The Politics of Bureaucratic Constraint: How Congress Uses Legal Language to Achieve Political Goals*

Matthew J. Denny†
Department of Political Science
Penn State University

April 22, 2017

Abstract

The traditional view of legislative control over bureaucratic officials is one of principals and agents, where limited time and expertise necessitate that legislators delegate some authority to these officials. This perspective recognizes the critical role for bureaucratic officials in developing and implementing policy. While it highlights inter-branch dynamics (who ultimately determines policy specifics), it places less emphasis on intra-branch dynamics (how variation in legislative control of the bureaucracy fits in to partisan conflict). In this study, I focus on the role of partisan politics in legislators' strategic decisions about what constraints to place on bureaucratic officials. To investigate this, I develop a text-based measure of the degree of constraint placed on administrative officials in a piece of legislation. My measure is based on an efficient decomposition of the mutual information contribution of legal "boilerplate" terms in the joint distribution over terms and document covariates in a given corpus of legislative texts. I apply this new measure to all bills introduced in the U.S. Congress between 1993 and 2014, and plan to validate it against expert legal coders. My findings suggest that Democrats use bureaucratic constraint to protect agencies whose missions align with their goals, while Republicans use it to advance a vision of smaller government.

Software: github.com/matthewjdenny/SpeedReader

*First version: April 2, 2017. This version: April 22, 2017. This research was supported by the National Science Foundation under IGERT Grant DGE-1144860.

†mdenny@psu.edu; 203 Pond Lab, Pennsylvania State University , University Park, PA 16802

1 Introduction

Members of Congress use a wide variety of tactics to achieve their policy goals. Some common tactics include: introducing legislation (Schiller 1995; Wawro 2001), leveraging leadership positions (Berry and Fowler 2016), exerting pressure on their peers (Bratton and Rouse 2011), appealing to the public (Grimmer 2013), and coordinating with interest groups (Hall and Deardorff 2006). An increasingly polarized and gridlocked Congress (see, for example: Binder 1999; Jones 2001; Poole and Rosenthal 2007) has only increased the pressure on legislators to find innovative strategies to advance their policy agendas. The tactics described above, and numerous others, have received consistent and fruitful attention from political scientists. Another important avenue for influencing policy outcomes involves strategically constraining or empowering specific bureaucratic agencies (see, for example: Epstein and O'Halloran 1999; Huber and Shipan 2002; Workman 2015). Here scholars have tended to focus on inter-branch dynamics (legislative control of the Bureaucracy), but have placed less of an emphasis on intra-branch dynamics (the use of specific constraints on, or delegation to, bureaucratic agencies as part of partisan congressional politics).

The thesis I develop in this study is that choices about legislative control over specific bureaucratic agencies are an important dimension of partisan conflict. To investigate this thesis, I collected a large corpus of U.S. congressional bills introduced between 1993 and 2014. I then developed a theoretically grounded, text-based method for identifying legalistic language in legislative texts, which I used to construct a measure of bureaucratic constraint at the level of individual bill sections. My findings indicate that Democratic and Republican lawmakers use bureaucratic constraint differently, and in line with their overall partisan goals. In particular, my findings suggest a more complex interpretation of the "ally principle"—that legislators will delegate more authority to bureaucrats officials under unified government (for a recent example, see: Palus and Yackee 2016)—in light of the party's different stances on the role of government (welfare state vs. small government). I find that when there is a Republican President, Democrats tend to use bureau-

cratic constraint in a manner that is consistent with protecting bureaucratic agencies in areas such as healthcare and education. On the other hand, Republicans tend to introduce more legislation aimed at constraining bureaucratic officials (across the board) when there is a Republican President than a Democratic one.

This study makes two main contributions. First, I highlight the way in which decisions about how much discretion to give bureaucratic officials are part of the larger political context, and fit in with existing partisan stances on the role of government. Second, I develop a flexible method for identifying legal boilerplate in legislative texts, and use it to create a set of "specificity" scores, which serve as a measure of the degree of constraint placed on bureaucrat's interpretation of a given bill section. This method is implemented in freely available software, and can be applied both to other legislative texts, but also more generally for finding terms that actively make it harder to distinguish between different categories of texts (e.g. parties, topics, time periods). Removing such "domain stopwords" can serve as an additional text preprocessing step, potentially leading to more interpretable results from common statistical text analysis methods (Handler et al. 2016).

I begin by discussing the literature on the relationship between Congress and the Bureaucracy, and situating my thesis about the role of partisan politics in that literature (Section 2). Next, I introduce a theory of legal language use in legislation (Section 3), and my method for finding such language in legislative texts (Sections 4, 5). The results of my analysis follow (Section 6), and I conclude with a discussion of their place in the literature (Section 7).

2 The Relationship Between Congress and The Bureaucracy

There is a rich literature seeking to model the interactions between legislators and bureaucrats (see, for example: Fiorina 1986; Mayton 1986; Calvert et al. 1989; Epstein and O'Halloran 1996a; Huber et al. 2001; Gailmard 2002; Spiller and Tommasi 2003; Moe 2006; Bressman 2007;

Gailmard 2009). Theoretical work has often conceptualized this relationship as a principal-agent problem (Krause 2011; Gailmard and Patty 2012). Legislatures must delegate some policy-making authority to the bureaucracy because legislators lack the time (Aranson et al. 1982; Kiewiet and McCubbins 1991), expertise (Alesina and Tabellini 2007), or political incentives (Alesina and Tabellini 2005) to develop precise solutions to all policy problems (Bawn 1995; Epstein and O'Halloran 1996a; Huber and Shipan 2002). However, agencies may not have incentives to enact policies in a way that the legislature would have preferred (Rourke and Others 1969; Wilson 1989; Meier and O'Toole 2006). Therefore, a great deal of scholarship has focused on whether, and the mechanisms by which the legislature can constrain the actions of the bureaucracy (Krause 2011). These mechanisms include congressional oversight (Aberbach 2001), limiting discretion (Huber and Shipan 2002), budgetary restrictions (MacDonald 2010), and the design of administrative procedures (Mccubbins et al. 1987; Moe 1989).

What these studies share in common is a theoretical perspective that (naturally) focuses on inter-branch dynamics more-so than intra-branch dynamics (particularly between members of the legislature). For example, there is a substantial literature that models legislative control of the Bureaucracy in terms of the divergence between the preferences of legislators and bureaucrats (see, for example: Calvert et al. 1989; Epstein and O'Halloran 1996b; Bendor and Meirowitz 2004; Krause 2011; Gailmard and Patty 2012). 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 bureaucratic 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.

But if the policy goals of legislators differ (which they do), then they may also prefer varying degrees of control over bureaucrats as a way to achieve those goals. For example, if Republican

lawmakers want reduced environmental regulation, they could explicitly advocate for a reduction in EPA funding, or pass legislation which precludes the EPA from regulating certain pollutants. But in addition, they may also place greater constraints on EPA officials even when introducing legislation on environmental protections they may generally agree with (like drinking water quality standards). These increased constraints will have the general effect of making the regulator's jobs more difficult, and reducing the overall scope of the agency in question.

More generally, if Republicans tend to favor a smaller and more constrained government than Democrats (Williamson et al. 2011), then we should expect that Republican lawmakers will tend to prefer a more constrained and controlled Bureaucracy than Democrats, all else equal. Writing for the journal *Regulation*, three years before he was nominated to the Supreme Court, Antonin Scalia provided a summary of this approach (emphasis added by the author):

> 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**. (Scalia 1983, p. 13)

This leads to my first hypothesis, that legislation introduced by Republicans will place more constraints on bureaucratic officials than legislation introduced by Democrats, on average.

However, these differences in preferred partisan constraints on bureaucratic officials will be mediated by which party controls the executive. Decades of research have shown how divided government effects the productivity of the Legislature, and its relationship to the Executive (Mayhew 1991; Howell et al. 2000; Bowling and Ferguson 2001; Binder 2015). The basic intuition is that when the Legislature and the Executive are controlled by different parties, they will have more trouble agreeing on policy and will therefore be less productive. In the context of bureaucratic politics, principal-agent theory suggests we should see some evidence of the "ally principle", whereby members of the party which controls the Executive will seek less constraints on the bureaucracy

than those of the opposition party. The idea is that members of the President's party can more easily trust that the executive branch will implement policies in a way they would have chosen, and in line with their political goals (Epstein and O'Halloran 1996b; Bawn 1997; Bendor and Meirowitz 2004). This allows them to write less specific, constraining legislation, and to delegate more implementation details to a "friendly" bureaucracy.

We should therefore expect that both Democrats and Republicans will seek to place relatively more constraints on bureaucratic officials when the Executive is controlled by the other party. Thus, while Republican lawmakers may prefer a more constrained bureaucracy in general, the "ally principle" tells us that they should prefer the most constraint under a Democratic President. This expectation seems reasonable enough, but some recent studies have found mixed to negative support for the ally principle (Gailmard and Patty 2012; Palus and Yackee 2016). In particular, Palus and Yackee (2016) find that state agency heads that are co-partisans with the majority party actually perceive slightly less policy discretion than agency heads who are members of the opposite party.

Taking account of partisan differences, and divided government paints a more complex picture of the political motives behind variation in legislative control of the bureaucracy, but there is one more important piece to the puzzle: variation across issue areas. It seems reasonable that Democratic lawmakers will have a different stance on oil drilling regulation (for example) than education policy. Furthermore, Williamson et al. (2011) argue that especially since the emergence of the Tea Party, the idea of monolithic opposition to government programs (and by extension the agencies that administer them) by Republican lawmakers is an oversimplification. The authors argue that "(Tea party Republicans) distinguish between programs perceived as going to hard-working contributors to US society like themselves and 'handouts' perceived as going to unworthy or freeloading people" (Williamson et al. 2011, p. 25). Therefore, we should expect Republicans and Democrats to moderate their stances of bureaucratic constraint based on their partisan views

on a particular issue. The question then becomes one of identifying which are important to which parties, and their stances on those issues.

Fortunately, the literature on "issue ownership" (see, for example: Therriault 2015; Petrocik et al. 2004; Petrocik 1996) provides an empirically grounded starting point for setting expectations about the partisan relationship between members of Congress and bureaucratic agencies. In particular, Egan (2013) provides a detailed analysis of party issue ownership from 1970-2011, relying on over 6,000 national survey questions to estimate issue ownership in each decade, and over the entire period. Table 1 depicts the issues that his analysis indicates are "owned" by Republicans (red) and Democrats (blue) on average, from 1970-2011. I then assigned each party a stance on these issues, to capture whether I expect them to seek more (negative) or less (positive) restrictive policies on that issue area. For example, I expect that Republicans will want to provide more leeway to bureaucrats on law enforcement issues, but less leeway on immigration.

Finally, if we put the influence of divided government on bureaucratic constraint in the broader context of partisan views on the role of government, and issue ownership, a full picture emerges. I expect that Republicans will tend to favor more restrictions on bureaucratic officials than Democrats, that members of the President's party will tend to favor less restrictions, and that each party will tend to want more restrictions on bureaucratic officials dealing with issues owned by the other party, or on which they have a negative stance. In many ways, these expectations simply follow from the idea that legislators seek to optimize their actions to meet their policy goals, but they deserve a more thorough explanation, which I provide below. In particular, I break my expectations down into four cases, based on the party of the Executive and members of the Legislature.

1. **Democratic President, Democratic Legislators:** I expect that Democratic legislators will seek to constrain agencies in the issue areas owned by Republicans (and those they have a negative stance on), but generally to place less constraints on bureaucratic agencies than their Republican counterparts. If, as Scalia argues, Democrat's goal is to govern, then we should expect that they will seek to empower the Bureaucracy to do just that, but still strategically place constraints on those agencies that align less with their partisan goals.

Issue	Margin	Stance
Domestic Security	14.5%	Positive
Military	13.9%	Positive
Immigration	8.5%	Negative
Inflation	8.5%	Negative
Crime	6.6%	Positive
Foreign Affairs	6.0%	Positive
Trade*	4.8%	Positive
Taxes	3.9%	Negative
Energy	2.6%	Negative
Education	10.4%	Positive
Jobs	12.0%	Positive
Health Care	12.4%	Positive
Social Security	14.4%	Positive
Environment	17.8%	Positive
Poverty	18.1%	Positive

Table 1: Issues (statistically significantly) owned by Republicans (red) and Democrats (blue), from Egan (2013, p. 67, *average* column). * Note that Trade was not statistically significantly Republican on average, but strongly Republican in the 2000's. **Margin** records the percentage margin for the given party, and **Stance** indicates the expected party stance towards bureaucratic discretion on the issue.

2. **Republican President, Democratic Legislators:** Here, I expect that as the opposition party, Democratic legislators will seek to protect agencies in the issue areas they own by introducing legislation that places more constraints on bureaucratic officials in those areas. The idea here is that if a Republican President is determined to dismantle the EPA (for example), they will appoint administrators who seek to reduce the overall scope and capacity of the agency. In response, Democratic lawmakers can attempt to prevent this reduction by placing lots of constraints on the agency head, forcing them to "do their job".

3. **Democratic President, Republican Legislators:** I expect that Republican legislators would like to constrain most bureaucratic agencies as much as possible when there is a Democratic President. This follows both from a general preference for smaller government, and a strategic incentive to limit an executive that has diverging policy goals. However, Republican lawmakers may also recognize that such legislation is unlikely to be successful, because the President can exercise their veto to prevent such constraints. This situation is different from the case of Democrats, because instead of trying to protect key agencies, they would simply be trying to constrain all of them. Such a situation could even create incentives to simply take more positions in legislation they introduce, while spending less effort crafting specific and constraining legal language than they would under a sympathetic Republican President.

4. **Republican President, Republican Legislators:** This is the instance where I expect the strongest efforts to constrain bureaucratic officials. With a sympathetic president, Republi-

	Democratic President	Republican President
Democratic Legislators	Reduce Constraints	Protect Key Agencies
Republican Legislators	Take Positions	Constrain the Bureaucracy

Table 2: Party strategies under Democratic and Republican presidents. Darker blue indicates less restriction on the Bureaucracy (in newly introduced legislation) while Darker red indicates more restriction on the Bureaucracy.

> can lawmakers are incentivized to coordinate on reducing the scope of bureaucratic agencies, particularly in policy areas owned by Democrats, and those that they have a negative stance on. Thus we see the "ally principle" flipped on its head — Republican lawmakers team up with their ally in the Executive to limit executive power.

These general viewpoints under presidents of both parties are illustrated in Table 2. To summarize, I expect that Democrat and Republican lawmakers will seek to place relatively more constraints on agencies in the issue areas owned by the opposite party, but that these patterns will be mediated (or even flipped) depending on which party controls the Presidency.

3 Using Specific Language to Constrain Bureaucrats

Legal language is the primary tool at a legislator's disposal in determining the degree to which bureaucratic officials are empowered or constrained by new legislation. Epstein and O'Halloran (1999) argue that there are two dimensions to this language. The first is language that explicitly delegates authority to bureaucratic officials, empowering them to directly set policy in some domain. The second is language that places constraints on bureaucratic discretion by requiring that they follow specific procedures, spelling out policy specifics, limiting the amount of time officials have to come up with a policy, etc. Huber and Shipan (2002) expand on the idea that legal language is a constraint on bureaucratic discretion by arguing that longer bills tend to be more constraining. The intuition here is that being more specific requires more words, so longer bills are more specific (and thus place greater constraints on bureaucratic discretion).

As an example, consider the following two passages taken from the Patient Protection and Af-

fordable 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))

These two passages encapsulate the core logic of Huber and Shipan, as it is reasonable to say that the first (much longer) passage is more restrictive than the second. Huber and Shipan (2002, ch. 3) provide a detailed justification, along with several case studies where they demonstrate that the length of (healthcare) legislation is inversely related to the degree of discretion it gives bureaucratic officials. However, their document length measure has some limitations, notably that it is primarily applicable to comparing single issue bills on the same issue (as Huber and Shipan suggest). This is because not all terms, phrases, and passages are created equal in terms of the degree to which they restrict or expand the discretion of bureaucrats in interpreting a piece of legislation. Additionally, with the passage of new laws and changes to existing laws over time, the bill length measure is also unlikely to be a valid comparison metric over time.

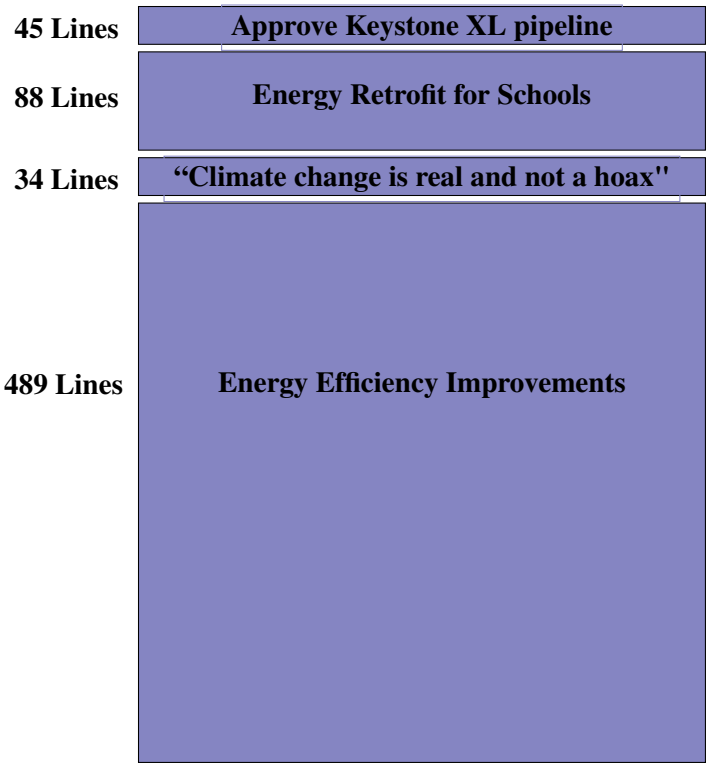
Furthermore, some legislation may cover many different substantive issues in the same bill (e.g. omnibus legislation), and the amount of text required to convey the purely substantive intent of piece of legislation may vary across issue areas. To illustrate the first point, consider the first bill introduced in the 114th Senate – 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 this bill, I found that the portion of the text related to approving the pipeline is actually quite small (approximately 5%)[1], and a number of other substantive provisions are included (see Figure 1). If we were to simply to count the number of words in this piece of legislation, we might conclude that it greatly restricted the discretion of bureaucrats with regard to the pipeline approval process; But the majority of the text in this bill is not even related to the pipeline. This has led some scholars to argue that full pieces of legislation are not the correct unit of analysis (Wilkerson et al. 2015), and sections should be considered separately.

Spurred by this example and many others, I take the approach of breaking bills into their constituent sections, and focusing on these individual sections as the unit of analysis. A given bill section is much more likely to focus on one aspect of a policy, so section length becomes a much more reasonable metric of bureaucratic constraint in this context. However, it could still be the case that even in a long bill section, most of the words are used to describe what the policy is about, as opposed to laying out specific constraints on the bureaucratic officials tasked with implementing it. Take, for example, Section 202 of the Dorothy I. Height and Whitney M. Young, Jr. Social Work Reinvestment Act (2013). This bill sought to improve the profession of social work, and this section dealt with grants to fund post-doctoral researchers.

> SEC. 202. RESEARCH GRANTS. (a) Grants Authorized.–The Secretary may award grants to not less than 25 social workers who hold a doctoral degree in social work, for post-doctoral research in social work– (1) to further the knowledge base about

[1]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.

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



effective social work interventions; and (2) to promote usable strategies to translate research into practice across diverse community settings and service systems. (b) Amounts.–The Secretary shall award the grants annually over a 4-year period. (c) Eligibility Requirements.–To be eligible for a grant under this section, a social worker shall– (1) demonstrate knowledge and understanding of the concerns of individuals and groups from different racial, ethnic, cultural, geographic, religious, linguistic, and class backgrounds, and different genders and sexual orientations; and (2) provide services and represent themselves as competent only within the boundaries of their education, training, licenses, certification, consultation received, supervised experience, or other relevant professional experience. (d) Minority Representation. At least 10 of the social workers awarded grants under subsection (a) shall be employed by a historically black college or university or minority-serving institution. (e) Authorization of Appropriations.–There is authorized to be appropriated $5,000,000 to the Secretary to award grants under this section.

Reading this section, it becomes clear that despite it relatively long length, it places very few constraints on the Secretary of Health and Human Services in establishing this program: they need to dispense at least 25 post-doctoral fellowships to qualified candidates (broadly defined), with at

least 10% of those fellowships going to minority-serving institutions. Thus, if we were to simply rely on bill section length as a measure of bureaucratic constraint, we might be lead to incorrect conclusions. This concern goes back to the core of what Huber and Shipan (2002) seek to measure with their document length approach: the amount of specific "legalese" in a document. From a measurement perspective, what we would really like to do is only measure the amount of specific legal language in each bill section and then use this as the basis for a measure of the degree of bureaucratic constraint implied in that section. I turn to the challenge of identifying legal language in the next section.

4 Finding Legal Boilerplate Using Information Theory

The text of a bill or bill section can be placed in one of three broad categories: (1) what the bill is about (topical content), (2) what the bill changes or creates (substantive policy actions), and (3) "boilerplate" (Yano et al. 2012; Wilkerson et al. 2015) — how the substantive policy actions are to be implemented, and other legal or administrative terminology. These categories are detailed in Table 3. In particular, the quantity of boilerplate language in a bill is exactly what Huber and Shipan (2002) seek to measure (by way of bill length) to get at the degree of bureaucratic discretion/constraint implied by a piece of legislation. So if we can distinguish between the different types of text in a bill, then we should be able to more accurately measure the latent concept of bureaucratic constraint, and use this measurement as the basis for a general, cross-domain metric of bureaucratic constraint for any piece of legislation.

For the most part, these categories of language are relatively distinguishable from one-another: a human coder will "know it when they see it". This lead Epstein and O'Halloran (1999) and Huber and Shipan (2002) (among others) to employ hand coding either as their primary measurement technique, or as validation. But human coding is not readily scalable to hundreds of millions of words, spread across hundreds of thousands of documents. Another approach would be to rely on human coding of a training set of bill sections to train a statistical model to estimate the de-

Category	Definition	Example
Topical Language	Confers information about the subject of the Bill.	{academic achievement standards}, {health care},{nuclear power plant}
Substantive Policy Language	Confers the action or policy change encapsulated in a piece of legislation or a particular provision.	{restrict abortion}, {reduce the deficit}, {criminalizing marijuana}, {Medicare benefit increases}
Boilerplate	Gives direction about legal interpretation or implementation.	{provision of this paragraph}, {specified in section}, {subsection (A)(1)}

Table 3: Qualitative definitions and examples of substantive policy language, topical language, and boilerplate in bill text.

gree of specific legal language used in the rest of the corpus. This is a relatively standard approach (Purpura and Hillard 2006), but still faces a several challenges. First, it still requires a large human-coded training set to provide good accuracy. Second, it may still be difficult to compare documents across issue areas because each issue may have a specific set of legal boilerplate terms, along with a shared pool of general legal boilerplate, so significant effort must be taken to ensure that these two categories are not conflated.

I take an alternative, more theoretically driven approach to identifying legal boilerplate language in legislative texts. It starts with a simple question: how do we expect legal boilerplate to be used in legislation? In the U.S. Congress, we expect that Democrats and Republicans will have different substantive policy goals, so substantive policy language should be used differently by members of each party. And topical language use should vary across issue areas. What we do not expect to change (between parties and to some extent, between issue areas), is the use of legal boilerplate. 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, their lawyers will likely need to do the same things to make this happen, regardless of party or issue area. This is made all the more likely be-

cause most law schools have a standard class on "administrative law" and/or "legislation", which goes over standard practices for crafting legal language in a bill (see, for example Eskridge et al. 2014).

To formulate my claim more precisely, I expect that: on average, in a given issue area and session of Congress (to account for changes in the U.S. Code over time), legal language will be used in the same way by Democrats and Republicans. Put another way, we would say that if we are given information about the use of a legal boilerplate word or phrase across documents, it will not help us distinguish what the document is about, or 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, information theory (Shannon 1948) provides a mathematical framework for formalizing the idea that legal boilerplate does not give information about issue and party categories, and thus for distinguishing legal boilerplate terms from topical and substantive policy language.

4.1 Average Conditional Mutual Information Vocabulary Partitioning

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 (Grimmer and Stewart 2013). 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, we can create a contingency table where each row records the counts of terms in bill sections sponsored by members of a given party, in a given issue area, in a given session of Congress. Intuitively, some columns (vocabulary terms) of these contingency tables will give us more information about which row we are in. For example, if we see a large count for the phrase "Affordable Care Act" in a particular row, it is likely that this row records the count of words in Republican sponsored bills about healthcare from 2013-2014

(when congressional Republicans repeatedly attempted to repeal it). However, other phrases may tell us far less, such as "create an oversight committee", because they appear in bills with many different substantive policy goals. What we want to know is how much information a term tells us about which category we are in.

If we think about our contingency table as a joint distribution over categories and vocabulary terms, we can calculate a measure on this joint distribution called its *mutual information* which quantifies how much information the vocabulary gives us about class labels, and vice versa. Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p_{x,y}(x,y) \log \left(\frac{p_{x,y}(x,y)}{p_x(x)\, p_y(y)} \right) \tag{1}$$

where p(x,y) is the joint probability distribution function of X and Y, and p(x) and p(y) are the marginal probability distribution functions of X and Y respectively. The mutual information of our contingency table will increase as we include more terms that distinguish between categories, and decrease as we include more terms that do not. Thus, we can use the mutual information of our contingency table as an objective function we seek to maximize to discover the partition of the vocabulary that tells us the most about which category we are in.

In the particular application considered in this study, a modification to this objective function will be necessary, as we wish to maximize the average mutual information between terms and categories (Democrat and Republican bill sections), conditional on the session of Congress, and major topic label. The categories in this case will be all sections from bills sponsored by members of each of the two major parties, within a major topic area and session of Congress. So for example, we would compare the aggregate term counts from all bill sections introduced by Democrats and Republicans, in legislation about Healthcare, introduced in the 110th Congress. The reason for disaggregating our data in this way is that some legal boilerplate terms may change in their usage

over time (perhaps with the advent of new technology, or because new laws are enacted which must then be referenced). Additionally, some terms may be boilerplate in some major topic areas, but actually be heavily partisan in others. Thus what we really want to find are terms that do not distinguish between parties *on average*, across each of these subdomains.

I introduce Average Conditional Mutual Information (ACMI) as a way to partition joint distributions under these criteria. Let $V : v \in \{v_1, v_2, ..., v_j\}$ be the vocabulary of unique terms indexed from 1 to J. Let the count of a particular term j in document i be denoted $v_{i,j}$. To capture the intuition outlined above, define an optimization problem over partitions of the vocabulary where we want to maximize the mutual information between the distribution over terms in the "substantive" partition S (topical and substantive policy language) and the conditional distribution over issue areas/terms. Those terms not included in the partition S will thus be labeled legal boilerplate. Let $c_{p,y} \in C$ be the subset of all bills that are sponsored by members of a particular party, in a particular session of Congress, and let $n_{p,y}$ be the number of bills in that subset. Furthermore, let the conditional distribution over K issue areas in that subset be $l_{p,y} \in L$. Then we can define the average conditional mutual information ACMI, between S and L as:

$$\text{ACMI}_j = \sum_{l_{p,y} \in L} \frac{1}{n_{p,y}} \sum_{k=1}^{K} \sum_{j=1}^{J} p_{v,l}(v_j, l_{p,y}^{(k)}) \log \left(\frac{p_{v,l}(v_j, l_{p,y}^{(k)})}{p_v(v_j) p_l(l_{p,y}^{(k)})} \right) \tag{2}$$

The maximization problem then becomes one over partitions in V (which terms to keep in the vocabulary):

$$\max_S \text{ACMI} \tag{3}$$

I do not need to place *a priori* constraints on the size or composition of S, as I do not have a clear theory about the relative size of the partitions. After applying this method, one is left with two partitions of the vocabulary, which can be used to identify boilerplate terms.

The approach described above focusses on distinguishing between partisan and non-partisan terms. Therefore, it is important to note another potential approach to finding partisan words —the "Fightin' Words" feature selection method of Monroe et al. (2008). Monroe et al. introduce a method for finding terms that statistically distinguish between documents written by members of different parties, which seems like a logical alternative method for finding boilerplate terms (those that do not distinguish between parties, on average). However, the feature selection method of Monroe et al. has two properties that make it less well suited for identifying boilerplate. First, it relies on a test for a statistically significant association between terms and categories. This test will place the highest confidence on estimates for terms that appear most frequently in the corpus, and therefore can end up treating common boilerplate terms as statistically significant predictors of category. This tendency can be controlled by the use of a strongly informative prior (Monroe et al. 2008, Section 3.5.1), but introduces a degree of freedom in the choice of strength of prior. Under the ACMI approach, the negative ACMI contribution serves as a natural threshold. Second, it is important to differentiate terms that do not distinguish between categories and those that actively make it harder to distinguish between categories (in an information theoretic sense). The Fightin' Words method will lump both types of terms together, while the ACMI approach will not (terms with a negative ACMI contribution will actively make it hard to distinguish between categories). Therefore, while the Fightin' Words feature selection of Monroe et al. (2008) is highly appropriate for finding partisan terms, it is less appropriate for finding legal boilerplate, as I have defined it.

5 Measuring Legal Constraints on Bureaucratic Officials (1993-2014)

My end goal was to calculate a text-based measure of the degree of legal constraint placed on bureaucratic officials in a given bill section. I selected a corpus of 97,428 U.S. Congressional bills (House and Senate) introduced in 103–113th sessions of Congress (January 3, 1993 – January 3, 2015). The collection and preprocessing of this corpus is described in detail in Supporting

Information A, but I provide a general overview of this process here. The corpus was originally collected from the congress.gov website by Handler et al. (2016). Bills were then broken up into sections, following Casas et al. (2017), and all non-substantive bill sections (table of contents, definitions, findings, etc.) were discarded. This resulted in 470,800 bill sections. These bill sections were then preprocessed into a standard unigram, and a phrase-based document-term matrix (Handler et al. 2016). I rely on the phrase based representation of the corpus in my analysis, because it better disambiguates the meanings of individual words (e.g. "social security" vs. "national security"), providing a more accurate basis measurement. The phrase based preprocessing resulted in a document-term matrix containing approximately 50,000,000 terms, with roughly 3,000,000 unique terms.

Each bill section was then linked to all available bill metadata (associated with the parent bill) from the Congressional Bills Project database (Adler and Wilkerson 2012). This dataset categorizes each bill into one of twenty one "major topic" categories[2] (which I refer to as **issue areas**), and includes information about the party of the bill sponsor, among other metadata. I took particular care when propagating the bill issue area labels, as a bill may contain sections that relate to multiple policy domains (Wilkerson et al. 2015). I collapsed the bill sections over their parent bill issue area labels to form a reference distribution over terms associated with each issue area, and the performed nearest-neighbor label propagation to all bill sections using these distributions (see Supporting Information B). This resulted in about 5% of bill sections changing labels (from their parent bill), with labor losing the most bill sections, and immigration and private bills gaining the most.

Having prepared the data for analysis, I then calculated ACMI contributions for every term (see Supporting Information C). A histogram of phrase ACMI contributions is illustrated in Figure 2.

[2]The categories are: "Agriculture", "Civil Rights", "Defense", "Domestic Commerce", "Education", "Energy", "Environment", "Foreign Trade", "Government Operations", "Health", "Housing", "Immigration", "International Affairs", "Labor", "Law and Crime", "Macroeconomics", "Private Bill", "Public Lands", "Social Welfare", "Technology", and "Transportation". See comparativeagendas.net/pages/master-codebook for a detail description of each label.

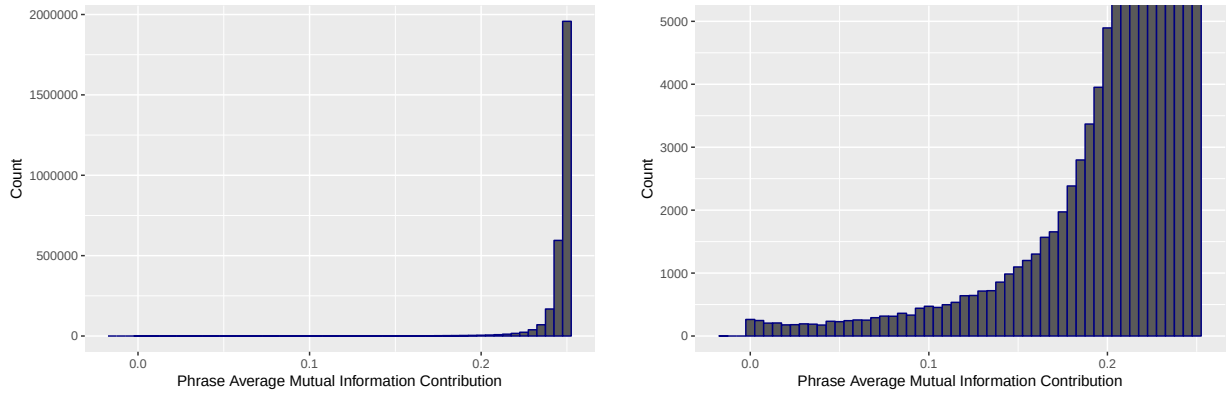


Figure 2: Histogram of phrase ACMI contributions. The figure on the left provides the full histogram, while the figure on the right truncates the y-axis at 5,000 to highlight lower contribution terms.

As we can see, most terms provide substantial information about topic and/or party, while a very small number of terms provided a small or negative ACMI contribution. More specifically, there were a total of 79 terms that had a negative ACMI score. These terms are displayed in Table 8 in Supporting Information D. Visual inspection of the terms indicated that they fit with the qualitative definition of legal boilerplate provided in Table 3. Terms like "defined in section", "striking subsection", and "inserting after paragraph" are all clearly legalese. However, as mentioned earlier in the paper, there may also be some issue-specific boilerplate terms, that the ACMI will not capture. Not accounting for these terms would bias any comparison between issue areas if these 79 terms are used differently across bill sections about those issues. Therefore, in forming the counts of legal boilerplate terms in each bill section, I also included all terms that had a negative mutual information contribution in at least 10 of 11 sessions of Congress for that bill section's issue area.

5.1 Specificity: A Measure of Legal Constraint in Legislation

Following Huber and Shipan (2002), one potential way to assess the degree to which a bill section places constraints on bureaucratic discretion is to simply count the number of boilerplate terms in that section. However, this may mischaracterize the degree of constraint in bill sections that also contain lots of topical or substantive policy language. This is because they may simply

need more legal boilerplate to meet basic formatting requirements if the bill section addresses a larger number of issues. To account for this potential issue, I define the *specificity* of a bill section (S), as the number of legal boilerplate terms in that section (including issue specific boilerplate), times the proportion of boilerplate terms in the section.

$$S = \#\text{ Boilerplate Terms in Section} \times \frac{\#\text{ Boilerplate Terms in Section}}{\#\text{ Terms in Section}} \tag{4}$$

Table 4 illustrates the correlations between the total number of terms in a bill section, the count and proportion of boilerplate terms, and the specificity score. As we can see, there is only a correlation of about 0.5 between the specificity score and the total term count, indicating that specificity is measuring a quantity that is unique from the raw term count.

	Boilerplate Terms	Boilerplate Prop.	Specificity	All Terms
Boilerplate Terms	1.00			
Boilerplate Prop.	0.08	1.00		
Specificity	0.81	0.22	1.00	
All Terms	0.86	-0.06	0.47	1.00

Table 4: Correlations between specificity and other potential measures of bureaucratic constraint. **All Terms** refers to the total term count in each bill section. **Boilerplate Prop.** refers to the proportion of boilerplate terms (relative to all terms), in each bill section, while **Boilerplate Terms** refers to the count in each section. Finally **Specificity** is just equal to **Boilerplate Prop.** $\times$ **Boilerplate Terms**.

The bill sections with the highest specificity scores will thus be those that both contain a large volume of legal boilerplate, and are also composed almost entirely of legal boilerplate. On the other hand, bill sections that contain a low volume of legal boilerplate but a high volume of topical and substantive policy language will receive the lowest specificity scores. Going back to the conceptions of bureaucratic constraint (or lack of discretion) discussed by Epstein and O'Halloran (1999) and Huber and Shipan (2002), a higher specificity score should be associated with more constraints (and less discretion) being placed on the bureaucratic officials tasked with implementing the policy.

Issue Area	Specificity Scores			Counts	
	Mean	Median	Max	Sections	Bills
Agriculture	6.0	2.5	229.0	13,584	2,682
Civil Rights	5.5	2.5	225.4	11,762	2,961
Defense	5.8	2.2	580.2	27,306	6,360
Domestic Commerce	6.5	2.7	1,390.3	34,911	6,843
Education	7.3	2.5	643.6	23,042	5,769
Energy	6.0	2.0	868.9	24,468	4,290
Environment	4.8	1.8	435.9	20,440	4,808
Foreign Trade	7.1	1.7	3,310.8	20,581	10,217
Govt. Operations	6.2	2.7	3,486.7	37,868	9,712
Health	10.1	4.0	486.7	57,135	11,427
Housing	7.0	2.8	488.8	11,129	2,595
Immigration	20.6	9.6	2,569.8	4,104	1,261
International Affairs	3.8	1.6	374.1	16,707	3,134
Labor	9.3	3.5	967.5	23,908	5,084
Law and Crime	6.2	2.6	3,348.0	33,209	6,242
Macroeconomics	12.1	4.7	904.1	23,423	6,685
Private Bill	6.6	3.2	385.7	4,836	3,190
Public Lands	4.2	1.5	1,214.7	36,141	9,083
Social Welfare	12.2	4.0	2,535.7	16,312	2,904
Technology	4.2	1.3	664.3	9,812	2,469
Transportation	6.3	2.2	2,735.6	20,122	3,902

Table 5: Descriptive statistics for bill section specificity scores.

Table 5 displays the mean, median, and maximum specificity scores for bill sections in 21 issue areas ("major topic" areas) as coded by Adler and Wilkerson (2012). As we can see from the table, technology related legislation has the lowest median specificity score, while immigration legislation has the highest. Looking at several bill sections with roughly median specificity scores (for that category) related to technology, I found that they mostly made general statements about how bureaucratic agencies should "promote research" and "seek academic and industry partnerships" to do basic science. On the other hand, when I looked at bill sections with roughly median specificity scores related to immigration, these tended to spell out in great detail how bureaucratic officials were to regulate H-1B visas, or other similar programs.

While the boilerplate terms recovered by the ACMI vocabulary partitioning method may seem

reasonable, and the specificity measure I define is theoretically defensible, it is still important to validate the scores against legal expert opinions. Therefore, I will be hiring several law students to read a large sample of bill sections and code them based on the relative degree of legal specificity and bureaucratic constraint implied in each section. I will use these coding results, along with a similar assessment of the legal boilerplate terms I detect to validate my specificity measure.

Beyond the need to human-validate the specificity scores I introduce, it is important to recognize some potential limitations and weaknesses of this approach. To begin with, it relies on metadata such as bill topic and the party of a bill's sponsor. If topics are incorrectly assigned to bill sections, this could potentially bias my results. Furthermore, changing the granularity or number of the topics used for calculating ACMI will also affect my results. I also only focus on the party of a bill's sponsor, but bills may have bipartisan initial cosponsors, and will likely collect cosponsors from both parties over time. A more nuanced approach would be to use more categories (e.g. Republican, Democratic, Bipartisan) or incorporate this information in an alternative approach. While I selected a particular approach to preprocessing my data that I felt was most appropriate for this task, my results are likely sensitive to these preprocessing choices (Denny and Spirling 2017). Adopting a fully supervised approach to detecting legal boilerplate might be able to avoid some of this sensitivity, while also catching corner cases of legal boilerplate language that my method might miss. Finally, my method is based on a theory about how legislators use language, and to the degree that they do not follow this theory, it will reduce the precision of my measurement method.

6 Analysis

In Section 2, I laid out three broad hypotheses about the relationship between Congress and the Bureaucracy. The first, was that Republicans will tend to seek more constraints on bureaucratic officials than Democrats, all else equal. The second, is that both Democrats and Republicans will make strategic choices about constraining bureaucratic officials based on issue ownership, and the

third is that all of these strategic choices will be filtered through the lens of which party controls the presidency. Below, I use the specificity measure developed in the previous section to test these hypotheses. Importantly, I continue to use bill sections as the unit of analysis, instead of aggregating up to bill themselves. This is the most appropriate approach, given that bills are often just vehicles for a collection of policy ideas (Wilkerson et al. 2015; Casas et al. 2017), and because they can deal with multiple issues.

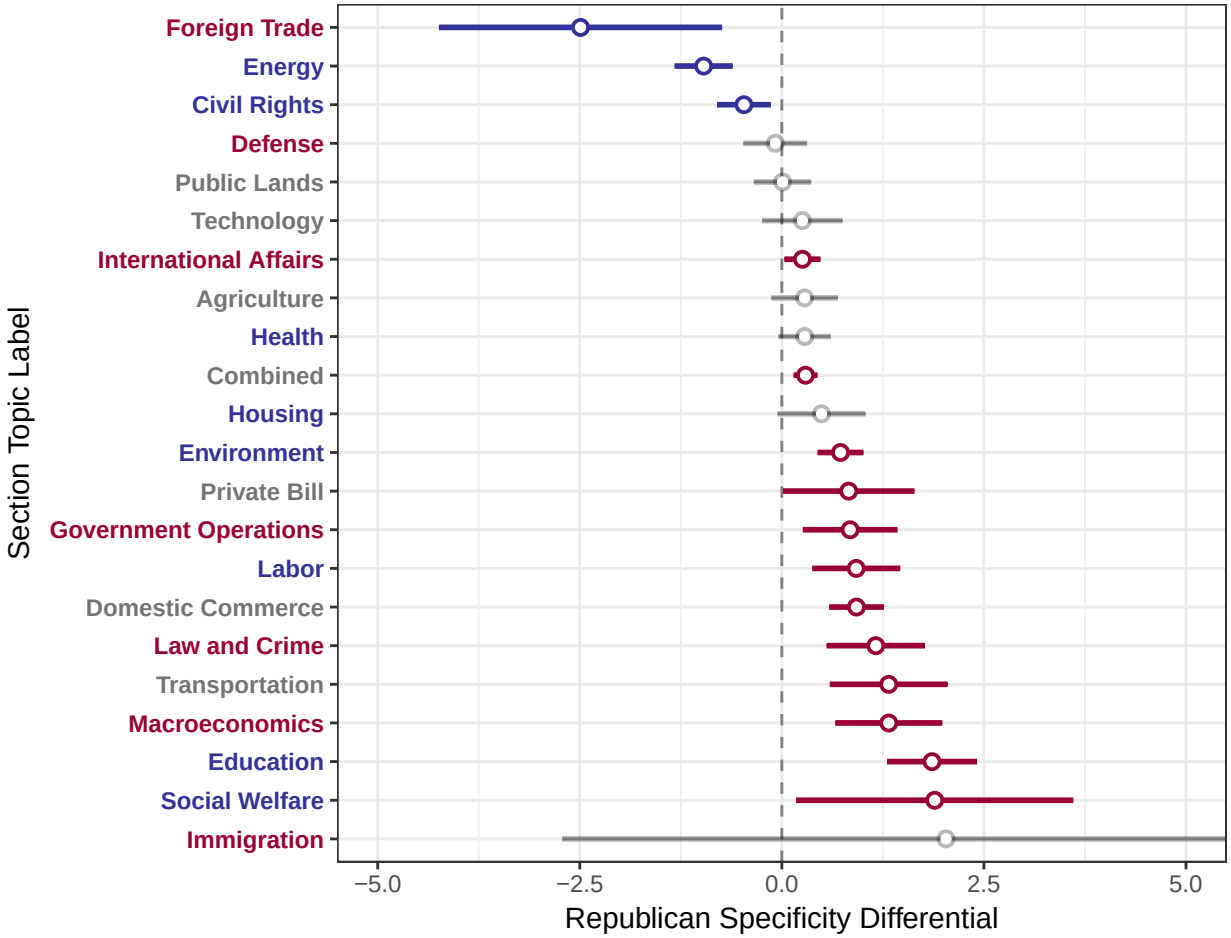


Figure 3: Party specificity differential by topic. Significant differences are highlighted in a darker color. Issues owned by Republicans (Democrats) are colored red (blue), and issues with ambiguous ownership are colored grey.

To explore the average overall constraints that each party places on bureaucratic agencies, I began by calculating average specificity score for sections belonging to all Democrat and Republican

sponsored bills, respectively. I also broke down these partisan differences by issue area, and the results are depicted in Figure 3. In this plot, each dot represents the party mean specificity differential for bill sections in that issue area, and the lines represent the 95% confidence interval for the differential. Dots to the right of the zero hash line indicate that sections from Republican sponsored bills in that issue area tend to have a higher specificity score, on average, than those from bills sponsored by Democrats. The dots and lines are colored red (blue) if that difference is statistically significant at the $\alpha = 0.05$ level (two-sided t-test).

The estimate for the "Combined" category represents the average difference across all bill sections, and as we can see, Republicans tend to use statistically significantly more constraining legal language than Democrats, on average. Moving on to issue area differences, Democrats tend to use significantly more constraining language in trade, energy, and civil rights legislation than their Republican counterparts. Going back to Table 1, we see that trade is a Republican issue during this period, while energy is an issue where Democrats would like to see more regulation than their Republican counterparts. The difference in civil rights specificity also makes a great deal of sense in the context of what such legislation is for: it is meant to ensure that regulatory agencies protect civil rights even in a hostile setting. Thus these results make a great deal of sense in the context of party issue ownership.

Turning to issue areas with a significant Republican specificity differential, we see a similar pattern emerging. For example, Republicans tend to introduce significantly more constraining legislation in issue areas owned by Democrats (Health, Environment, Labor, Social Welfare). Republicans also tend to introduce more constraining legislation in some issue areas they own, but where they would like to see a more constrained Bureaucracy (Macroeconomics and Immigration). Some of these differences do not make as much intuitive sense, such as Republicans preferring more constraining legislation on law and crime, but we have yet to look at these results in the context of which party controls the Presidency.

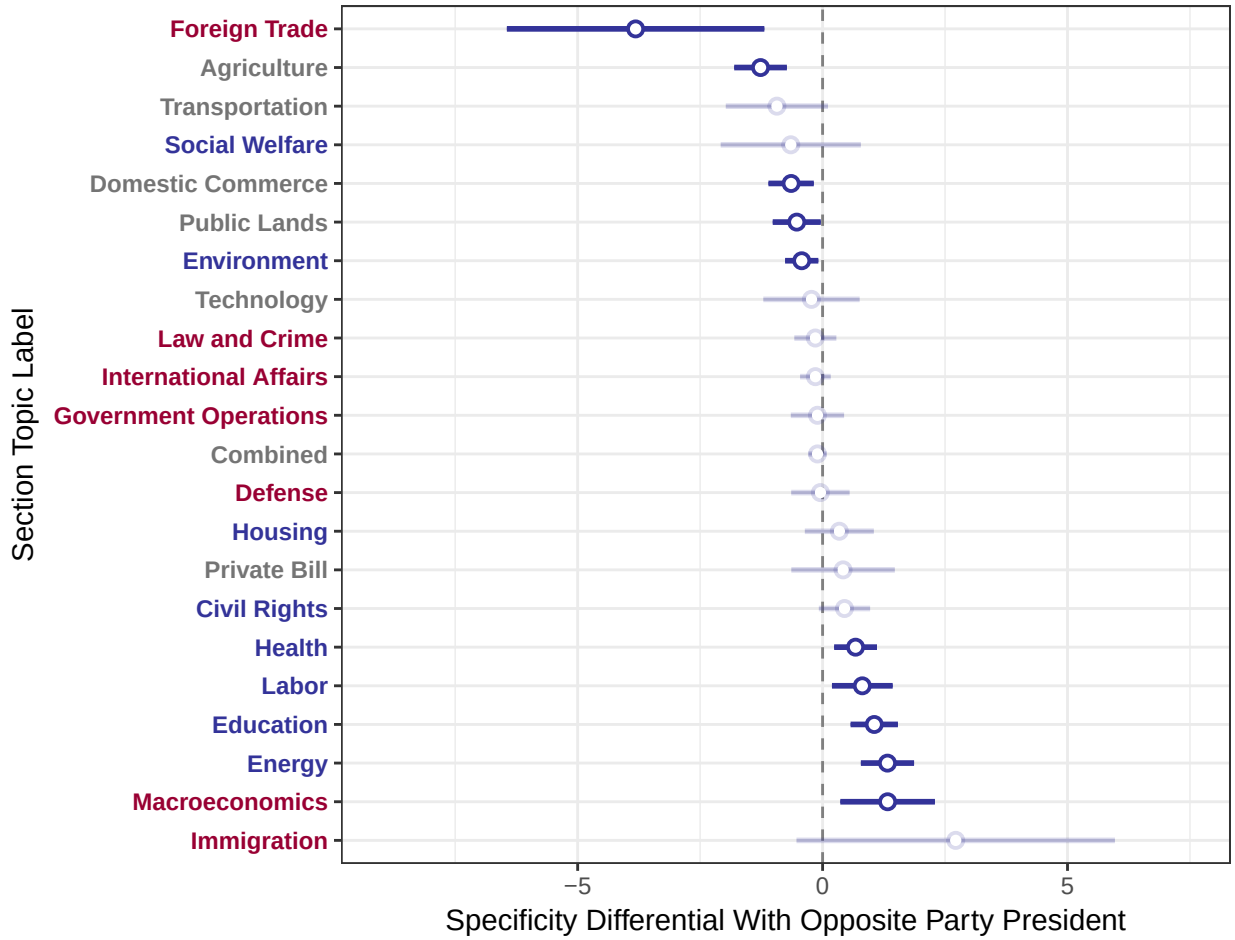


Figure 4: Specificity differential for bill sections introduced by Democrats with a Democrat President (negative) vs. Republican President (positive), by topic. Significant differences are highlighted in a darker color. Issues owned by Republicans (Democrats) are colored red (blue), and issues with ambiguous ownership are colored grey.

Figures 4 and 5 show party specificity differentials for Democrats and Republicans respectively, but now broken down by which party controls the Presidency. In each figure, a positive specificity differential means that members of that party tend to introduce legislation that places more constraints on bureaucrats when the other party controls the presidency, while a negative differential means they tend to place more constraints on bureaucratic agencies when their own party is in the White House. In these figures, the darker colored lines indicate a significant effect, while the lighter colored lines indicate non-significant differences.

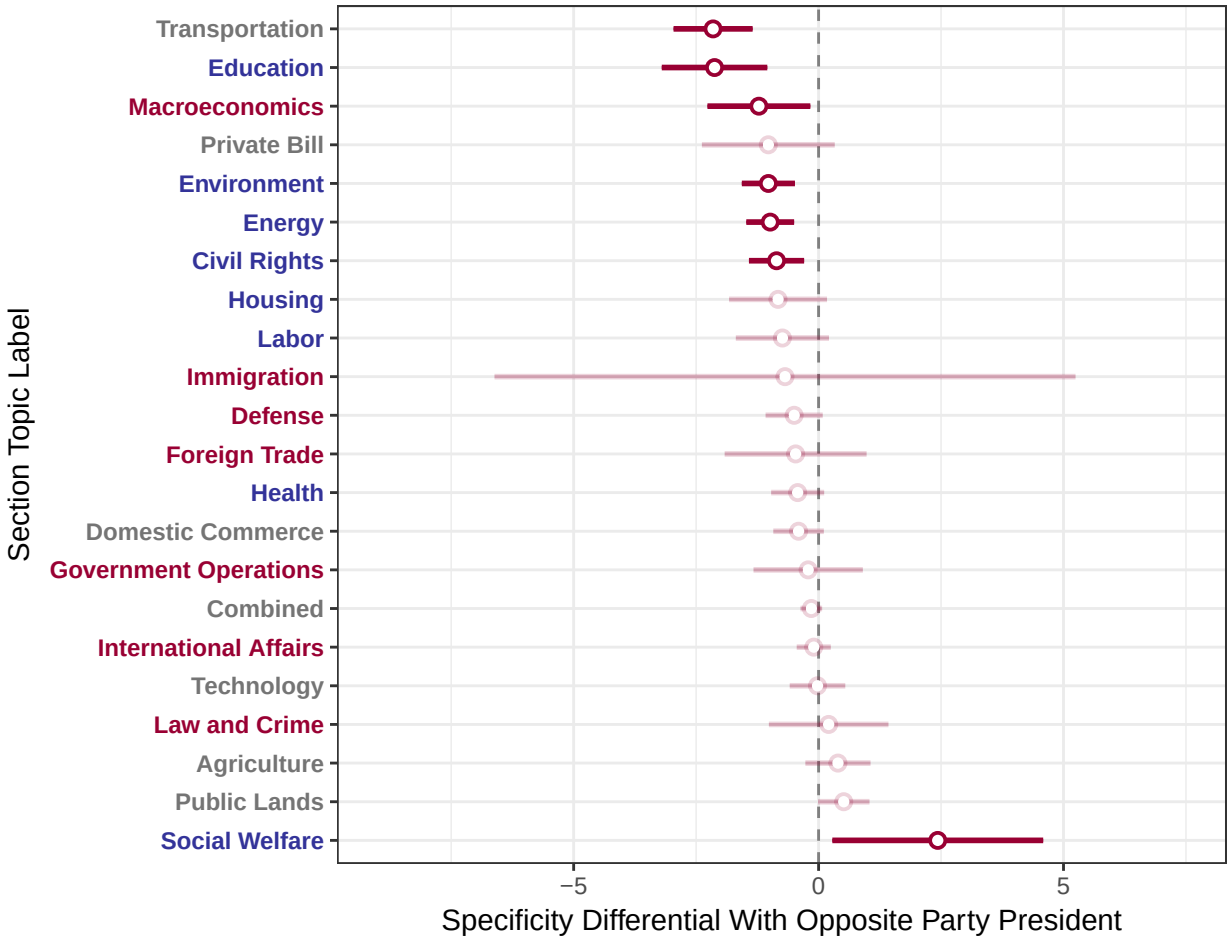


Figure 5: Specificity differential for bill sections introduced by Republicans with a Republican President (negative) vs. Democrat President (positive), by topic. Significant differences are highlighted in a darker color. Issues owned by Republicans (Democrats) are colored red (blue), and issues with ambiguous ownership are colored grey.

When disaggregating by which party controls the Presidency, more differences between Republicans and Democrats emerge. Democrats seem to place more constraints on bureaucrats in Republican issue areas when they have a Democratic President, and tend to place more constraints on bureaucrats in their own issue areas when there is a Republican President. This behavior is consistent with the theory outlined in Section 2, where Democratic legislators seek to strategically constrain bureaucratic officials in issue areas owned by Republicans when they have a sympathetic president, and to protect agencies in issue areas they own (by placing more constraints on them)

when there is a Republican President. However, on average (see the "Combined" differential) Democrats are no more likely to constrain bureaucratic officials when either party controls the Presidency.

Looking at the results for Republicans, a somewhat different picture emerges. Republican lawmakers seem to generally place more constraints on bureaucratic officials when they have a Republican President than under a Democratic President (although the difference is not statistically significant). They seem to focus primarily on constraining bureaucratic agencies that deal with issue areas owned by Democrats, while not doing much to "protect" agencies that deal with issues they own when there is a Democratic President. If we are to believe Scalia (1983, p. 13), then we can perhaps see this as an economy of effort on the part of Republican lawmakers (because they know it will be difficult to constraint the Bureaucracy under a Democratic President).

7 Discussion

The relationships between members of Congress and bureaucratic officials have received decades of attention from political scientists, and with good reason: they are 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 trading in information and expertise (Workman 2015), and members of Congress seeking to induce bureaucrats to implement policies in a desired way. This literature has captured an important dimension of strategic interaction between bureaucrats and members of Congress, but has paid less attention to the strategic, partisan dynamics influencing the constraints members of Congress seek to place on bureaucratic officials. In this study, I tied together the literature on Bureaucratic politics, partisan politics, and issue ownership to offer a more nuanced view of the relationship between members of Congress and bureaucrats. I then developed and validated a new measure of specificity in legislative texts, and used it to assess my theoretical predictions about partisan variation in bureaucratic constraint across issue areas and

divided government. My results indicate the Republican and Democratic lawmakers strategically constrain bureaucratic agencies as part of their overall strategy for achieving their partisan goals. Furthermore, I show that the two parties have fundamentally different approaches to applying constraints on bureaucratic agencies.

More specifically, my results indicate that Republicans introduced legislation that was more specific, and thus placed more constraints on bureaucratic agencies during the period under study (1993-2014). However, these constraints were not evenly distributed across all issue areas, but tended to be focused on issue areas typically "owned" by Democrats (education, environment, labor, social welfare, etc.). Furthermore, Republican lawmakers tend to attempt to place the most restrictions on bureaucratic agencies when they control the White House. As for Democrats, the results indicate that they tend to use constraints on bureaucratic officials in more of a defensive role: seeking to place constraints on bureaucratic officials on key issues they own when there is a Republican president, and on key Republican issues when there is a Democratic president.

There are a number of avenues for extending this work. First, I will make the specificity scores I calculate available for use in future studies. These could allow scholars to drill down further into the relationships between particular lawmakers and particular bureaucratic agencies. The ACMI partitioning method (and the attendant software) may also have applications outside of this study. Because the method I introduce for detecting boilerplate is extremely computationally efficient, it can easily be applied to a wide variety of corpora, or potentially as a preprocessing step before running a topic model. It could also be used to discover topical content or partisan policy language, providing a fast and theoretically based alternative to machine learning classifiers or topic models. Finally, my findings about the partisan and strategic politics of bureaucratic constraint suggest the need for renewed attention to the relationship between members of Congress and bureaucratic officials. Future work should examine this relationship at even greater granularity, and bring in campaign dynamics, for example.

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 (Epstein and O'Halloran 1999). Furthermore, I am limited to only looking at a 22 year period due to data availability issues. It would be valuable to extend the study back further to include more Republican Presidents, and forward to the Trump administration. It is also possible that a mixed approach combining my ACMI method and other machine learning classifiers might yield better results in identifying legal boilerplate. Additionally, it would be valuable to buttress the claims I make here (based on an aggregate level measure) with agency official interviews, to see how my measure relates to their experience.

References

Aberbach, Joel D. *Keeping A Watchful Eye: The Politics of Congressional Oversight*. Brookings Institution Press, 2001.

Adler, E. Scott and John Wilkerson. Congressional Bills Project: (1980-2012), 2012. http://www.congressionalbills.org/index.html.

Alesina, Alberto and Guido Tabellini. Why Do Politicians Delegate? 2005. http://www.nber.org/papers/w11531.

Alesina, Alberto and Guido Tabellini. Bureaucrats or politicians? Part I: A single policy task. *American Economic Review*, 97(1):169–179, 2007.

Aranson, Peter H, Ernest Gellhorn, and Glen O Robinson. Theory of legislative Delegation. *Cornell Law Review*, 68(1):1–67, 1982. http://scholarship.law.cornell.edu/clr/vol68/iss1/1.

Bawn, Kathleen. Political Control Versus Expertise: Congressional Choices about Administrative Procedures. *The American Political Science Review*, 89(1):62–73, 1995.

Bawn, Kathleen. Choosing Strategies to Control the Bureaucracy: Statutory Constraints, Oversight, and the Committee System. *Journal of Law, Economics, & Organization*, 13(1):101–126, 1997. http://www.jstor.org/stable/765129.

Bendor, Jonathan and Adam Meirowitz. Spatial Models of Delegation. *American Political Science Review*, 98(2):1–19, 2004. http://www.journals.cambridge.org/abstract{_}S0003055404001157.

Berry, Christopher R. and Anthony Fowler. Cardinals or Clerics? Congressional Committees and the Distribution of Pork. *American Journal of Political Science*, 60(3):692–708, 2016.

Binder, Sarah. The Dysfunctional Congress. *Annual Review of Political Science*, 18(1):85–101, 2015. http://www.annualreviews.org/doi/abs/10.1146/annurev-polisci-110813-032156.

Binder, Sarah A. The Dynamics of Legislative Gridlock, 1947-96. *American Political Science Review*, 93(3):519–533, 1999. http://www.jstor.org/stable/10.2307/2585572.

Bowling, Cynthia J. and Maragret R. Ferguson. Divided Government, Interest Representation, and Policy Differences: Competing Explanations of Gridlock in the Fifty States. *Journal of Politics*, 63(1):182–206, 2001. http://onlinelibrary.wiley.com/doi/10.1111/0022-3816.00064/abstract.

Bratton, KA and SM Rouse. Networks in the Legislative Arena: How Group Dynamics Affect Cosponsorship. *Legislative Studies Quarterly*, 36(3):423–460, 2011. http://onlinelibrary.wiley.com/doi/10.1111/j.1939-9162.2011.00021.x/full.

Bressman, Lisa Schultz. Procedures as Politics in Administrative Law. *Columbia Law Review*, 107(8):1749–1821, 2007.

Calvert, Randall L, Mathew D Mccubbins, and Barry R Weingast. A Theory of Political Control and Agency Discretion. *American Journal of Political Science*, 33(3):588–611, 1989. http://www.jstor.org/stable/2111064.

Casas, Andreu, Matthew J. Denny, and John Wilkerson. Legislative Effectiveness 2.0: A Method for Accurately Identifying the Bills that Become Law as Provisions of Other Bills. In *Midwest Political Science Association Anual Meeting*, 2017.

Denny, Matthew J. and Arthur Spirling. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Available at SSRN 2849145*, 2017. http://ssrn.com/abstract=2849145.

Denny, Matthew J., Brendan O. Connor, and Hanna Wallach. A Little Bit of NLP Goes A Long Way: Finding Meaning in Legislative Texts with Phrase Extraction. In *Midwest Political Science Association Anual Meeting*, 2015.

Egan, Patrick J. *Partisan Priorities: How Issue Ownership Drives and Distorts American Politics*. Cambridge University Press, New York, 2013.

Epstein, David and Sharyn O'Halloran. Divided Government and the Design of Administrative Procedures: A Formal Model and Empirical Test. *The Journal of Politics*, 58(02):373–397, 1996a. http://www.jstor.org/stable/2960231.

Epstein, David and Sharyn O'Halloran. Divided Government and the Design of Administrative Procedures: A Formal Model and Empirical Test. *The Journal of Politics*, 58(2):373–397, 1996b. http://www.jstor.org/stable/2960231.

Epstein, David and Sharyn O'Halloran. *Delegating Powers: A Transaction Cost Politics Approach to Policy Making under Separate Powers*. Cambridge University Press, 1999.

Eskridge, William N., Abbe R. Gluck, and Victoria F. Nourse. *Statutes, Regulation, and Interpretation: Legislation and Administration in the Republic of Statutes*. West Academic Publishing, St. Paul, MN, 2014. https://searchworks.stanford.edu/view/10745254.

Fiorina, Morris P. Legislator Uncertainty, Legislative Control, and the Delegation of Legislative Power. *Journal of Law, Economics, & Organization*, 2(1):33–51, 1986. http://www.jstor.org/stable/764915.

Gailmard, S. Expertise, Subversion, and Bureaucratic Discretion. *Journal of Law, Economics, and Organization*, 18(2):536–555, 2002. http://jleo.oupjournals.org/cgi/doi/10.1093/jleo/18.2.536.

Gailmard, Sean. Discretion rather than rules: Choice of instruments to control bureaucratic policy making. *Political Analysis*, 17(1):25–44, 2009.

Gailmard, Sean and John W. Patty. Formal Models of Bureaucracy. *Annual Review of Political Science*, 15(1):353–377, 2012.

Grimmer, Justin. *Representational Style in Congress: What Legislators Say and Why It Matters*. Cambridge University Press, 2013.

Grimmer, Justin and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, jan 2013. http://pan.oxfordjournals.org/cgi/doi/10.1093/pan/mps028.

Hall, Richard L and Alan V Deardorff. Lobbying as Legislative Subsidy. *American Political Science Review*, 100(1):69–84, 2006. http://www.jstor.org/stable/27644332.

Handler, Abram, Matthew J. Denny, Hanna Wallach, and Brendan O'Connor. Bag of What? Simple Noun Phrase Extraction for Text Analysis. In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science at the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016. https://brenocon.com/handler2016phrases.pdf.

Howell, William, Scott Adler, Charles Cameron, and Charles Riemann. Divided government and the legislative productivity of Congress, 1945-94. *Legislative Studies Quarterly*, pages 285–312, 2000.

Huber, John D. and Charles R. Shipan. *Deliberate Discretion? The Institutional Foundations of Bureaucratic Autonomy*. Cambridge University Press, 2002. https://doi.org/10.1017/CBO9780511804915.

Huber, John D, Charles R Shipan, and Madelaine Pfahler. Legislatures and Statutory Control of Bureaucracy. *American Journal of Political Science*, 45(2):330–345, 2001.

Jones, David R. Party Polarization and Legislative Gridlock. *Political Research Quarterly*, 54(1): 125–141, 2001. http://prq.sagepub.com/content/54/1/125.short.

Kiewiet, D Roderick and Mathew D McCubbins. *The logic of delegation*. University of Chicago Press, 1991.

Krause, George A. Legislative Delegation of Authority to Bureaucratic Agencies. *The Oxford Handbook of American Bureaucracy*, pages 1–44, 2011.

MacDonald, Jason A. Limitation Riders and Congressional Influence over Bureaucratic Policy Decisions. *American Political Science Review*, 104(04):766–782, 2010. http://dx.doi.org/10.1017/S0003055410000432.

Mayhew, David R. *Divided We Govern*. Yale University, 1991.

Mayton, William T. The Possibilities of Collective Choice: Arrow's theorem, Article I, and the Delegation of Legislative Power to Administrative Agencies. *Duke Law Journal*, 35(6):948–969, 1986. http://scholarship.law.duke.edu/dlj/vol35/iss6/2.

Mccubbins, Mathew D, Roger G Noll, and Barry R Weingast. Administrative Procedures as Instruments of Political Control. *Journal of Law, Economics, and Organization*, 3(2):243–277, 1987. http://www.jstor.org/stable/764829.

Meier, Kenneth J and Laurence J O'Toole. *Bureaucracy in a democratic state: A governance perspective*. JHU Press, 2006.

Moe, Terry M. The Politics of Bureaucratic Structure. In *Can the Government Govern?*, number September, pages 267–329. The Brookings Institution, Washington D.C., 1989.

Moe, Terry M. Political Control and the Power of the Agent. *Journal of Law, Economics, and Organization*, 22(1):1–29, 2006. http://www.jstor.org/stable/3555032.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16: 372–403, 2008.

Palus, Christine Kelleher and Susan Webb Yackee. Clerks or Kings? Partisan Alignment and Delegation to the US Bureaucracy. *Journal of Public Administration Research and Theory*, page muw001, 2016. http://jpart.oxfordjournals.org/content/early/2016/02/21/jopart.muw001.short?rss=1.

Petrocik, John R. Issue Ownership in Presidential Elections, with a 1980 Case Study, 1996. http://www.jstor.org/stable/2111797.

Petrocik, John R., William L. Benoit, and Glenn J. Hansen. Issue Ownership and Presidential Campaigning, 1952-2000. *Political Science Quarterly*, 18(4):599–626, 2004. http://www.jstor.org/stable/30035698.

Poole, Keith T. and Howard Rosenthal. On Party Polarization in Congress. *Daedalus*, 136(3): 104–107, 2007. http://www.mitpressjournals.org/doi/pdf/10.1162/daed.2007.136.3.104.

Purpura, Stephen and Dustin Hillard. Automated Classification of Congressional Legislation. *Proceedings of the 2006 international conference on Digital government research*, pages 219–225, 2006. http://www.purpuras.net/dgo2006PurpuraHillardClassifyingCongressionalLegislation.pdf.

Rourke, Francis E and Others. Bureaucracy, politics, and public policy. 1969.

Scalia, Antonin. Regulatory Reform: The Game Has Changed. *Regulation*, 5:13–15, 1983.

Schiller, Wendy J. Senators as Political Entrepreneurs: Using Bill Sponsorship to Shape Legislative Agendas. *American Journal of Political Science*, 39(1):186–203, 1995. http://www.jstor.org/stable/2111763.

Shannon, Claude E. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 1948. http://dl.acm.org/citation.cfm?id=584093.

Spiller, Pablo T and Mariano Tommasi. The Institutional Foundations of Public Policy: A Transactions Approach with Application to Argentina. *Journal of Law, Economics, and Organization*, 19(2):281–306, 2003. http://jleo.oxfordjournals.org/content/19/2/281.abstract.

Therriault, Andrew. Whose Issue is it Anyway? A New Look at the Meaning and Measurement of Issue Ownership. *British Journal of Political Science*, 45(4):929–938, 2015. https://doi.org/10.1017/S000712341400057X.

Wawro, Gregory. *Legislative entrepreneurship in the US House of Representatives*. University of Michigan Press, 2001.

Wilkerson, John, David Smith, and Nicholas Stramp. Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach. *American Journal of Political Science*, 59(4):943–956, 2015.

Williamson, Vanessa, Theda Skocpol, and John Coggin. The Tea Party and the Remaking of Republican Conservatism. *Perspectives on Politics*, 9(1):25–43, 2011.

Wilson, James Q. *Bureaucracy: What government agencies do and why they do it*. Basic Books, 1989.

Workman, Samuel. *The Dynamics of Bureaucracy in the US Government: How Congress and Federal Agencies Process Information and Solve Problems*. Cambridge University Press, 2015. https://doi.org/10.1017/CBO9781107447752.

Yano, Tae, Noah a Smith, and John D Wilkerson. Textual Predictors of Bill Survival in Congressional Committees. *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 793–802, 2012.

Supporting Information A Data Preparation and Preprocessing

The data used in this study were originally collected by Denny et al. (2015); Handler et al. (2016) from the congress.gov website, using URLs provided in the Congressional Bills Project database (Adler and Wilkerson 2012). The bill text was captured by scraping the web page associated with each bill. This also provided information about earlier versions of a bill (where applicable). These earlier versions were also collected. In this study, I make use of only the versions of bills as they were originally introduced in the House or Senate. This was done in order to better attribute the language in each bill to the bill sponsor and initial cosponsors, as later versions may have been amended by other members of Congress. This resulted in a total of 97,428 bills introduced in 103–113th sessions of Congress (January 3, 1993 – January 3, 2015).

Next, these bills were broken down into individual sections, which were the unit of analysis in this study. Importantly, I follow Casas et al. (2017) in removing a number of procedural bill sections, so that only substantive sections remain. First, the frontmatter such as title, date, and cosponsors are removed along with the end of the bill, which records the place of signature and other information. Additionally, the *Table of Contents*, *Findings, Definitions*, and *Authorization of Appropriations* sections were removed. The focus of this study is measuring the differential constraints legislators place on administrative officials, and these sections are likely to be almost entirely boilerplate without necessarily constraining those officials. A bill containing a long definitions section might be very specific, or just deal with numerous topics. While it may seem counterintuitive to focus only on substantive sections, these are precisely places we should look for constraints on the bureaucracy. This process resulted in a total of 470,800 substantive bill sections across 97,428 bills.

The bill sections were tokenized using the OpenNLP R package, which treats special characters such as punctuation and parentheses as separate tokens, while also attempting to keep terms like "White House" as a single term. After tokenization, a set of 1, 2, and 3-gram phrases were extracted using the "PhrasesNoCoord" grammar included in the development version of the phrasemachine R package (Handler et al. 2016). This grammar extracts both noun and verb-argument phrases, meaning that it will capture all categories of language described in Section 4. I 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"), however, some double counting is inevitable and is evident in the data. For this double counting to bias my results, it would have to disproportionately effect boilerplate or substantive terms, and inspection of the resulting phrases and counts gives no indication that this is the case.

In both cases, I chose to keep all tokens (phrases) I extracted and did not discard any character classes (such as punctuation). Denny and Spirling (2017) show that choices of whether to discard punctuation, stopwords, numbers, etc. can have significant effects on pairwise document distances, which are strongly related to the Mutual Information based measurement approach I take. Therefore, I deemed it most prudent to minimize preprocessing in order to maximize the raw information contained in the documents. The one step I did take was to lowercase all terms. This was done because I want to treat capitalized and lowercase version of the word similarly for the purposes of

my measurement approach. The phrase extraction algorithm does naturally exclude some character classes like punctuation because it does not match the POS filter. After this process, I was left with two representations of the corpus: *unigrams* and *phrases*.

One of the key assumptions underlying the ACMI vocabulary partitioning approach is that boilerplate terms are used in a similar (legalistic) way across different categories. This assumption necessitates further collapsing some classes of common boilerplate terms into a single category. For example, the *phrases* vocabulary contains hundreds of different bigram noun phrases of the form "section a", "section 27b", etc. Some of these might be specific to a particular bill, yet all bigrams of the form "section XXX" share a common referent (a block of text) and thus (I argue) should be treated as the same for the purposes of identifying boilerplate. Therefore, I created a number of term classes along the lines of "section XXX" to collapse these many different instantiations of a common term (e.g. "section") into one term in the vocabulary. Below, I list the unigram term classes for which this was done:

1. **Parentheses:** Parentheses and brackets were collapsed into their own special category because they most often indicate that a particular section, part, paragraph, etc. was being referenced ("part (a)", "section (1)(a)(iii)", etc.).

2. **Punctuation:** All other punctuation and special characters were collapsed into their own category.

3. **Numbers:** All numbers, including dollar amounts, comma separated numbers, and decimals.

4. **Text Unit Identifiers:** All numbers with a letter or letters attached at the end ("1024a", "27c", etc.), roman numerals ("xiv", "iii", etc.), and individual letters ("b","h", etc.). These are used to identify various, sections, subsections, titles, paragraphs, parts, etc. (e.g. "section 20b", "title iv", "paragraph c").

For phrases, this process was slightly more extensive:

1. **Numbers:** All numbers, including dollar amounts, comma separated numbers, and decimals.

2. **Text Unit Identifiers:** All numbers with a letter or letters attached at the end ("1024a", "27c", etc.), roman numerals ("xiv", "iii", etc.), and individual letters ("b","h", etc.). These are used to identify various, sections, subsections, titles, paragraphs, parts, etc. (e.g. "section 20b", "title iv", "paragraph c").

3. **Text Units:** All bi and trigrams of the form "section XXX" (e.g. "section 24", "section a") are collapsed into a single vocabulary term and their counts are combined. The same is done for subsections, titles, subtitles, parts, subparts, paragraphs and subparagraphs, each with their own collapsed vocabulary term.

4. **Time Units:** All bigrams of the form "XX days" are collapsed into a single vocabulary entry. The same is done for months years and "XX year", as a separate category, as "XX years" often refers to a term for which some provision will be active, while "XX year" often refers to a particular date.

Additionally, for the phrases, I also completely removed several classes of characters that were originally output by the phrase extraction algorithm. I did this because there is inherent ambiguity in POS tags, leading even the best POS taggers to occasionally make mistakes. No parentheses or punctuation should be captured by the filter, these were manually removed. I also removed several garbage unigrams that were classified as nouns by the POS tagger but are in fact XML or HTML escape characters, or the "com" in ".com". I also removed "ih" and "is", as these often refer to bill identifiers (S-27-IS, H-3486-IH, etc.), so their meaning cannot be disambiguated (especially in the case of "is").

1. **Parentheses:** Parentheses and brackets were removed.
2. **Punctuation:** All other punctuation and special characters were removed.
3. **Junk Terms:** The following "junk" terms were removed: "lt", "gr", "thn", "eq", "gt", "com", "sc", "ih", "is".

Descriptive statistics for each of these two representations are presented in Table 6. As we can see, using the phrase filter results in an almost 75% reduction in the number of terms in the document term matrix. It is also important to note that because the phrase extraction method I use relies on parts-of-speech tag patterns, different terms may receive different POS tags in different situations (depending on the context in which they are used) and there will be some inherent ambiguity in the tags for some tokens (because even state of the art POS taggers, like the one used here, only achieve approximately 97% accuracy). This means that there will be some differences in the number of terms such as numbers extracted by simply taking all unigrams and when using the phrase extraction method.

Supporting Information B Topic Label Propagation

One of the important assumptions underlying the method I introduce for identifying boilerplate language is that I have reasonably accurate major topic labels for each bill section. Such labels were assigned to each bill by Purpura and Hillard (2006) using a machine learning classifier trained on the manually annotated abstracts of each bill, and are included in the Congressional Bills Project metadata (Adler and Wilkerson 2012). One simple approach would be to directly propagate a bill's major topic label to each section associated with that bill. However, this will misclassify instances where a bill section addresses a different topic than the abstract of the bill (Wilkerson et al. 2015).

To the degree that bill sections are assigned incorrect topic labels, this is likely to lead my method to underestimate the probability that any given term is boilerplate, and possibly introduce bias. Underestimating the probability that a given term is boilerplate is likely because even some boilerplate terms may be used differently across different topic areas, therefore making them appear more informative if bill sections with lots of different major topics are mixed together. On the other hand, bias may be introduced if bills about some major topics are more likely to have sections included in them that are about other major topics (such as labor policy, which might also include sections about the economy, healthcare, civil rights, immigration, etc.) than others.

	Unigrams	Phrases
Number of Tokens	193,383,448	51,651,519
Unique Terms	64,519	2,932,647
Median Terms/Document	181	48
Mean Terms/Document	410.8	109.7
Parentheses	18,212,880	
Numbers	23,535,249	3,384,038
Punctuation	16,222,456	
Section IDs	4,941,924	2,867,481
Section XX		520,262
Subsection XX		19,846
Title XX		50,956
Subtitle XX		31,961
Part XX		47,169
Subpart XX		12,037
Paragraph XX		16,953
Subparagraph XX		7,080
XX Year		3,134
XX Years		78,974
XX Month(s)		35,301
XX Day(s)		94,846

Table 6: Descriptive statistics for unigram and phrase representations of the Bills Corpus. Term counts for special vocabulary categories are provided where applicable.

I address the issue of potentially mislabeled bill sections by performing nearest-neighbor label propagation. I begin by constructing empirical distributions over terms for each major topic label using direct label propagation from bills to sections. For example, I take the sum over all sections associated with bills whose major topic was "Health" for each term, then normalize, and this is the "reference distribution" for the "Health" topic. I construct one of these distributions for each of the 21 topics in my dataset. Next, I normalize the term counts in each bill section (so that they sum to 1), and calculate the distance between each section's term distribution and each topic's reference distribution. Calculating these distances is extremely computationally intensive, due to the large number of documents and vocabulary distribution size. Therefore, I adopted a lazy evaluation strategy that provides a many orders of magnitude speedup, while still capturing the relevant similarity between distributions.

More formally, let $\mathbf{r}^{(t)}$ be the reference distribution for topic $t \in T$, where each $r_v^{(t)}$ records the empirical probability of term $v \in V$ in topic t. Furthermore, let $\mathbf{s}^{(n)}$ be the empirical distribution over terms in bill section $n \in N$, and let N_v be the count of term v in the entire corpus. In this application, I use what I will refer to as the "one sided" inverse term-frequency weighted Euclidean distance (D) between $\mathbf{s}^{(n)}$ and $\mathbf{r}^{(t)}$. To calculate D, we only compare the non-zero entries in $\mathbf{s}^{(n)}$ to

their counterparts in $\mathbf{r}^{(t)}$ (one sided evaluation where we ignore entries in $\mathbf{r}^{(t)}$ that are not in $\mathbf{s}^{(n)}$), and weight the distance contributions by $1/N_v$.

$$D^{(t,n)} = \sum_{v \in \{\mathbf{s}^{(n)} \neq 0\}} \frac{\left(s_v^{(n)} - r_v^{(t)}\right)^2}{N_v} \tag{5}$$

This distance is efficient to calculate because lookups need only be done for the reference distribution and not the other way around, facilitating the use of a small hash-maps and making the problem $\mathcal{O}(NT)$ time complexity as opposed to $\mathcal{O}(NVT)$ (resulting in a several million times speedup in my application). D is decreasing in the similarity between the distributions and decreasing in the size of the probability mass in $\mathbf{r}^{(t)}$ that is not concentrated on terms in $\mathbf{s}^{(n)}$. Furthermore, terms that are very frequent in the corpus (like domain stopwords) provide a lower contribution to D than less frequently used terms. This highlights the importance of matching on less frequent and likely more informative terms.

For each bill section, D was calculated for each topic, using both unigram and phrase features. This entire process took under 30 seconds on a standard laptop[3]. A topic label was then selected for each bill section by choosing $t = \min(D^{(,n)})$. Results of this label propagation process are displayed in Table 7. The first numeric column on the left displays the count of number of sections assigned to each topic label using direct propagation of bill labels to their sections. The 0% column for unigrams and phrases displays the results of direct label propagation based on the distance metric described above. As we can see, phrase based label propagation produces results which are substantially more similar to the direct label propagation results (79.3% agreement) than unigram based label propagation (57.8% agreement). Manual inspection of the results suggests that this is because phrases tend to be more informative about a document's content that unigrams.

One potential problem with this approach is best illustrated with an example. Suppose a healthcare bill includes a section devoted to regulating health insurance. Such a section might use a number of health and finance related terms. It could be the case that the term distribution ends up being slightly closer to the distribution over terms in domestic commerce or macroeconomics legislation than health legislation. Thus, we might end up accidentally misclassifying a bill section that really should have remained in the topic category of its parent bill. To address this issue, I experimented with a small point mass prior on the topic label associated with a bill section's parent bill. The two priors I selected were 0.5% and 1%, meaning that $D^{(t,n)}$ (where t is equal to the topic label for the parent bill) was reduced by 0.5% and 1% respectively. In the case of unigrams, we can see that the application of even such a small prior brings the agreement rate to almost one, immediately which is undesirable. However, in the case of phrases, the priors increase agreement much more slowly.

I take a conservative approach and use the 1% phrase-based propagated topic labels as a basis for the rest of my analysis. The orange colored cells in Table 7 indicate a reduction of at least 10% in the number of sections assigned to a particular topic label (as compared to direct label propagation), while the blue colored cells indicate an increase of at least 10%. The largest trends we see

[3] Source code is **[available here]**.

Topic	Direct	Unigram-Based Propagation			Phrase-Based Propagation		
		0%	0.5%	1%	0%	0.5%	1%
Labor	28,309	9,188	28,219	28,281	19,220	20,194	23,908
Govt. Operations	38,995	25,189	38,901	38,981	32,332	36,508	37,868
Housing	10,353	16,054	10,368	10,355	13,020	12,039	11,129
Domestic Commerce	36,761	22,311	36,656	36,700	27,795	32,652	34,911
Energy	24,373	22,792	24,390	24,372	24,100	24,339	24,468
Social Welfare	16,425	15,045	16,411	16,395	19,092	16,204	16,312
Environment	20,872	18,367	20,847	20,857	20,638	19,945	20,440
Macroeconomics	21,167	29,504	21,182	21,156	22,793	24,984	23,423
Health	59,432	37,493	59,402	59,417	48,670	54,072	57,135
Public Lands	36,490	37,677	36,468	36,479	35,222	36,156	36,141
Law & Crime	34,613	25,872	34,607	34,621	29,771	31,899	33,209
Agriculture	12,945	17,661	12,942	12,939	15,875	14,090	13,584
Technology	9,178	20,272	9,173	9,173	13,230	11,263	9,812
Education	21,393	28,627	21,364	21,367	22,530	24,374	23,042
Transportation	20,124	19,771	20,110	20,114	20,697	19,989	20,122
Defense	28,167	21,297	28,160	28,180	26,751	26,330	27,306
Civil Rights	10,845	25,747	10,854	10,843	18,580	13,788	11,762
Foreign Trade	20,188	21,548	20,279	20,211	22,168	21,022	20,581
Intl. Affairs	16,469	26,961	16,458	16,460	19,955	17,816	16,707
Private Bill	2,162	16,367	2,229	2,182	8,399	6,750	4,836
Immigration	1,539	13,057	1,780	1,717	9,962	6,386	4,104
Agreement Rate:		0.578	0.999	0.999	0.794	0.884	0.948

Table 7: Counts of the number of bill sections assigned to each major topic label. **Direct** indicates counts for direct label propagation from the bill to its sections. Unigram and phrase based propagation indicate label propagation based on unigram and phrase distribution similarity to each reference category (based on bill labels). **0%** indicates topic counts for direct minimum distance propagation. 0.5% and 1% indicate topic label propagation results corresponding to point mass priors of 0.5% and 1% on the topic of the bill that a section came from. **Agreement Rate** indicates the agreement rate in section topic labels between direct label propagation and phrase or unigram based propagation with the appropriate prior.

are a reduction in the number of sections assigned to the **Labor** topic label, and an increase in the number of sections assigned to the **Private Bill** and **Immigration** topic labels. Manual inspection of the results indicated that a number of bill sections originally in the labor category were actually about worker visas, and that a number of provisions similar to those in private bills were inserted in bills ostensibly about other topics, across the board. In the future I intend to undertake a more extensive human-coding based validation of these results, but for now, I simply apply them in my downstream analyses.

Supporting Information C Efficient Optimization: Decomposing Mutual Information

Having prepared the text data for analysis, I needed to calculate the ACMI contribution for each term in the vocabulary of the phrase representation of the corpus. This would involve removing about three million unique terms, one at a time, from over 200 category pair distributions (Democrat vs. Republican sponsored bills in each of 11 sessions of Congress, for 21 different issue areas), and calculating the mutual information of the remaining distribution. This would prove to be prohibitively computationally expensive to do in a naive fashion (taking the better part of a decade of CPU time). To speed up this calculation, I decomposed the mutual information contribution of each term in the vocabulary to facilitate extremely efficient calculation of ACMI_j. This decomposition is described below, and resulted in a many orders of magnitude speedup (with computation taking a few-hundred seconds on a laptop). I was left with an ACMI contribution for each term in the vocabulary, as well as their contributions in each session of Congress/issue area pair.

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:

$$I(C;V) = \sum_{v \in V} \sum_{c \in C} p_{c,v}(c,v) \log \left(\frac{p_{c,v}(c,v)}{p_c(c)\, p_v(v)} \right) \tag{6}$$

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:

$$\Delta(I)_{-v} = I(C;V_{-v}) - I(C;V) \tag{7}$$

If $\Delta(I)_{-v}$ is positive, then term v makes and 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:

$$DE_v = p_{c,v}(c_1,v) \log \left(\frac{p_{c,v}(c_1,v)}{p_c(c_1)\, p_v(v)} \right) + p_{c,v}(c_2,v) \log \left(\frac{p_{c,v}(c_2,v)}{p_c(c_2)\, p_v(v)} \right) \tag{8}$$

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

$$D = \frac{\sum(C;V)}{\sum(C;V) - \sum v} \tag{9}$$

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:

$$\log(xy) = \log(x) + \log(y) \tag{10}$$

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:

$$D_c = \frac{\sum(c_i;V)}{\sum(c_i;V) - (c_i;V_v)} \tag{11}$$

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).

$$I(C;V_{-v}) = D \times \left[I(C;V) - D + \sum_i \sum_{I(C_i;V_{-v})\neq 0} \log(D_c) \right] \tag{12}$$

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.

Supporting Information D Boilerplate Terms

Tables 8 and 9 contain all terms identified as making an average negative contribution to mutual information across all categories for phrases and unigrams respectively. As we can see, in the case of unigrams, these terms are a mix of standard English stopwords (“a”, “and”, “of”, “by”, “etc”.) and legal boilerplate (section headers, “regulations”, “limitations”, etc.). In the case of phrases, these tend to be much more concentrated on legal boilerplate (“pursuant to section”, “striking subsection”, “provision of law”, etc.). The negative terms for each category are too numerous to display here but are available in the replication materials.

{general}, {pursuant to section}, {an}, {in}, {period beginning}, {provided in subsection}, {described in subsection}, {the}, {internal revenue}, {internal revenue code}, {revenue code}, {section shall be}, {described in paragraph}, {defined in section}, {accordance with section}, {the secretary}, {united states}, {for purposes}, {a}, {for}, {such section}, {states code}, {term}, {united states code}, {described in section}, {notwithstanding}, {amendment made}, {act shall be}, {such regulations}, {effective date}, {shall take effect}, {take effect}, {such act}, {such date}, {other provision}, {provision of law}, {date of enactment}, {sec}, {fiscal years}, {fiscal year}, {public law}, {one year}, {striking subsection}, {made by subsection}, {short title}, {act may be}, {year period}, {established under section}, {et seq}, {except}, {security act}, {act is}, {striking paragraph}, {new subsection}, {following new subsection}, {amendments made}, {striking the period}, {inserting after paragraph}, {new paragraph}, {following new paragraph}, {first sentence}, {day period}, {no}, {amended by section}, {year period beginning}, {amendment}, {NUMBERS}, {SECTION IDS}, {SECTION}, {SUBSECTION}, {PART}, {TITLE}, {SUBTITLE}, {PARAGRAPH}, {XX YEARS}, {XX MONTHS}, {XX DAYS}

Table 8: Phrase Boilerplate – All Categories

{sec}, {requirement}, {a}, {in}, {general}, {to}, {subject}, {section}, {an}, {shall}, {be}, {total}, {of}, {during}, {any}, {period}, {for}, {one}, {or}, {more}, {the}, {following}, {and}, {order}, {such}, {with}, {if}, {has}, {that}, {under}, {subparagraphs}, {paragraph}, {at}, {end}, {beginning}, {on}, {date}, {taken}, {subparagraph}, {subsection}, {not}, {by}, {unless}, {otherwise}, {may}, {necessary}, {taking}, {pursuant}, {this}, {result}, {amount}, {which}, {is}, {based}, {require}, {available}, {pay}, {than}, {except}, {as}, {provided}, {where}, {regulations}, {secretary}, {act}, {s.c}, {title}, {providing}, {paid}, {additional}, {required}, {without}, {part}, {nothing}, {provide}, {would}, {notice}, {case}, {less}, {days}, {before}, {take}, {duties}, {make}, {so}, {appropriate}, {same}, {are}, {number}, {both}, {request}, {manner}, {sufficient}, {it}, {states}, {within}, {regarding}, {purposes}, {time}, {given}, {who}, {will}, {their}, {reason}, {approved}, {information}, {limitation}, {basis}, {described}, {from}, {considered}, {protection}, {purpose}, {other}, {terms}, {conditions}, {prior}, {limitations}, {construed}, {have}, {been}, {each}, {receive}, {state}, {law}, {report}, {determines}, {after}, {receiving}, {percent}, {defined}, {code}, {specified}, {being}, {against}, {individual}, {made}, {person}, {relating}, {authority}, {ensure}, {provisions}, {accordance}, {annual}, {program}, {submit}, {there}, {enforcement}, {equal}, {cost}, {interest}, {clause}, {rate}, {was}, {determined}, {respectively}, {including}, {public}, {agency}, {federal}, {jurisdiction}, {costs}, {addition}, {payment}, {administrative}, {sections}, {special}, {years}, {into}, {united}, {later}, {last}, {determining}, {agencies}, {application}, {procedures}, {apply}, {private}, {definitions}, {term}, {means}, {laws}, {individuals}, {et}, {seq}, {exceed}, {only}, {respect}, {applicable}, {continue}, {least}, {determination}, {established}, {department}, {service}, {chapter}, {amended}, {adding}, {new}, {referred}, {but}, {government}, {months}, {meaning}, {authorized}, {office}, {services}, {upon}, {while}, {extent}, {consistent}, {carry}, {out}, {contents}, {funds}, {striking}, {establishment}, {conduct}, {availability}, {used}, {first}, {congress}, {members}, {enactment}, {follows}, {representatives}, {include}, {all}, {meet}, {thereafter}, {sentence}, {use}, {notwithstanding}, {effect}, {amendment}, {national}, {provision}, {requirements}, {effective}, {through}, {subsections}, {its}, {no}, {current}, {president}, {between}, {related}, {year}, {included}, {those}, {preceding}, {activities}, {assistance}, {development}, {authorization}, {whether}, {should}, {also}, {submitted}, {implementation}, {persons}, {day}, {amendments}, {consultation}, {amounts}, {regard}, {issue}, {paragraphs}, {process}, {conforming}, {technical}, {inserting}, {redesignating}, {short}, {cited}, {determine}, {system}, {located}, {form}, {were}, {implement}, {establish}, {written}, {enter}, {whose}, {received}, {increased}, {PARENTHESES}, {NUMBERS}, {PUNCTUATION}, {SECTION IDS}

Table 9: Unigram Boilerplate – All Categories