an approach that excludes matched n-grams would (in the same example) would arbitrarily conclude that the second occurrence does not have a match (even when it was the second that did have a match, in actuality).

The resulting vectors capture a lot of information about each document’s similarity to the other. Instead of simply comparing the proportion of n-grams that are shared, we can also compute statistics that also consider the locations of the shared n-grams. For example, we expect the matched n-grams of a hitchhiker to be located in a compact area of the law. The statistics computed for the current study are listed below (many more are possible).
bill\textsubscript{1} refers to the bill that did not become law, and bill\textsubscript{2} refers to the law.\textsuperscript{1} Note that below, n = 5 in all cases except the first bullet point.

- **Shared n-grams**: For each bill-law pair, we compute the simple proportion of shared n-grams in bill\textsubscript{1} that have a match in bill\textsubscript{2} and vice versa. We do this for unigrams, bi-grams, trigrams, 4-grams, 10-grams, and 20-grams (12 metrics in all). These statistics do not rely on the sequence based approach, and are instead supplemental.

- **Addition Scope**: This is calculated as the simple proportion of n-grams in bill\textsubscript{2} that do not have a match in bill\textsubscript{1}.

- **Deletion Scope**: This is calculated as the simple proportion of n-grams in bill\textsubscript{1} that do not have a match in bill\textsubscript{2}.

- **Scope**: This is calculated as mean of Deletion Scope and Addition Scope and gives a general characterization of the degree of difference between the two bills. The remaining statistics do leverage information about matching n-gram location.

- **Maximum Match Length (bill\textsubscript{1})**: The longest contiguous overlapping n-gram match in bill\textsubscript{1}. This captures the size of the “biggest chunk” of shared text in bill\textsubscript{2} from bill\textsubscript{1}.

- **Mean Match Length (bill\textsubscript{1})**: The mean length of contiguous overlapping n-gram matches in bill\textsubscript{1}.

- **Mean Match Length (bill\textsubscript{2})**: The mean length of contiguous overlapping n-gram matches in bill\textsubscript{2}.

- **Number of Matching Blocks (bill\textsubscript{1})**: The number of separate matching n-gram sequences in bill\textsubscript{1}.

- **Number of Non-Matching Blocks (bill\textsubscript{1})**: The number of separate non-matching n-gram sequences in bill\textsubscript{1}.

- **Number of Matching Blocks (bill\textsubscript{2})**: The number of separate matching n-gram sequences in bill\textsubscript{2}.

- **Number of Non-Matching Blocks (bill\textsubscript{2})**: The number of separate non-matching n-gram sequences in bill\textsubscript{2}.

- **Average Deletion Size**: The average length of non-matching sequences (the purple sequences in Figure 4.1) of overlapping n-grams in bill\textsubscript{1}.

\textsuperscript{1}Only bill versions published prior to the law’s enrollment date are considered.
• **Proportion of Possible Deletions**: The proportion of separate non-matching n-gram sequences in $\text{bill}_1$ relative to the possible separate non-matching sequences (if one token were different every n-gram size + 1 tokens).

• **Deletion granularity**: We start by dividing the average length of non-matching sequences (the purple sequences in Figure 4.1) by the total number of overlapping n-grams in $\text{bill}_1$. When this proportion is equal to one, none of the text of $\text{bill}_1$ is present in $\text{bill}_2$. When it is zero, $\text{bill}_1$ is identical to $\text{bill}_2$. To calculate the deletion granularity (from $\text{bill}_1$ to $\text{bill}_2$), we subtract this proportion from 1.
Appendix I
Active Learning with a Massive Ensemble

These statistics are then combined as features/variables in logistic regression models predicting whether a given bill was a hitchhiker on a given law. As discussed in the main text, the initial challenge was that there is no corpus of hitchhikers to train on so we needed to develop our own. The first step in this process was to use a simple bigram algorithm (Dice) to find all bill-law pairs where at least 80% of the bill’s unique words (after preprocessing) matched words in the law. This filter reduced the candidate pairs by 99% (from about 400 million to about 5 million). We then identified a single law that matched 164 bills at this 80% threshold level (HR-146, the Omnibus Public Lands Management Act of 2009).

One of the authors examined and labeled these cases (using WCopyFind) and found 89 of the 164 to be true hitchhikers. The next step was to use these 164 examples to train regression models to predict additional likely cases that could also be labeled and added to the corpus. We constructed over 1,500 different models using the statistics described above.\(^1\) We then trained these models on the initial corpus and used the best of them to predict additional likely hitchhiker cases.

In this first iteration, the 99 models that had precision and recall above 90% predicted 480 additional hitchhikers in the 111th Congress.\(^2\) Twelve graduate students, one undergraduate, and one faculty member then labeled these cases (once again using WCopyFind to visually compare how the texts overlapped). This was easier than expected as we observed perfect agreement for the 10% of cases labeled by two or more individuals.

We then retrained all 1,561 models using this larger corpus of 640 examples. In the second iteration, 39 models that exceeded the high performing threshold predicted just 5

\(^1\) All possible 1-to-3 variable combinations for a total of: \(\sum_{n=1}^{3} \frac{21!}{n!(21-n)!} = 1,561\) models.

\(^2\) Precision and recall are calculated using an n-fold approach that averages results across 300 partitions of the corpus into 80% train and 20% test sets.
additional cases. We labeled these cases and iterated the process two more times. The final ensemble of 22 high performing models - subsequently used to predict hitchhikers across all ten Congresses - had 92% precision and 95% recall. Closer inspection revealed that most of the false positive predictions (8%) were cases where a substantial portion (but not all) of the bill was in the law. The rest were very short bills that contained very similar language (such as duty suspension bills or continuing appropriations resolutions). The false negative cases (5%) tended to be cases where the annotator still judged it to be a hitchhiker case even though there was a fair amount of language difference between the overlapping text of the bill and law.

Table I.1: Summary of the Active Learning Process.

<table>
<thead>
<tr>
<th></th>
<th>Iteration 1</th>
<th>Iteration 2</th>
<th>Iteration 3</th>
<th>Iteration 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Corpus Size</td>
<td>164</td>
<td>644</td>
<td>649</td>
<td>651</td>
</tr>
<tr>
<td>True Positives &amp; Negatives</td>
<td>(P=89,N=75)</td>
<td>(477,167)</td>
<td>(481,168)</td>
<td>(483,168)</td>
</tr>
<tr>
<td># High Performing Models</td>
<td>99</td>
<td>39</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>New Hitchhiker Predicted</td>
<td>480</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Precision</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>95%</td>
<td>94%</td>
<td>95%</td>
<td></td>
</tr>
</tbody>
</table>

As a final step we used the same corpus of 650 labeled cases to compare the performance of several off the shelf algorithms. Their average recall was higher (99%) but their precision was lower.

3Cosine similarity, Dice coefficient, WDiff, Smith-Waterman, Needleman-Wunsch.
was much lower (75%). This indicates that compared to the other methods, our approach is conservative. It is much less likely to make false positive predictions (92% versus 75%) at the expense of making a few more false negative predictions (95% versus 99%).
Appendix J
Logistic Regression Models

In this appendix we first present a table of descriptive statistics for all the variables included in the logistic regression models presented in the paper. Then in the following figure we show the coefficients (and standard errors in parenthesis) for the two logistic regression models for which we plotted marginal effects in Figure 4.4. Finally, in Figure J.1 we explore potential significant heterogeneous effects across Congresses. We do not observe however any clear temporal trend, and despite some isolated exceptions, we find our findings to be robust across time (size and direction of the coefficients).

Figure J.1: Key coefficients of interest when estimating a separate model for each Congress.
<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>0.00</td>
<td>1.00</td>
<td>0.593</td>
<td>0.491</td>
<td>1.00</td>
</tr>
<tr>
<td>Committee Chair</td>
<td>0.00</td>
<td>1.00</td>
<td>0.06</td>
<td>0.237</td>
<td>0.00</td>
</tr>
<tr>
<td>Subcommittee Chair</td>
<td>0.00</td>
<td>1.00</td>
<td>0.087</td>
<td>0.281</td>
<td>0.00</td>
</tr>
<tr>
<td>Committee Rank Member</td>
<td>0.00</td>
<td>1.00</td>
<td>0.029</td>
<td>0.167</td>
<td>0.00</td>
</tr>
<tr>
<td>Subcommittee Rank Member</td>
<td>0.00</td>
<td>1.00</td>
<td>0.05</td>
<td>0.217</td>
<td>0.00</td>
</tr>
<tr>
<td>Committee Member</td>
<td>0.00</td>
<td>1.00</td>
<td>0.45</td>
<td>0.498</td>
<td>0.00</td>
</tr>
<tr>
<td>Years in Congress</td>
<td>0.00</td>
<td>51.00</td>
<td>13.899</td>
<td>10.567</td>
<td>4.00</td>
</tr>
<tr>
<td>Extremism</td>
<td>0.00</td>
<td>1.26</td>
<td>0.442</td>
<td>0.175</td>
<td>0.00</td>
</tr>
<tr>
<td>Bills Sponsored</td>
<td>1.00</td>
<td>232.00</td>
<td>36.105</td>
<td>27.828</td>
<td>16.00</td>
</tr>
<tr>
<td>Female</td>
<td>0.00</td>
<td>1.00</td>
<td>0.16</td>
<td>0.366</td>
<td>0.00</td>
</tr>
<tr>
<td>African American</td>
<td>0.00</td>
<td>1.00</td>
<td>0.058</td>
<td>0.233</td>
<td>0.00</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.00</td>
<td>1.00</td>
<td>0.021</td>
<td>0.144</td>
<td>0.00</td>
</tr>
<tr>
<td>Number of Co-sponsors (log)</td>
<td>0.00</td>
<td>6.07</td>
<td>1.398</td>
<td>1.462</td>
<td>0.00</td>
</tr>
<tr>
<td>Unified Congress</td>
<td>0.00</td>
<td>1.00</td>
<td>0.72</td>
<td>0.45</td>
<td>1.00</td>
</tr>
<tr>
<td>Gridlock Interval</td>
<td>0.38</td>
<td>0.65</td>
<td>0.56</td>
<td>0.09</td>
<td>-</td>
</tr>
<tr>
<td>Senate</td>
<td>0.00</td>
<td>1.00</td>
<td>0.353</td>
<td>0.478</td>
<td>0.00</td>
</tr>
<tr>
<td>Reauthorization bill</td>
<td>0.00</td>
<td>1.00</td>
<td>0.017</td>
<td>0.127</td>
<td>0.00</td>
</tr>
<tr>
<td>Revenue Bill</td>
<td>0.00</td>
<td>1.00</td>
<td>0.248</td>
<td>0.432</td>
<td>0.00</td>
</tr>
<tr>
<td>Companion Bill</td>
<td>0.00</td>
<td>1.00</td>
<td>0.025</td>
<td>0.156</td>
<td>0.00</td>
</tr>
<tr>
<td>Administration Bill</td>
<td>0.00</td>
<td>1.00</td>
<td>0.003</td>
<td>0.056</td>
<td>0.00</td>
</tr>
<tr>
<td>Congress</td>
<td>103.00</td>
<td>113.00</td>
<td>-</td>
<td>-</td>
<td>110.00</td>
</tr>
<tr>
<td>Major Policy Agendas Topic code</td>
<td>1.00</td>
<td>21.00</td>
<td>-</td>
<td>-</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Table J.1: Model data: descriptive statistics table
Table J.2: Results for two logistic regression models predicting whether a bill becomes a stand alone law (LAW) and, conditional on that not happening, whether it becomes law as a hitchhiker (HITCHHIKER). We include Congress (103 to 113th) and topic (Policy Agendas major topic) fixed-effects, although for simplicity we do not include the fixed-effect coefficients in the table.

<table>
<thead>
<tr>
<th></th>
<th>LAW</th>
<th>HITCHHIKER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>0.9093* (0.0748)</td>
<td>0.3487* (0.0626)</td>
</tr>
<tr>
<td>Committee Chair</td>
<td>1.1978* (0.0663)</td>
<td>0.7482* (0.0691)</td>
</tr>
<tr>
<td>Subcommittee Chair</td>
<td>0.6693* (0.0628)</td>
<td>0.3945* (0.0635)</td>
</tr>
<tr>
<td>Committee Rank Member</td>
<td>0.4587* (0.1536)</td>
<td>0.3094* (0.1229)</td>
</tr>
<tr>
<td>Subcommittee Rank Member</td>
<td>0.5354* (0.1345)</td>
<td>0.0701 (0.1131)</td>
</tr>
<tr>
<td>Committee Member</td>
<td>0.1742 (0.1138)</td>
<td>0.2828* (0.0865)</td>
</tr>
<tr>
<td>Committee Member x Majority</td>
<td>0.1065 (0.1277)</td>
<td>0.1986 (0.1031)</td>
</tr>
<tr>
<td>Years in Congress</td>
<td>0.0084* (0.002)</td>
<td>0.0054* (0.0019)</td>
</tr>
<tr>
<td>Extremism</td>
<td>-0.3276* (0.1226)</td>
<td>-0.7196* (0.1201)</td>
</tr>
<tr>
<td>Bills Sponsored</td>
<td>-0.0051* (0.001)</td>
<td>-0.0049* (9e-04)</td>
</tr>
<tr>
<td>Female</td>
<td>-0.2029* (0.066)</td>
<td>-0.0328 (0.0575)</td>
</tr>
<tr>
<td>African American</td>
<td>0.0284 (0.0993)</td>
<td>-0.1762 (0.105)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.3237* (0.1275)</td>
<td>0.1429 (0.1298)</td>
</tr>
<tr>
<td>Number of Co-sponsors (log)</td>
<td>0.0555* (0.0141)</td>
<td>0.0784* (0.0138)</td>
</tr>
<tr>
<td>Unified Congress</td>
<td>0.4257* (0.0535)</td>
<td>0.6473* (0.0548)</td>
</tr>
<tr>
<td>Gridlock Interval</td>
<td>3.2796* (0.2935)</td>
<td>1.2349* (0.2662)</td>
</tr>
<tr>
<td>Senate</td>
<td>0.042 (0.0529)</td>
<td>0.3455* (0.0517)</td>
</tr>
<tr>
<td>Reauthorization bill</td>
<td>0.9503* (0.0911)</td>
<td>0.5626* (0.1097)</td>
</tr>
<tr>
<td>Revenue Bill</td>
<td>-0.5211* (0.0739)</td>
<td>-0.1614* (0.0661)</td>
</tr>
<tr>
<td>Revenue Bill x Senate</td>
<td>-2.5615* (0.2927)</td>
<td>0.1637 (0.0929)</td>
</tr>
<tr>
<td>Companion Bill</td>
<td>1.5338* (0.0796)</td>
<td>0.8541* (0.091)</td>
</tr>
<tr>
<td>Administration Bill</td>
<td>2.6897* (0.1545)</td>
<td>2.2986* (0.197)</td>
</tr>
<tr>
<td>Constant</td>
<td>-6.9682* (0.2578)</td>
<td>-4.8159* (0.2106)</td>
</tr>
</tbody>
</table>

| N           | 84,913 | 82,009 |
| AIC         | 21,509 | 23,763 |

Note: *p<0.05
Appendix K
Hitchhikers Bills for two Target Law Examples
<table>
<thead>
<tr>
<th>Affordable Care Act (HR-3590) – 111th Congress</th>
<th>Financial Freedom Act of 1999 (HR-2488, 106th Congress)</th>
</tr>
</thead>
<tbody>
<tr>
<td>111-HR-1010</td>
<td>106-HR-1039</td>
</tr>
<tr>
<td>111-HR-1402</td>
<td>106-HR-1127</td>
</tr>
<tr>
<td>111-HR-1415</td>
<td>106-HR-1172</td>
</tr>
<tr>
<td>111-HR-1460</td>
<td>106-HR-1194</td>
</tr>
<tr>
<td>111-HR-1570</td>
<td>106-HR-1546</td>
</tr>
<tr>
<td>111-HR-20</td>
<td>106-HR-1616</td>
</tr>
<tr>
<td>111-HR-2006</td>
<td>106-HR-1703</td>
</tr>
<tr>
<td>111-HR-2223</td>
<td>106-HR-1914</td>
</tr>
<tr>
<td>111-HR-2301</td>
<td>106-HR-1955</td>
</tr>
<tr>
<td>111-HR-2358</td>
<td>106-HR-1986</td>
</tr>
<tr>
<td>111-HR-2525</td>
<td>106-HR-2018</td>
</tr>
<tr>
<td>111-HR-3138</td>
<td>106-HR-2400</td>
</tr>
<tr>
<td>111-HR-3242</td>
<td>106-HR-2416</td>
</tr>
<tr>
<td>111-HR-3256</td>
<td>106-HR-2430</td>
</tr>
<tr>
<td>111-HR-3468</td>
<td>106-HR-264</td>
</tr>
<tr>
<td>111-HR-3556</td>
<td>106-HR-423</td>
</tr>
<tr>
<td>111-HR-362</td>
<td>106-HR-487</td>
</tr>
<tr>
<td>111-HR-3648</td>
<td>106-HR-607</td>
</tr>
<tr>
<td>111-HR-3688</td>
<td>106-HR-630</td>
</tr>
<tr>
<td>111-HR-4313</td>
<td>106-HR-7</td>
</tr>
<tr>
<td>111-HR-444</td>
<td>106-HR-739</td>
</tr>
<tr>
<td>111-HR-479</td>
<td>106-HR-813</td>
</tr>
<tr>
<td>111-HR-756</td>
<td>106-HR-859</td>
</tr>
<tr>
<td>111-S-1022</td>
<td>106-HR-865</td>
</tr>
</tbody>
</table>

127
Bibliography


URL https://doi.org/10.1017/CBO9781107447752


URL http://www.jstor.org/stable/20680226


URL http://www.jstor.org/stable/27644332


URL http://link.springer.com/article/10.1007/BF00140770


URL http://prq.sagepub.com/content/60/4/607.short

132


URL http://www.jstor.org/stable/20647961


URL https://searchworks.stanford.edu/view/10745254

URL http://dl.acm.org/citation.cfm?id=584093

   URL http://aclweb.org/anthology/W14-3101


   URL http://www.jstor.org/stable/3647679


   URL http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=265371


URL http://www.jstor.org/stable/2647822


URL http://www.jstor.org/stable/40263477


URL https://doi.org/10.1371/journal.pone.0176999


URL http://www.cambridge.org/asia/catalogue/catalogue.asp?isbn=9781107023185{\&}ss=fro


URL http://prq.sagepub.com/content/58/2/295.


URL http://nlp.stanford.edu/fsnlp/

URL https://mitpress.mit.edu/books/transmission-information

URL http://www.aclweb.org/anthology/J90-1003

URL http://repository.cmu.edu/lti/212/


140


URL http://dl.acm.org/citation.cfm?id=2101264


URL http://plagiarism.bloomfieldmedia.com/wordpress/software/wcopyfind/
Vita
Matthew J. Denny

Employment
Research Scientist, Facebook Core Data Science (2019 - present)
Research Affiliate, Georgetown University McCourt School Massive Data Institute (2019 - present)

Education
Ph.D., Political Science and Social Data Analytics, Penn State University 2019
M.A., Political Science, Penn State University 2017
M.A., Political Science, University of Massachusetts Amherst, 2015
M.S., Resource Economics, University of Massachusetts Amherst, 2013
B.A., Economics, Commonwealth Honors College, University of Massachusetts Amherst, 2012

Select Publications

