Computational Data Sciences and the Regulation of Banking and Financial Services

Sharyn O'Halloran, Marion Dumas, Sameer Maskey, Geraldine McAllister, and David K. Park

October 17, 2016

Abstract

The development of computational data science techniques in natural language processing (NLP) and machine learning (ML) algorithms to analyze large and complex textual information opens new avenues to study intricate policy processes at a scale unimaginable even a few years ago. We apply these scalable NLP and ML techniques to analyze the United States Government's regulation of the banking and financial services sector. First, we employ NLP techniques to convert the text of financial regulation laws into feature vectors and infer representative "topics" across all the laws. Second, we apply ML algorithms to the feature vectors to predict various attributes of each law, focusing on the amount of authority delegated to regulators. Lastly, we compare the power of alternative models in predicting regulators' discretion to oversee financial markets. These methods allow us to efficiently process large amounts of documents and represent the text of the laws in feature vectors, taking into account words, phrases, syntax, and semantics. The vectors can be paired with predefined policy features, thereby enabling us to build better predictive measures of financial sector regulation. The analysis offers policymakers and the business community alike a tool to automatically score policy features of financial regulation laws to and measure their impact on market performance.

1 Introduction

This paper combines observational methods with computational data science techniques to understand the design of financial regulatory structure in the United States. The centerpiece of the analysis is a database encoding the text of financial regulation statutes from 1950 to 2010. Among other variables, we identify the amount of discretionary authority Congress delegates to executive agencies to regulate the banking and financial services sector. The analysis requires aggregating measures from thousands of pages of text-based data sources with tens of thousands of provisions, containing millions of words. Such a large-scale manual data tagging project is time consuming, expensive and subject to potential measurement error.

To mitigate these limitations, we employ Natural Language Processing (NLP), such as parsing, Machine Learning (ML) techniques, such as Naive Bayes and feature selection models, and topic modeling, to complement the observational study. While none of these techniques alone are unique, the combination of these techniques applied to financial regulation is unique. These methods allow us to efficiently process large amounts of complex text and represent them as feature vectors, taking into account a law's topic, words, and phrases. These feature vectors can be easily paired with predefined policy attributes specified in the manual coding. The purpose of this paper is to analyze

how these methods can be used to build predictive models of financial regulation from the text of the laws.

In a previous paper, O'Halloran et al. (2015) analyze how traditional observational studies could be enhanced by computational analysis. The results show that feature selection models that combine observational methods with computational data science techniques greatly improve the accuracy of the measurements. We expand the previous analysis by computationally assigning laws to categories via topic modeling to further enhance the accuracy of the predictive model over feature selection models alone. Furthermore, we show how these techniques can be used to develop robust standard errors and thereby facilitate hypothesis testing about the design of financial regulations.

The analysis provides policymakers with a tool to combine the insights of subject matter experts with the advantages of computational analysis of text to score financial regulation laws to understand their impact on market performance. This paper thereby offers a new path, illustrating how triangulating different methods can facilitate the measurement of otherwise expensive and difficult to code institutional variables. This in turn furthers our understanding of important substantive public policy concerns.

The paper proceeds in the following steps. Part 2 sets the stage by presenting the illustrative example that structures the subsequent analysis: how to test hypotheses derived from the political economy of regulatory design and more specifically financial market regulation. We review the main hypotheses in the field and demonstrate the traditional observational approach to measurement by detailing the coding method used to construct the financial regulation database. We then discuss the limits of such observational approaches and how advances in computational data sciences can mitigate some of these shortcomings. Part 3 presents the computational methods used in the paper: a combination of NLP to parse all the documents of the financial regulation laws into feature vectors, topic modeling to cluster laws into relevant policy sub-domains, and finally supervised models to predict the outcome variable of interest. Part 4 presents our results, applying the ML algorithms to compare the power of alternative models in predicting regulatory discretion. Part 5 discusses the significance of the findings in light of the research design challenges discussed earlier. Conclusions and future developments close the paper.

2 Measurement and Inference in Testing Theories of Financial Market Regulation

2.1 The Why and How of Financial Regulation

What explains the structure of financial regulation? Where, how and by whom policy is made significantly impacts market outcomes.[1] When designing financial regulation laws, Congress specifies

[1] A number of studies show that government institutions matter for the regulation of markets. Keefer (2008) argues that competitive governmental structures are linked with competitive markets. In particular, separation of powers and competitive elections are correlated with strong investor protection and lending to the private sector. Barth, Caprio, and Levine (2006) show countries that encourage private enforcement of banking laws and regulation (e.g., through litigation) rather than direct control or no regulation at all, have the highest rates of financial sector development and therefore capital formation. Historical studies of financial development in the United States tell similar stories. Kroszner and Strahan (1999) show that the relative political strength of winners from deregulation (large banks and smaller, bank-dependent firms) and the losers (small banks and insurance firms) explains the timing of bank branching deregulation across states in the United States. Haber (2008) argues that governments free from outside political competition will do little to implement regulations in the banking sector.

the rules and procedures that govern bureaucratic actions. The key is how much discretionary decision making authority Congress delegates to regulatory agencies. In some cases, Congress delegates broad authority, such as mandating the Federal Reserve to ensure the "safety and soundness of the financial system." Other times, Congress delegates limited authority, such as specifying interest rate caps on bank deposits.

A recurring theme in the political economy literature of regulatory design is that the structure of policy making is endogenous to the political environment in which it operates.[2] Epstein and O'Halloran (1999) show that Congress delegates policymaking authority to regulatory agencies when the policy preferences of Congress and the executive are closely aligned, policy uncertainty is low, and the cost (political and otherwise) of Congress setting policy itself is high. Conflict arises because of a downstream moral hazard problem between the agency and the regulated firm, which creates uncertainty over policy outcomes.[3]

Application of these theoretical insights to financial regulation is well-motivated. Banking is a complex policy area where bureaucratic expertise is valuable and market innovation makes outcomes uncertain. Morgan (2002), for instance, shows that rating agencies disagree significantly more over banks and insurance companies than over other types of firms. Furthermore, continual innovation in the financial sector means that older regulations become less effective, or "decay," over time. If it did not delegate authority in this area, Congress would have to continually pass new legislation to deal with new forms of financial firms and products, which it has shown neither the ability nor inclination to do.[4] Overall, then, we have the following testable hypotheses: Congress delegates more discretion when: 1) The preferences of the President and Congress are more similar; and 2) Uncertainty over market outcomes (moral hazard) is higher.

Groll, OHalloran, and McAllister (2015) expand on this work, addressing whether policymakers regulate financial markets on their own or delegate regulatory authority to government agencies when faced with uncertainty about firm-specific investments and systemic risk at the financial services level. The executive is better informed and knows the exact correlation but puts greater weight on the social cost of a possible bailout.[5]

They conclude that Congress delegates regulatory authority when (1) the preferences of the executive and Congress are more similar, (2) the costs of a bailout are high, (3) there is more uncertainty about investment risk and systemic risk, and (4) Congress's bailout concern is low relative to the executive's. Further, financial services are more heavily regulated when firm-specific investments and systemic risks are uncertain. But when interbranch preferences differ or perceived

[2]For early work in this area, see, for example, McCubbins and Schwartz (1984) and McCubbins, Noll and Weingast (1987; 1989).

[3]Excellent technical work on the optimal type of discretion to offer agencies is provided by Melumad and Shibano (1991) and Alonso and Matouschek (2008), and Gailmard (2009). A series of studies examine the politics of delegation with an executive veto (Volden, 2002), civil service protections for bureaucrats (Gailmard and Patty, 2007; 2012), and executive review of proposed regulations (Wiseman, 2009), among others. See also Bendor and Meirowitz (2004) for contributions to the spatial model of delegation and Volden and Wiseman (2011) for an overview of the development of this literature.

[4]Maskin and Tirole (2004) and Alesina and Tabellini (2007) also emphasize the benefits of delegation to bureaucrats and other non-accountable officials.

[5]The argument is quite intuitive: When the financial system experiences a shock, then constituents are more likely to hold the president and the executive accountable than any individual member of Congress. For formal proofs of these propositions, the reader is referred to Groll, OHalloran and McAllister (2015). A similar argument is made in trade policy as constituents hold the president and the executive more accountable for the overall economic conditions, which explains more free-trade oriented positions by the executive than Congress. See, for example, OHalloran (1994).

systemic risk is low, Congress may allow risky investments to be made that, *ex post*, it wished it had regulated.

To illustrate the trade-off between policy differences and market uncertainty, the analysis focuses on Congresss and the executive's preferences and information differences. Congress prefers, in general, lower regulation thresholds than the executive because it puts less weight on the cost of a possible bailout. This is referred to as the preference difference between Congress and the executive, which is part of Congress's trade-off between the information advantage of the executive and the difference in preferred policy.

Figure 1 illustrates the implications when Congress would regulate on its own and when it would delegate to the executive. The shaded area indicates situations in which Congress delegates, while outside this area Congress makes policy on its own. Firms do not internalize a potential systemic failure in the absence of regulation and would make any investment that yields them a nonnegative expected return–that is, any investments at or above r^M in Figure 1. When Congress regulates, then any investments that would yield Congress a negative expected social return including the cost of a possible bailout–that is, any returns below r^C in Figure 1–would be banned and all others would be allowed. When Congress delegates, then, the executive sets a requirement in a similar spirit: banning investments with a negative expected social return that accounts for the cost of a possible bailout and the executive's salience. The executive knows the correlation and can therefore set a standard for uncorrelated investments, $\underline{r}^E$ and for correlated investments that of Congress's preferred threshold of $\overline{r}^C$ because of limited discretion from Congress. The delegation to the executive with discretion follows then from these two thresholds. All three regulatory standards are increasing in (1) the actual cost of a bailout, (2) the salience of a possible bailout, and (3) the perceived likelihood of correlated investments. But the stringency decreases when the likelihood of successful investments increases.

[Figure 1 about here.]

Focusing on the preference difference and expertise, the executive puts relatively more weight on the bailout cost than Congress, and therefore Congress values the executive's expertise most when there is no preference difference and delegates more discretion to the executive. But as the executive puts an increasing weight on the bailout cost, or Congress a decreasing weight, the preference difference increases and Congress delegates less to the executive because of its higher standards, which would imply a movement along the vertical axis in Figure 1. In other words, the delegation area shrinks as the disagreement between Congress and the executive increases and Congress prefers to regulate on its own. However, when the costs of a potential bailout increase or Congress's uncertainty over correlated investments increases, the area of delegation expands as Congress gains relatively more from the executive's expertise increase.

The preference difference also has implications for investments that are highly risky but high-returning when Congress perceives a low likelihood of correlated investments. In such situations, Congress prefers to regulate the financial investments on its own–that is, Congress would set a return requirement that is actually below the executive's standard for uncorrelated investments $\underline{r}^E$ in Figure 1. The reason is that the executive's expertise about correlated investments is not expected to be valuable to Congress, and the executive's standard is perceived as too stringent given the preference difference.

Overall, then, we have the following testable hypotheses:

1. Congress delegates more discretion when:
   (a) The policy preferences between Congress and the executive become more similar;
   (b) Firms investment risks become more uncertain, and;
   (c) There is more uncertainty about investment risk; and
   (d) The costs of a bailout are higher.

2. The more Congress cares about bailout costs, the higher are
   (a) Congress's preferred level of regulation,
   (b) The executive's discretion and regulation, and
   (c) Overall levels of regulation.

2.2 Financial Regulation Laws as Data

The previous section established that we need to test the hypothesis that regulatory design responds to the political preferences of Congress and the executive. Testing such hypotheses is challenging because it requires measuring discretionary authority. Measuring policy or institutional features is generally difficult, because these variables have no intrinsic scale and are not directly observable, arising instead from a combination of many rules. These rules are qualitative and thus require parsing texts. In our motivational example, the key variable of interest is the amount of discretionary authority Congress delegates to regulatory agencies to set policy (Discretion Index). It depends not only on the amount of authority delegated (Delegation Ratio) but also on the administrative procedures that constrain executive actions (Constraint Ratio). In what follows, we explain the process used to construct a measure of agency discretion. This process illustrates how measurement often proceeds in the social sciences, absent the help of computational tools.

We create a new database comprising all U.S. federal laws enacted from 1950 to 2010 that regulate the financial sector. The unit of analysis is an individual law, which specifies the rules and producers that regulate the actions of financial market participants. The database contains 120 public laws. The average corpus of text of a legislative summary is 6,278 words.[6] Because the Discretion Index is a combination of the Delegation Ratio and the Constraint Ratio, we present the measurement as a two-step process.

Delegation Ratio Delegation is defined as authority granted to an executive branch actor to move policy away from the status quo.[7] For each law, we code if substantive authority is granted to executive agencies, the agency receiving authority (for example, the U.S. Securities and Exchange Commission, Treasury, etc.), and the location of the agency within the administrative hierarchy (for example, independent agency, cabinet, etc.).

[6]The analysis relies on legislative summaries provided by *Congressional Quarterly* and contained in the Library of Congresss Thomas legislative database.

[7]For example, the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010 (Dodd-Frank Act) delegated authority to the the Federal Deposit Insurance Corporation to provide for an orderly liquidation process for large, failing financial institutions.

To measure delegation, each law in our database was manually read independently, its provisions numbered, and all provisions that delegated substantive authority to the executive branch were identified and counted.[8]

From these tallies, we calculate the delegation ratio by dividing the number of provisions that delegate to the executive by the total number of provisions. In the database, each law contains an average of 27 provisions of which 11 delegate substantive authority to four executive agencies. The average delegation ratio across all laws then is 0.41 or 11/27. The histogram of delegation ratios is shown in Figure 2a. As indicated, the distribution follows a more or less normal pattern, with a slight spike for those laws with 100% delegation (these usually had a relatively small number of provisions).

[Figure 2 about here.]

Constraint Ratio Executive discretion depends not only on the amount of authority delegated but also on the administrative procedures that constrain executive actions. Accordingly, we identify 14 distinct procedural constraints associated with the delegation of authority and note every time one appears in a law.[9] Including all 14 categories in our analysis would be unwieldy, so we investigated the feasibility of using principal components analysis to analyze the correlation matrix of constraint categories. As only one factor was significant, first dimension factor scores for each law were calculated, converted to the [0,1] interval, and termed the constraint index. Each law on average contained three constraints of the possible 14, yielding an overall constraint ratio of 0.21. The histogram of constraint ratios is shown in Figure 2b

Discretion Index From these data, we calculate an overall discretion index. For a given law, if the delegation ratio is D and the constraint index is C, both lying between 0 and 1, then total discretion is defined as $D * (1 - C)$ — that is, the amount of unconstrained authority delegated to executive actors.[10] The more discretion an agency has to set policy, the greater the leeway it has to regulate market participants. Lower levels of agency discretion are associated with less regulation.

As an illustration for how this measure is calculated, the Dodd-Frank Act contains 636 provisions of which 314 delegate authority to 46 executive agencies, yielding a delegation ratio of 0.5. The law also indicated ten procedural constraints out of a possible 14, yielding a constraint index of 0.7 (10/14). Combining delegation and constraints ratios produces a discretion index of $0.5 * (1 - 0.7) = 0.1$.

To verify the robustness of our estimates and confirm that our choice of aggregation methods for constraints does not unduly impact our discretion measure, Figure 3 shows the average discretion index each year calculated four different ways. As the time series patterns are almost identical, our choice of method number four (continuous factors, first dimension) is not crucial to the analysis that follows.

[8] To ensure the reliability of our measures, each law was coded independently by two separate annotators. It was reviewed by a third independent annotator, who noted inconsistencies. Upon final entry, each law was then checked a fourth time by the authors. O'Halloran et. al. (2015) provides a detailed description of the coding method used in the analysis.

[9] Examples of procedural constraints include spending limits, and legislative action required, etc. See O'Halloran et. al (2015) for a detail description of these constraints.

[10] See Epstein and O'Halloran (1999) for a complete discussion of this measure.

[Figure 3 about here.]

As a basic check on our coding of delegation and regulation, we compare the distribution of the discretion index for laws that regulated the financial industry overall, and laws that deregulated. We would expect that laws regulating the industry would delegate more discretionary authority, and Figure 4 shows that this is indeed the case. The average discretion index for the 24 laws that deregulate is 0.24, as opposed to 0.35 for the 85 laws that regulate, in line with the hypotheses discusses above.[11]

[Figure 4 about here.]

Do differences in executive discretion, in turn, affect the financial industry? It is in general difficult to determine a measure of the degree to which regulation is successful, but Philippon and Reshef (2012)'s measure of excess wages in the financial industry serves as a useful proxy; the greater the degree of regulation, the lower excess wages should be.

Accordingly, Figure 5a overlays total discretion delegated each year with Phillipon and Reshef's excess wage measure. The trends seem to be in the correct direction: excess wages began to increase in the 1960s, then a spate of regulation drove them down. They increased again the late 1980s, and again were lowered by a spike in regulation. Finally wages started rising precipitously during the 1990s and on to the first decade of the 2000s, but there was no corresponding rise in regulation to meet them, so they just kept increasing to the levels we see today. The trends in Figure 5a thus accord with the notion that delegating discretionary authority to executive branch officials does constrain the level of excess wages in the financial industry.[12] They also pose a puzzle: why was the spate of financial innovation in the 1960s — this decade saw the explosion of credit in the economy, including the widespread use of credit cards and the creation of credit unions — met with a regulatory response, while the most recent innovations — derivatives, non-bank lenders and the rise of the shadow banking system — were not? We postpone our suggested answer until the concluding section of this paper.

[Figure 5 about here.]

Figure 5a also indicates that the trend in recent decades has been for Congress to give executive branch actors less discretion in financial regulation. Since the Great Society era of the 1960's, and on into the early 1970s, the total amount of new executive branch authority to regulate the financial sector has generally declined. The exceptions are a few upticks in discretion which coincide with the aftermaths of well-publicized financial crises and scandals, including the Savings and Loan crisis, the Asian crisis, and the Enron scandal. Otherwise, the government has been given steadily less authority over time to regulate financial firms, even as innovations in that sector have made the need for regulation greater than ever, and even as the importance of the financial sector in the national economy has greatly increased as illustrated in Figure 5b.

[11] The remaining laws neither regulated nor deregulated.

[12] Further, the patterns also seem consistent with the notion that regulations "decay" over time as new financial instruments appear to replace the old one. If one estimates a Koyck distributed lag model $y_t = \alpha + \beta x_t + \beta\phi x_{t-1} + \beta\phi^2 x_{t-2} + \ldots + \epsilon_t$ via the usual instrumental variables technique, then $\beta = -0.025$ and $\phi = 0.49$, indicating that regulations lose roughly half their effectiveness each year. See Wooldridge (2006), pp. 635–637 for details of the estimation technique.

What is the source of this decrease in discretion? As shown in Figure 6a, the amount of authority delegated to oversee the financial sector has remained fairly constant over time, perhaps decreasing slightly in the past decade. The trends in Figure 5, then, are due mainly to a large and significant increase in the number of constraints placed on the regulators' use of this authority. In addition, we find that the number of actors receiving authority has risen significantly over the time period studied, as also shown in Figure 6a. However, the location of these agencies in the executive hierarchy has changed as well, away from more independent agencies to those more directly under the president's control as illustrated in Figure 6b.

[Figure 6 about here.]

Finally, we investigated the impact of this changed regulatory structure on overall market performance, statistically analyzing the impact of greater agency discretion on yearly changes in the Dow Jones Industrial Average and on the number and size of bank failures. We found, somewhat surprisingly, that financial markets react positively to higher levels of regulation. The Dow Jones average was half a percent higher, on average, in years in which Congress gave executive actors authority to regulate financial institutions, as compared to years in which no authority was given. Furthermore, the number and size of bank failures decreased significantly following regulatory reform.[13]

Overall, then, our preliminary analysis suggests that the current rules defining financial regulation may create a web of interlocking and conflicting mandates, making it difficult for regulators to innovate the rules and standards governing the financial industry, while at the same time opening regulatory agencies to industry capture. The problem is not lack of regulation, then, but that regulators have little discretion. Modern laws delegate less,constrain more and split authority across more agencies than their predecessors. This has created a situation where many areas of financial activity are heavily regulated by the Federal government, but those charged with oversight are hamstrung by overlapping jurisdictions, the need for other actors to sign off on their policies, or outright prohibitions on regulatory actions by Congress.

2.3 Limitations of the Observational Method

Section 2.2 defined the discretion index as $D * (1 - C)$, where D is the delegation ratio and C is the constraint index. As noted earlier, the process discussed in section 2.2 is a standard political economy approach to measurement for the testing of hypotheses. The above analysis adopts a research design based on observational methods, which potentially suffer from a number of well-known shortcomings. First, observational studies assume that all variables of interest can be measured according to a pre-specified coding rule. For example, the analysis posits that discretion can be calculated as a combination of delegation and constraints. In constructing these measures, the coding rules invariably impose a structure on the text, indicating some words or phrases as delegation and others as constraints. This can lead to coding bias.

Second, this approach can lead to substantial coding error, especially as it is scaled to larger datasets. Indeed, annotating the original data is extremely time consuming, especially when derived from disparate text-based sources, as we do here. The resources needed to extract the appropriate information, train annotators, and code the data can prove prohibitive and is prone to error. For example, consider again the Dodd-Frank Act, which covers the activities financial institutions can

[13] See Groll, O'Halloran and McAllister (2015)

Table 1: Comparison of Observational Study and New Machine Learning Method

	Observation Study		Machine Learning	
	Process	*Disadvantages*	*Process*	*Advantages*
Coding	Coding rules Human coders Checking consistency	Labor and time costs,and coding error in large corpus Coding bias Measurement uncertainty not quantified	Natural language processing Data can be words, semantic units, relations, data structure Validation on test set or cross-validation	Efficiency and consistency Detection of implicit/laten factors Quantification of measurement uncertaintiy
Analysis	Hypothesis testing Regression analysis Correlation on variables	Number of hypothesis Number of variables Scaling and sensitive to outliers	Various Naive Bayes Modesl Comparing model accuracies Comparison with human coding	Not limited by data amount Optimization of complexity Flexibility beyond coding rules
Internal Validity	Low Panoply of variables in single analysis	Missing underlying structure Missing important variables Imposed functional form	High Analysis of words, relations, semantic dependencies	No functional form Low generalization errors No overfitting

undertake, how these institutions will be regulated, and the regulatory architecture itself. Recall the law contains 636 major provisions, of which 314 delegate authority to some 46 federal agencies. In addition, the Act has a total of 341 constraints across 11 different categories, with 22 new agencies created. If we process the text of this law by the coding method detailed above, data annotators, trained in political economy theories, would read and code the provisions based on the rulebook provided. In effect, coders would have to read 30,000 words – the length of many novels. Unlike novels, however, legislation is written in complex legal language, which must be interpreted correctly and in painstaking detail. Consequently, there is the possibility that data annotators will introduce noise when coding laws. Measurement error is in some way inevitable, but the rule-based approach to coding prevents us from quantifying the uncertainty arising from these errors.

Third, standard econometric techniques, upon which many political economic studies rely, including the one conducted here, face difficulty in analyzing high dimensional variables that could theoretically be combined in a myriad of ways. For example, Figure 3 shows four possible alternatives to calculate the discretion index by varying the weights assigned to the different categories of procedural constraints.

In the following section, we explore data science methods and identify the techniques best suited to address the limitations of traditional observational methods. The computational methods presented below provide potential improvements over manual coding from a set of defined rules, including:

- We are not limited by the amount of data we can process.
- We are not limited to a handful of coding rules to quantify each law for building the discretion model.
- We can take into account the raw text of the law to explore word combinations, syntactic and dependency relations, and identify other sets of features that otherwise would be difficult to encode manually.

Table 1 provides a comparison of observational methods and data science techniques along three criteria: coding legislation, analysis, and internal validity. The overview illustrates the shortcomings of using only manual, rules-based coding methods and the way these new methods can enhance observational studies. In sum, computational analysis helps lessen error, quantify the uncertainty arising from these errors, find additional variables and patterns (data features), and improve the predictive power of models.

It is important to keep in mind that, while NLP and ML techniques can analyze extensive text-based data to test theories of policymaking and regulatory design, they rely on the critical data and hypotheses initially produced by subject matter experts to inform or seed the model and train complex algorithms. Therefore, data science techniques can be seen as a complement to observational studies and theoretical analysis.

3 Methods: Computational Coding of Financial Regulation Laws

We next describe the computational models used to predict the level of agency discretion. We show that both unsupervised and supervised algorithms can be combined to provide sparse representations of large datasets of laws and build predictive models of a statute's regulatory structure. In particular, we seek to determine what factors or "features" of a law predict agency discretion and also build a model that predicts discretion with high accuracy. Identifying the key features, words or word patterns, that predict the level of agency discretion in a given law helps refine and develop better proxies for institutional structure. We will compare how different types of features (computationally selected features, observational features and topics identified as latent variables in an unsupervised model) each contribute to increasing the accuracy of the predictive model. Our analysis thereby offers a novel approach to analyzing institutional design. Only recently have data science techniques been applied to study financial regulation and public policy more broadly.

First, however, we need to represent the passages of the legal documents in a format that is suitable for ML methods. We employ NLP techniques to convert the text of the laws into feature vectors. Some of the many different ways to encode text into features are listed here:

- Bag of Words: A bag of words model represents text as a feature vector, where each feature is a word count or weighted word count (McCallum and Nigam, 1998).
- Tag Sequences: Sentences or chunks of text are tagged with various information, such as Parts of Speech (POS) tags (Brill, 1992) or Named Entities (NEs, see Nadeau and Sekine, 2007), which can be used to further process the text.
- Graphs: Documents or paragraphs of the documents can be represented in graphs, where nodes can model sentences, entities, paragraphs, and connections represent relations between them (Mihalcea and Radev, 2011).
- Semantic Representation: A sequence of words mapped into an organized structure that encodes semantics of the word sequences (Griffiths, Steyvers and Tenenbaum, 2007).

These methods can be applied to represent text, thereby allowing machines to extract additional information from the words (surface forms) of the documents. Depending on the problem being addressed, one or more of these tools may be useful. We next explain the representation form adopted for the computational experiments below.

3.1 Data Representation Using Natural Language Processing

We must represent each individual law in a form suitable for ML algorithms to take as inputs. We first convert the raw text of an individual law in feature representation format. For the current

experiment, we convert the text of the financial regulation laws to Word Vectors. We describe the process of converting text into feature vectors below.

Step 1: Data Cleaning - For each law, we first clean the text to remove any words that do not represent core content, including meta information such as dates, public law (P.L.) number and other metadata that may have been added.

Step 2: Tokenization - After cleaning the data, we tokenize the text. Tokenization in NLP involves splitting a block of text into a set of tokens (words). This involves expanding abbreviations *(Mr. > Mister)*, expanding words *(I've > I have)*, splitting punctuation from adjoining words *(He said, > He said ,)* and splitting text using a delimiter such as white space *(bill was submitted > [(bill) (was) (submitted)])*.

Step 3: Normalization - Once tokenized, we must then normalize the data. The normalization of data involves having consistent tokenization across the same set of words.

Step 4: Vocabulary - In order to represent text in the form of feature vectors we must find the total vocabulary of the corpus appended with the additional vocabulary of the language.

Step 5: Vector Representation - Once we have defined the vocabulary, we can treat each word as adding one dimension in the feature vector that represents a block of text.

Let d_i be the document i. Let $y = w_1, w_2, ..., w_n$ be the vector representation of that document d_i, where w_k represent the existence of word w_k in the document d_i. Let us take an example piece of text from the Dodd-Frank Act, contained in section 1506.

d_i=*"..the definition of core deposits for the purpose of calculating the insurance premiums of banks"*. Let n be the total vocabulary size. The vector representation y for this document d_i will consist of a vector of length n where all values are set to zero except for the words that exist in document d_i. The total vocabulary size n tends to be significantly bigger than the number of unique words that exist in a given document so the vector tends to be very sparse. Hence, the vector y for document d_i is stored in sparse form such that only non-zero dimensions of the vector are actually stored. The vector of d_i will be

$$y = \{definition = 1.0, representation = 1.0, core = 1.0, purpose = 1.0, calculate = 1.0, \\ insurance = 1.0, premium = 1.0, bank = 1.0\}.$$

This is a binary vector representation of the text d_i. Given that the word is present in the document, we can in fact keep track of the word count in the given document d_i and store counts in the vector rather than storing the binary number representing it. Correspondingly, this generates a multinomial vector representation of the same text. If we take the entire Dodd-Frank Act as d_q, rather than sample text, and store counts for each word, we yield the vector representation of the Act as:

$$y = \{sec = 517.0, financial = 304.0, securities = 106.0, requires = 160.0, federal = 154.0, \\ requirements = 114.0, ..., inspection = 2.0\}.$$

Step 6: $TF * IDF$ Transformation - Once we represent the document in raw word vector format, we can improve the vector representation format by weighting each dimension of the vector with a corresponding term known as *Inverse Document Frequency* (IDF) (Spärck, 1972). An IDF transformation takes into account giving less weight to words that occur across all documents. For example, if the word *SEC* occurs frequently in all laws then the word *SEC* has less distinguishing power for a given class than *house*, which may occur less frequently, but is strongly tied to a given class. We re-weight all the dimensions of our vector d_q by multiplying them with the corresponding IDF score for the given word. We can obtain IDF scores for each word w_j by creating an IDF vector that can be computed by Equation 1.

$$IDF(w_j) = log(\frac{N}{\#count - of - Doc - with - w_j}), \tag{1}$$

where N is the total number of documents in the corpus and $\#count - of - Doc - with - w_j$ is the total number of documents with the word w_j. If the word w_j occurs in all documents then the IDF score is 0.

3.2 Unsupervised Model: Topical Modeling of Financial Regulation Laws

Unsupervised models can help discover underlying clusters in the data, such as groups of documents that address the same issue. In the context of our analysis, clustering is particularly useful because we expect that the discretion level will vary systematically by policy issue, depending on the level of political risk (see section 2.1). This prompts us to explore ways of clustering our corpus of financial regulation laws according to finer policy domains in order to include these policy domains as features in the supervised model described in the next section. More generally, the clustering of laws by policy domains should prove useful in new and larger corpora of laws as a way to describe substantive issues covered by the laws. Clustering can also facilitate the coding of institutional features, as the relevance of different institutional features can vary depending on the policy domain (e.g. health versus financial regulation).

Different models exist to describe corpora as a set of clusters or topics, in which topics are distributions over word frequencies or other scores such as TF-IDF scores. An example includes k-means (Steinbach, Karypis and Kumar, 2000). Topic models is a family of models particularly well suited for this task. These models start from the same word vector representation of the document described earlier. They then model this word vector as a distribution over topics (for example, a financial law can be about both securities and commodities) and topics as distributions over words. Conversely, one can think of each word of a document as belonging to a topic and each document being a mixture of multiple topics, a realistic description for many documents, including laws. One of the first and most widely used topic models is the Latent Dirichlet Allocation model (LDA), proposed by Blei, Ng and Jackson (2003), of which other topic models are close variants. We thus use it to present the basic structure of topic models.

In a topic model with K topics, each topic k has a distribution β_k over words of the vocabulary. β_k is thus a vector of probabilities of observing each word, summing to 1. Each document d_i has a

distribution over topics θ_i (a vector with elements $\theta_{i,k}$ for $k = 1, ...K$). $w_{i,j}$ is the observed word in the jth word position of document d_i, with topic assignment $z_{i,j}$ (the vector of topic assignment for the whole document being denoted z). The generative model for a document d_i with word vector y_i can be described by the following steps:

1. Draw the per topic proportions from a Dirichlet prior

$$\theta \sim Dir(\alpha)$$

2. For each of the n words in document d_i (where j is the jth word in the document d_i):

   (a) Draw the topic assignment

   $$z_{i,j} \sim Multinomial(\theta_i)$$

   (b) Draw the word given the topic assignment:

   $$w_{i,j} \sim Multinomial(\beta_{z_{i,j}})$$

The resulting log-likelihood for a document represented with word vector y_i is:

$$log\, p(y_i|\alpha, \delta) = log \int \left(\sum_{k=1}^{K} \prod_{j=1}^{n} p(w_j|z_{i,j} = k, \beta_k) p(z_{i,j} = k|\theta_i) \right) p(\theta_i|\alpha) d\theta_i$$

Different implementations exist to obtain the posterior distribution of the parameters of this model based either on sampling algorithms (most commonly Gibbs sampling as presented in Griffiths and Steyvers, 2004) or variational expectation-maximization algorithms.

Many other models have been proposed on the basis of LDA. These models relax some of the assumptions in LDA and build into the basic model other features and dependencies to uncover more complex structures in documents (for an excellent recent review, see Blei, 2012). Among others, dynamic topic models allow the distribution of topics and the content of topics to change over time or according to the documents authors. A large suite of models integrate other types of meta-data. Bayesian non-parametric topic models uncover hierarchies of topics, from general to more fine-grained sub-topics. Correlated topic models allow correlations between topics (some topics are more likely to co-occur in a document than others).

In this project, we use the Latent Dirichlet Allocation model fitted by a variational expectation-maximization (VEM) algorithm as implemented in the `topicmodels` package in R (Grun and Hornik (2011) an R interface to Blei's C implementation of VEM for fitting LDA). We also experimented with the Structural Topic Model (STM) of Roberts, Stewart and Tingley (2014), also fit with a VEM algorithm and implemented in the R package `stm`. STM is a topic-model developed with comparative political economy applications in mind. It incorporates document covariates as variables affecting the prevalence of topics in each document (for example, certain topics may be more prevalent in a Republican presidential speech than in a Democratic one) and the word content for each document (a Republican candidate may frame the same topic in different ways than a Democratic candidate). The generative model of the STM differs somewhat from LDA, since the per-document topic distribution is drawn from a logistic normal distribution (allowing covariates to influence the mean). The per-topic word distribution is drawn from an exponential distribution (also allowing the inclusion of covariates). The STM package is versatile, with many features to estimate models, select, explore and visualize them. We will present the results from the STM in 4.5 and show that it successfully identifies relevant policy sub-domains, which additionally help improve the performance of our supervised models, to which we now turn.

3.3 Supervised Model: Naive Bayes

We frame our problem of predicting the level of agency discretion in a given law as a classification problem. The ML approach to supervised classification tasks is to train a model based on features of the data to predict observations' membership into the classes of interest labeled by a human. For this dataset, we denote the set of discretion classes as C_i, where i ranges from 0 to 5, the total number of classes used to tag individual laws for the *Level of Discretion.*

The Level of Discretion C_i in a given law is a subjective measure of how much discretionary authority is given to the agency in that law only. It is coded from 0 to 5, with 0 indicating that no discretionary authority was given to executive agencies to regulate financial markets and 5 meaning that significant discretionary authority was given. The *Discretion Level*, as a subjective measure, is different from the Discretion Index computed in section 2.2. The latter index is derived from theory (based on the delegation ratio and constraint index, which are variables deemed salient on the basis of theoretical models of political economy). We use the *Discretion Level* instead of the *Discretion Index* because we want a measure that is independent from the coding rules dictated by theory, which could be wrong, and instead reflect a human's intuitive understanding of the text as a whole. Additionally, using subjective judgment as the *gold standard* that algorithms have to predict is a standard practice when ML models are built. With this in mind, let C_i be the level of discretion that we are trying to predict for a given document (law) y.

Many different machine learning algorithms are used in document/text classification problems. One of the most commonly applied algorithms is a Naive Bayes method. We build a Naive Bayes Model for predicting discretion level for each of the laws y.

We must compute $p(C_i|y)$ for each of the classes (discretion levels) and find the class C_i. $p(C_i|y)$ can be obtained by Equation 2

$$p(C_i|y) = \frac{p(C_i)p(y|C_i)}{p(y)} \tag{2}$$

To find the best class C_i, we compute the argmax on the class variable:

$$i^* = arg\max_i p(C_i|y). \tag{3}$$

To compute $p(C_i|y)$, we use Bayes rule to obtain $p(C_i|y) = \frac{p(y|C_i)*p(C_i)}{p(y)}$. Since our task is to find argmax on C_i, we simply need to find C_i with the highest probability. As the term $p(y)$ is constant across all different classes, it is typically ignored. Next, we describe how we can compute $p(y|C_i)$ and $p(C_i)$.

$p(C_i)$ is the prior probability of class C_i. This term is computed on the training set by counting the number of occurrences of each class. In other words, if N is the total number of documents in training and N_i is the number of documents from class i, then $P(C_i) = \frac{N_i}{N}$.

In order to compute the probability $p(y|C_i)$, we assume that document y is comprised of the following words $y = \{w_1, w_2, \ldots, w_n\}$, where n is the vocabulary size. We make a conditional independence assumption that allows us to express $p(y|C_i) = p(w_1, \ldots, w_n|C_i)$ as

$$p(w_1, \ldots w_n|C_i) = \prod_{j=1}^{n} P(w_j|C_i). \tag{4}$$

We compute $P(w_j|C_i)$ by counting the number of times word w_j appears in all of the documents in the training corpus from class C_i. Generally, *Add-one Smoothing* is used to address the words that never occur in the training document. Add-one smoothing is defined as follows: Let N_{ij} be the number of times word w_j is found in class C_i and let $P(w_j|C_i)$ be defined by equation 5, where n is the size of the vocabulary.

$$P(w_j|C_i) = \frac{N_{ij} + 1}{\sum_i N_{ij} + n} \tag{5}$$

Given a test document y, for each word w_j in y, we look up the probability $P(w_j|C_i)$ in this test document and substitute it into equation 5 to compute the probability of y being predicted as C_i. In section 4, we describe the Naive Bayes Model built from different sets of features, thereby allowing us to compare the performance of our model in various settings.

4 Results: Comparing Models

As explained in section 3.3, our purpose is to find characteristics of financial regulation laws that predict agency discretion. The computational analysis approach lets the data identify those policy features or attributes that most accurately predict outcomes rather than test hypothesizes about the impact of theoretically motivated independent explanatory variables. In this section, we compare different versions of the Naive Bayes model, incrementally enriching and refining the feature set used to build the model. As noted in section 3.3, to evaluate the performance of a given ML model in predicting agency discretion, each law is assigned a discretion level, ranging from 0 to 5, which serves as the target answer. We then compare the predictions yielded by each of the ML models against the baseline or target value. Finally, we compute a summary metric. Here we use the F-statistic, which indicates the accuracy of alternative models in correctly classifying each law relative to the baseline.

4.1 Naive Bayes Model 1: Computer Generated Features

The first Naive Bayes Model is based on the document vectors where the data is all the text found in the financial regulatory laws, which includes more than 12,000 distinct words. Each word is a parameter that must be estimated across each of the six classes. We took the raw text of the laws and converted it into document vectors as described in the previous section and estimated the parameters of Naive Bayes Model. This model produced an accuracy of 37% with an F-Measure of 0.38.

Our baseline system is a model that predicts Class 0 for all documents. Absent any other information, the best prediction for a document is a class that has the highest prior probability, which is 0.26 for Class 0. We should note that the Naive Bayes Model 1 based solely on text features does better than the baseline model by 11%.

Table 2 shows the prior probabilities for the six classes of Discretion.

4.2 Naive Bayes Model 2: Manually Coded Features

We first compare the model with features extracted from the raw text derived from the coding rules outlined above. We take the same set of laws and their corresponding coding rules as features. We identified more than 40 features from the coding rules, including the Number of Provisions with

Table 2: Class and Prior Probability

Class	Prior Probability
0	0.26
1	0.14
2	0.25
3	0.24
4	0.08
5	0.07

Delegation, constraints such as Reporting Requirements, Time Limits, et cetera. We next created a second Naive Bayes Model using these hand-labeled coding rules as features. Naive Bayes is a general classification algorithm that can take any type of feature vectors as inputs. For Model 2, we again estimated the parameters using the same set of laws that was used to estimate the parameters for building Model 1, and produced an accuracy of 30.0% and F-Measure of 0.40. Interestingly, the raw text model produced a higher level of accuracy than the model built solely from the coding rules. When we build a Naive Bayes Models with manually hand coded features the model parameters are estimated in a similar fashion as stated in Equation 4 except instead of words w_j we have hand coded features h_k as described in Equation 6.

$$p(h_1, \ldots, h_m|C_i) = \prod_{k=1}^{m} P(h_k|C_i). \tag{6}$$

4.3 Naive Bayes Model 3: Combining Manual and Computational Features

Naive Bayes Model 3 combines the purely raw text approach of examining all of the text and the manual approach in which we use the coded features extracted by annotators from the texts. We again estimated the parameters as described in Section 4. This model produces an accuracy of 41% and an F-measure of 0.42. These results indicate that a combination of both raw text and manual approaches performs better than either individual approach. When we combine the features we are pooling both sets of w_j and h_k features into same pool. For the estimation of $p(w_i, \ldots, w_n, h_k, \ldots, h_m|C_i)$ we again assume conditional independence among features given the class allowing us to efficiently compute $p(w_i, \ldots, w_n, h_k, \ldots, h_m|C_i)$ using the following equation $\prod_{j=1}^{n} P(w_j|C_i).\prod_{k=1}^{m} P(h_k|C_i)$.

4.4 Naive Bayes Model 4: Feature Selection Model

The number of parameters for Model 1 is almost the same size as the vocabulary of the corpus, while the total number of parameters for Model 2 equals the number of manually-labeled coding rules. It is likely that the raw text-based features can be overwhelming for a small number of manually-labeled features. Therefore, we built a fourth Naive Bayes Model where we ran a feature selection algorithm on the combined set of features.

Feature selection algorithms select a subset of features based on different constraints or on the maximization of a given function. We used a correlation-based feature selection algorithm, which selects features that are highly correlated within a given class, but with low correlation across classes, as described in Hall (1998). The feature selection algorithm picked up a feature

set containing 47 features, including some features from the manually-produced coding rules and a few word-based features as well. Some of the words selected by the feature selection algorithm of Discretion Level include: *auditor, deficit, depository, executives, federal, prohibited, provisions, regulatory, and restrict.*

Model 4 produced the highest level of accuracy at 67% with an F-measure of 0.68. This increase in accuracy is explained in part by the smaller feature set that remains after we discard a number of word-based features. The smaller feature set allows us to better estimate the parameters with our data set of 120 laws thereby reducing the data sparsity problem. The best model produced a high degree of accuracy only after careful feature selection and model design.

4.5 Naive Bayes Model 5: Feature Selection Model with Topics

In the previous model, we combined a selection of word features and three manually-coded features. In this last model, we further enrich the analysis by including topics identified by topical modeling as additional features. Topical modeling allows us to validate the model beyond the accuracy measure. Indeed, we argue that different policy sub-domains should have different Discretion Levels, depending on the risk involved for politicians. Thus, we will test whether the topics indicative of more risk indeed predict higher discretion.

Before showing how the topics affect the performance of the Naive Bayes model, we present the results of the topic modeling itself, which shed light on the content of the corpus, and how these results are to be evaluated.

To fit a topic model, one must stipulate the number K of topics (unless we use a Bayesian non-parametric topic model where K is also inferred from data). An analysis typically explores corpora by fitting the topic model with different numbers of topics and examining their relevance for the analysis of interest. Figure 7 presents words representative of the topics for a model where $K = 5$. It shows both the seven words with the highest probability of appearing in a text of a given topic, as well as the seven words that are most representative of each topic according to their high FREX score, a measure that is similar in spirit to $TF - IDF$, as it combines the exclusivity of a word to a topic and its prevalence.[14]

[Figure 7 about here.]

Through examination of the words, we can interpret the underlying topics inferred by the model. Topic 1 concerns traded securities, including futures and swap agreements. We label it *securities_futures.* Topic 2 concerns the regulation of banks, specifically those laws that seek to insure the safety and soundness of the financial system. We label it *banking_safety.* Topic 3 concerns mortgages and the funding of housing and urban development. We label it *housing.*

[14] We define the exclusivity score of a word j for a topic k as the ratio of its probability of occurring in topic k to its probability of occurring in other topics. Thus $\phi_{k,j} = \frac{\beta_{k,j}}{\sum_{i \neq k} \beta_{i,j}}$. We then define the $FREX_{k,j}$ score as the harmonic mean of the words rank in the distribution of exclusivity scores for topic k (which frequency distribution is denoted $\phi_{k,\cdot}$) and the word's rank in the distribution of word frequencies for topic k (which frequency distribution is denoted $\mu_{k,\cdot}$). Thus:

$$\text{FREX}_{k,j} = \left(\frac{\omega}{\text{ECDF}_{\phi_{k,\cdot}}(\phi_{k,j})} + \frac{(1-\omega)}{\text{ECDF}_{\mu_{k,\cdot}}(\mu_{k,j})} \right)^{-1} \tag{7}$$

where ω is the weight for the exclusivity (which is set to 0.5 by default) and $\text{ECDF}_{x_{k,\cdot}}$ is the empirical cumulative density function applied to the values x over the first index, giving us the rank. See Airoldi, Blei, Erosheva, and Fienberg, 2015, p. 280.

Human Annotated Topic Labels	*Topics Identified by the Topic Model* positive association	negative association
banking	banking_safety bank fraud_foreign banks welfare_banking	securities_futures
regulation	banking_safety *interacted with* bank fraud_foreign banks	securities_futures
consumer protection	housing banking_security	securities-futures bank fraud_foreign banks
commodities	welfare_banking	banking_safety housing bank fraud_foreign banks
securities	securities_futures	banking_safety housing
mortgage_lending	housing welfare_banking	banking_safety securities_futures bank fraud_foreign banks

Table 3: Table showing the association between annotated labels (human coded topics on the left) and the latent topics identified by the topic model.

Topic 4 concerns banking regulation, specifically bank fraud and the relationship of U.S. banks and banking policies to foreign banks and policies. We label it *bank fraud_foreign banks*. Topic 5 concerns the relationship of financial institutions to welfare programs, and market protection. We label it *banking_welfare*. Using our domain expert judgment, these topics appear coherent and meaningful.

With this dataset, we can further evaluate the quality of the topics because we have human-coded information on topics. For each law, annotators determined whether it addresses each of the following six policy issues (and a law can address more than one): banking, securities, commodities, regulation, consumer protection and mortgage lending. To validate the topics, we regressed each of these annotated labels on $\theta_{1:5}$, that is, the posterior of the topic proportions (for the five latent topics of the topic model). If the latent topics are coherent, these regressions should show thematically logical association between the annotated labels and the latent topics. Table 3 shows that this is indeed the case. For each annotated topic label in the original dataset, the table shows which latent topics identified by the topic model are significantly associated with it (either positively or negatively). For example, we see that laws with annotated label *banking* are strongly associated with latent topics *banking-security*, *bank fraud - foreign banks*, *welfare-banking*, but negatively associated with *security*. Conversely, the annotated label *securities* is positively associated with the topic *securities-futures*, but negatively associated with *banking-safety* and *housing*. We see from these results that the topics identified by the topic model are coherent and reliable indicators of existing policy sub-domains.

How do we know that $K = 5$ is an appropriate number of topics? There is no "best" number of topics, since the appropriate resolution depends on the interpretation and insights sought by

the analyst. However, K could be too small (lumping topics that are quite different) or too large (splintering the data into groups that are difficult to interpret because not truly distinguishable). When the number of topics is small, these problems appear by inspection (as we have done above), but when there are many topics, quantitative indicators of the quality of the model are useful. We, therefore, quickly present a few below.

The appropriate methods of validating a topic model is an active topic of research (Blei 2012). One approach is to compute the probability of held-out data. Figure 8a shows the perplexity of held-out data in 10-fold cross-validations for models with different number of topics K. Perplexity is $p = -exp(\frac{\sum_{i \in heldout\ set} p(y_i|\beta_{1:K},\alpha)}{total\ \#\ words\ in\ \ heldout\ set})$. A lower perplexity indicates a better fit. For our data, the perplexity declines rapidly as a function of K while $K \leq 4$, after which increases in K have a diminishing impact on this measure of fit.

Recent research demonstrates that probability-based measures of fit, such as perplexity, correlate poorly with human judgment of the coherence and interpretability of topics (Chang et al. 2009). In complement to probability-based measures, researchers have proposed measures that capture the goals of the model, namely finding coherent and exclusive topics. Mimno et al. (2011) introduced a measure of *semantic coherence* of topics. Semantic coherence computes how frequently the most probable words of topic k co-occur in the same document (relative to the baseline frequency of these words). This is an intuitive measure of how clustered the words representing topic k are in the corpus. However, semantic coherence will decrease with the number of topics (since the most frequent words in each topic become more likely to appear in multiple topics). It must be traded against *exclusivity*, a measure of how exclusive the words of a topic are to that particular topic[15]. Figure 8b plots the exclusivity and the semantic coherence of each topic as we vary K from 4 to 7. Intuitively, models whose topics are on the frontier of exclusivity versus semantic coherence are of better quality (Roberts et al. 2014). In our case, we see that $K = 4$ and $K = 6$ yield each one topic that scores low either on semantic coherence or on exclusivity. Topics obtained with $K = 5$ and $K = 7$ are closer to the frontier. In fact, $K = 7$ seems slightly superior. Examining the resulting topics, we observed that topics 1 and 5 are split in finer categories that are easy to interpret in the context of financial regulation. However, for reasons of parsimony given the size of the corpus, we keep $K = 5$ in the experiments of section 4.

[Figure 8 about here.]

We now turn to the final predictive model, in which the topics of the topic model are added to the previously selected features. We wish to know whether the policy sub-domains are useful in predicting the discretion level. To answer this question, we associate each law with its dominant topic (the topic k for which $\theta_{i,k}$ is highest), as derived above, and add this topic as an attribute in the Naive Bayes Feature Selection Model 4, for a total of 48 features.

Adding topics as additional features increased the classification accuracy of the model to 70.83%. This model produces an F-measure of 0.71. Furthermore, a chi-square test rejects the null hypothesis of independence between the discretion classes and the topics. These results reinforce our notion that combining different types of features generated through alternative methods strengthens the quality of the model. Here we add the latent variables inferred by a topic model, corresponding to different policy domains within financial regulation legislation.

[15]Exclusivity of word j to topic k was defined in an earlier footnote, and is $\phi_{k,j} = \beta_{k,j} / \sum_{i \neq k} \beta_{i,j}$. The exclusivity score for the whole topic is the sum of these $\phi_{k,j}$ word scores for all words in a topic.

As mentioned earlier, we can go further and apply these topics as a test that the model is using these features as predicted by theory. Coming back to the political economy hypotheses outlined in the first part of the paper, we expect that politicians will grant less discretion in policy areas that entail more political risk. Topics 1 and 2, which concern securities, consumer protection and the stability of the financial system are high risk (coded H). Topics 3 and 5, concerning the financing of homes and welfare programs is medium risk (coded M), while topic 4, the regulation of bank fraud and foreign banks is low political risk (coded L). If we use this risk level as an attribute instead of the topic (keeping our number of attributes at 48), we recover the identical classification accuracy of 70.83. This indicates that the risk levels are meaningful predictors.

We run an ordinal logistic regression of the discretion level on the risk level of the laws policy domain to analyze the influence of the risk level of the policy domain on the discretion level. For each discretion level C_i, Figure 9 shows the first difference between being a law in a High Risk policy domain versus a Low Risk policy domain. This quantity indicates the changes in the expected probability of the law having level C_i when the we shift the variable "political risk" from low to high. For example, for discretion level 0 ($C_i = 0$ the lowest discretion level) the figure represents the difference in the expected probability that a law has discretion level 0 given that it addresses a High Risk policy domain versus a Low Risk policy domain: $E(C_i = 0|H) - E(C_i = 0|L)$. When Congress and the executive disagree over policy saliency, we see from Figure 9 that laws concerning High Risk policy domains are less likely to have high levels of discretion, in line with our hypothesis. This demonstrates that the features affect the prediction of discretion in a way that is consistent with theory.

[Figure 9 about here.]

5 Discussion and Conclusions

The results of the five models are summarized in Table 4. Model 5 performs best, yielding a 71% accuracy level. It includes three types of features: computationally selected features, manually-coded features, and topics, all which complement each other by capturing different attributes embedded in the financial regulation laws.

Table 4: Naive Bayes Models

Feature Type	Accuracy(%)	F-Measure
Model 1: Computer Generated Text Features (C)	36.66	0.38
Model 2: Manual-Coded Variables/Features (M)	30.00	0.40
Model 3: C + M	40.83	0.42
Model 4: Feature Selection (C + M)	66.66	0.68
Model 5: Feature Selection with Topics (C + M + T)	70.83	0.71

Let us consider how the computational techniques described here would concretely modify the research design of an observational study in political economy. First, we see that the Naive Bayes Model allows us to quantify the measurement uncertainty inherent in mapping complex texts onto an ordinal or categorical variable. This uncertainty can in turn be integrated in the statistical analysis of the data to better quantify our uncertainty. Hence, these methods can be useful even with a small hand-annotated corpus, simply to put errors bars on the coding of the data.

Second, by quantifying how the measurement uncertainty resulting from the coding changes as we modify the type of features used, the methods allow us to decide how to spend scarce resources to develop our datasets. In the example used in this paper, the hand-annotated *Discretion Index* (in Section 2.2) is very time consuming and prone to errors as each provision of each text needs to be carefully annotated for whether it delegates authority and the number of procedural constraints. This approach is not scalable. In contrast, Models 1, 2 and 3 are perfectly scalable, as they require no human coding. Yet, they lead to high levels of uncertainty. Models 4 and 5 include manually-coded features which are much less time consuming and less prone to error than the Discretion Index and they help reduce the uncertainty. We can include these different uncertainty levels in power calculations and decide the degree to which it is best to increase the quality of the coding by including more or less costly human-annotated features or increase the quantity of data by eschewing the more costly human-annotated features and relying more heavily on computationally generated features.

This paper is the first to our knowledge to apply this combination of supervised and unsupervised models to legal texts for the purpose of testing theories in political economy. We have focused on the simplest text data representation, the bag-of-words approach. We have found that this data representation and topic modeling successfully identify coherent policy sub-domains. We have also found that these word features help improve Naive Bayes classifiers in predicting institutional variable, such as regulatory discretion. While the level of accuracy attained is sufficient for some analyses,[16] it is only 71% in our best model. This indicates that a bag-of-words approach has limits.

Subsequent research to enhance the methods used in this paper include feature-level improvements, such as N-grams with high term-hood (e.g. Wong et al. 2008); transforming the word features and the texts into synsets, with Wordnet; and identifying features based on Neural Nets and Deep Nets (e.g., Mikolov et al. 2013). A second class of enhancements include improving the algorithm by adopting random forests, support-vector machines or Maximum Entropy Classifiers that use features combining words, semantic labels and part-of-speech-tags (Nigam et al., 1999).

As we know, computational data science captures complex patterns and interactions that are not easily recognized by manual coding rules. However, these NLP and ML techniques have not been used to analyze text-based data to test theories of regulatory design. These computational methods enable us to represent the text of a given law as a feature vector, where each feature represents a word or weighted terms for words. We also apply topic modeling to associate topics with institutional features, such as regulatory discretion. Each of these techniques provide potential improvements over manual-coding and inferences from a set of defined coding rules. Yet these computational models rely on the critical data initially produced by subject matter experts to inform or seed the model and train these complex algorithms. Computational data science techniques, therefore, are an important and critical complement to observational studies.

A research strategy that uses more than one technique of data collection and interpretation can improve the validity of analyzing high-dimensional datasets commonly found in political economy studies of financial regulation. The practical implications of the analysis are manifold. The analytical methods developed enable governments and financial market participants alike to: 1) automatically score policy choices and link them to various indicators of financial sector performance; 2) simulate the impact of various policies or combinations of policy under varying economic and political conditions; and 3) detect the rate of change of market innovation by comparing trends of policy efficacy over time. The analysis will help governments to better evaluate the effect of the

[16] For a test of these hypotheses using standard regression analysis, see Groll, O'Halloran and McAllister (2015).

policy choices they confront, as well as assist the financial community to better understand the impact of those choices on the competitive environments they face.

References

[1] Airoldi, Edoardo M., David M. Blei, Elena A. Erosheva, and Stephen E. Fienberg (2015). Handbook of Mixed Membership Models and Their Applications (Chapman and Hall/CRC Handbooks of Modern Statistical Methods) (Page A). CRC Press. Kindle Edition.

[2] Alesina, Alberto and Guido Tabellini. 2007. "Bureaucrats or Politicians? Part I: A Single Policy Task." *The American Economic Review* 97, no. 1: 169–179.

[3] Alonso, Ricardo and Niko Matouschek. 2008. "Optimal Delegation." *The Review of Economic Studies* 75, no. 1: 259–93.

[4] Barth, James R., Gerard Caprio, Jr., and Ross Levine. 2006. *Rethinking Banking Regulation: Till Angels Govern*. New York: Cambridge University Press.

[5] Bendor, Jonathan and Adam Meirowitz. 2004. "Spatial Models of Delegation." *American Political Science Review* 98(2):293-310.

[6] Blei, David M. 2012. "Probabilstic topic models", *Communications of the ACM* 55, no.4:77-84.

[7] Blei, David M and Andrew Y. Ng, and Michael Jordan. 2003. "Latent Dirichlet Allocation", *Journal of Machine Learning Research* 3:993-1022.

[8] Brill, E., 1992. A simple rule-based part of speech tagger. InProceedings of the workshop on Speech and Natural Language (pp. 112-116). Association for Computational Linguistics.

[9] Chang, Johnathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems* (22):288296.

[10] Epstein, David and Sharyn O'Halloran. 1999. *Delegating Powers* New York: Cambridge University Press.

[11] Epstein, David and Sharyn O'Halloran. 2009. "Avoiding Financial Katrinas: Systemic Risk as a Common Pool Problem." Working Paper, Columbia University.

[12] Gailmard, Sean. 2009. "Discretion Rather Than Rules: Choice of Instruments to Constrain Bureaucratic Policy-Making." *Political Analysis* 17(1): 25–44.

[13] Gailmard, Sean. 2009. "Multiple Principals and Oversight of Bureaucratic Policy-Making." *Journal of Theoretical Politics* 21(2): 161–86.

[14] Gailmard, Sean and John W. Patty. 2007. "Slackers and Zealots: Civil Service, Policy Discretion, and Bureaucratic Expertise." *American Journal of Political Science* 51(4): 873–89.

[15] Gailmard, Sean and John Patty. 2012. "Formal Models of Bureaucracy." *Annual Review of Political Science* 15: 353–77.

[16] Griffiths, Thomas L., and Mark Steyvers. 2004. "Finding scientific topics." *Proceedings of the National Academy of Sciences* 101.suppl 1: 5228-5235.

[17] Griffiths, Thomas L., Mark Steyvers, and Joshua B. Tenenbaum. 2007. "Topics in Semantic Representation." *Psychological Review* 114(2):211–244.

[18] Groll, Thomas, Sharyn OHalloran, and Geraldine McAllister. Delegation and the Regulation of Financial Markets mimeo (2015).

[19] Hornik, Kurt, and Bettina Grn. 2011 "topicmodels: An R package for fitting topic models." *Journal of Statistical Software* 40(13): 1-30.

[20] Haber, Stephen. 2008. "Political Institutions and Financial Development: Evidence from the Political Economy of Bank Regulation in Mexico and the United States", In *Political Institutions and Financial Development*, eds Haber, North, and Weingast. Stanford: Stanford University Press.

[21] Hall, M.A. 1998. "Correlation-based Feature Subset Selection for Machine Learning." *Phd Thesis*. University of Waikato.

[22] Keefer, Philip. 2008. "Beyond Legal Origin and Checks and Balances: Political Credibility, Citizen Information, and Financial Sector Development," In *Political Institutions and Financial Development*, eds. Haber, North, and Weingast. Stanford: Stanford University Press.

[23] Kroszner, Randall and Philip Strahan. 1999. "What Drives Deregulation? Economics and Politics of the Relaxation of Bank Branching Restrictions." *Quarterly Journal of Economics* 114 (4): 1437–67.

[24] Maskin, Eric and Jean Tirole. 2004. "The Politician and the Judge: Accountability in Government." *American Economic Review* 94 (4): 1034–54.

[25] McCallum, A. and Nigam, K., 1998, July. A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).

[26] McCubbins, Mathew D. and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28, no. 1: 165–79.

[27] McCubbins, Mathew D., Roger Noll and Barry Weingast. 1987. "Administrative Procedures as Instruments of Political Control." *Journal of Law, Economics and Organization* 3: 243–77.

[28] McCubbins, Mathew D., Roger Noll and Barry Weingast. 1989. "Structure and Process, Politics and Policy: Administrative Arrangements and the Political Control of Agencies." *Virginia Law Review* 75: 431–82

[29] Melumad, Nahum D., and Toshiyuki Shibano. 1991. "Communication in settings with no transfers." *RAND Journal of Economics* 22(2): 173–98.

[30] Mihalcea, R. and Radev, D., 2011. Graph-based natural language processing and information retrieval. Cambridge University Pres

[31] Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders and Andrew McCallum. 2011. "Optimizing semantic coherence in topic models." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

[32] Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff, 2013. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems.

[33] Morgan, D.P.. 2002. "Rating Banks: Risk and Uncertainty in an Opaque Industry." *The American Economic Review* 92(4): 874-888.

[34] Nadeau, D. and Sekine, S., 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), pp.3-26.

[35] Nigam, Kamal and Lafferty, John and McCallum, Andrew, 1999. Using maximum entropy for text classification. IJCAI-99 Workshop on Machine Learning for Information Filtering.

[36] O'Halloran, Sharyn. 1994. Politics, Process, and American Trade Policy. Ann Arbor: University of Michigan Press.

[37] O'Halloran, Sharyn, Geraldine McAllister and Kaiping Chen. 2014. Working Paper, Columbia University.

[38] O'Halloran, Sharyn, et al. 2015. "Data Science and Political Economy: Application to Financial Regulatory Structure." Forthcoming. In Howard Rosenthal , ed. *Big Data and Political Economy.* New York: Russell Sage Press.

[39] Philippon, Thomas and Ariell Reshef. 2012. "Wages and Human Capital in the U.S. Financial Industry: 1909-2006." *Quarterly Journal of Economics* 112(4): 1551-1609.

[40] Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2014 "stm: R package for structural topic models." R package version 0.6 1 (forthcoming in *Journal of Statistical Software*

[41] Roberts, Margaret E, Brandon M. Steward, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand2014. "Structural Topic Models for OpenEnded Survey Responses." *American Journal of Political Science* 58(4): 1064-1082.

[42] Steinbach, Michael, George Karypis, and Vipin Kumar. 2000. "A Comparison of Document Clustering Techniques." Proceedings of the 6th ACM SIGKDD, World Text Mining Conference, Boston, MA.

[43] Spärck Jones, K. 1972. "A statistical interpretation of term specificity and its application in retrieval" *Journal of Documentation* 28: 11-21.

[44] Volden, Craig. 2002. "A Formal Model of the Politics of Delegation in a Separation of Powers System." *American Journal of Political Science* 46(1):111-133.

[45] Volden, Craig and Alan Wiseman. 2011. "Formal Approaches to the Study of Congress." In Eric Schickler and Frances Lee, eds. *Oxford Handbook of Congress.* Oxford: Oxford University Press pp. 36-65.

[46] Wiseman, Alan E.. 2009. "Delegation and Positive-Sum Bureaucracies." *The Journal of Politics* 71(3): 998–1014.

[47] Wong, Wilson and Liu, Wei and Bennamoun, Mohammed, 2008. Determination of unithood and termhood for term recognition. *Handbook of Research on Text and Web Mining Technologies*. IGI Global.

Captions of Figures and Tables

Figure 1: Regulation and Delegation

Figure 2: Histogram of the Constraint Ratio and Number of Constraint Categories.

Figure 3: Four Measures of Executive Discretion.

Figure 4: Distribution of Discretion Index for Laws that Deregulate and Laws that Regulate.

Figure 5: Regulatory Discretion and Financial Sector Wages and Size over Time.

Figure 6: Trends in Constraints, Delegation and Agencies' Independence.

Figure 7: Topic model for the 120 laws with 5 topics. Each topic is summarized by its seven most frequent words and its seven words with the highest FREX score. Larger words are more frequent. Each word's position on the x-axis indicates its exclusivity to the topic.

Topic 1 terms and acronyms – Gramm-Leach-Bliley: a 1999 law that deregulated financial markets. NRSRO: Nationally Recognized Statistical Rating Organization, credit rating agency. SIPC: Securities Investor Protection Corporation. CFTC: Commodity Futures Trading Corporation, an agency that regulates futures and options markets.

Topic 2 terms and acronyms – RTC: Resolution Trust Corporation, a federal agency that operated between 1989 and 1996 and administered insolvent federal savings and loan institutions. FDIC: Federal Deposit Insurance Corporation, an agency which provides deposit insurance for banks. FED: Federal Reserve Bank. FSLIC: Federal Savings and Loan Insurance Corporation, an agency that until 1989 administered the insurance of federal savings and loan institutions. CRA: Community Reinvestment Act, a law to incentivize banks to meet the financial needs of lower income communities, particularly regarding mortgage lending. SAIF: Savings Association Insurance Fund.

Topic 3 terms and acronyms – HUD: U.S. Department of Housing and Urban Development. FNMA: Federal National Mortgage Association (Fannie Mae), providing mortgage-backed securities. FHA: Federal Housing Administration, a federal agency setting construction standards and insuring loans for home building.

Topic 4 terms: Patman refers to Wright Patman who was the chair of the Senate Committee on Banking and Currency for ten years.

Topic 5 terms and acronyms – FY: fiscal year, a term used in appropriations bill.

Figure 8: Measures used to compare topic models as we vary the number of topics K.

Figure 9: Probability distribution of the first difference of falling in each discretion class between a law covering a high risk policy domain versus a low risk policy domain.

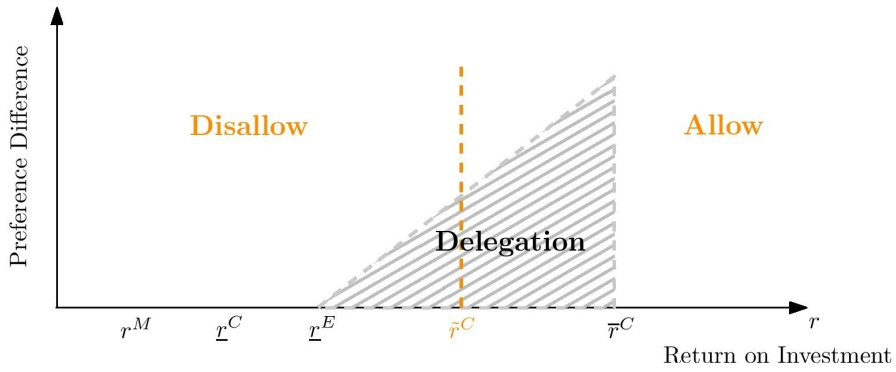
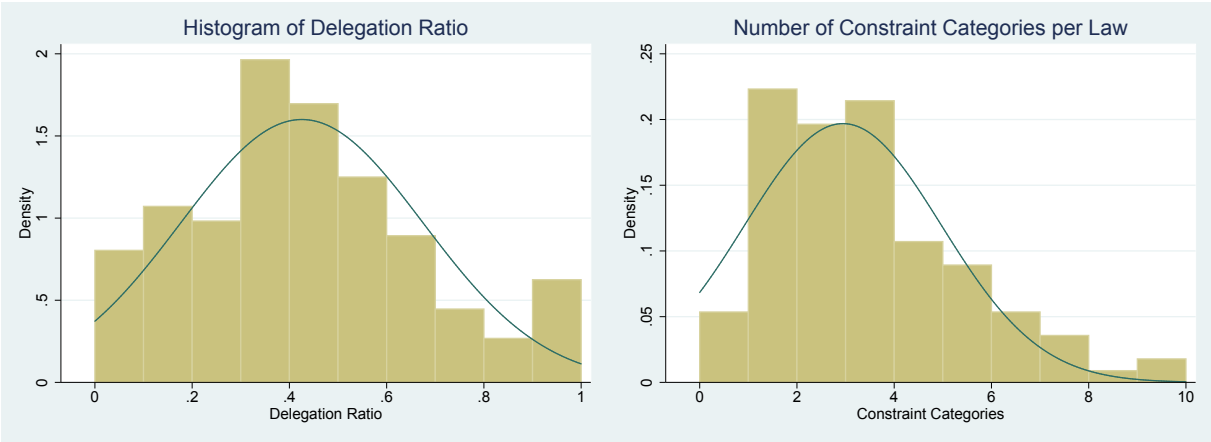


Figure 1: Regulation and Delegation.



(a) Delegation Ratio.

(b) Constraint Ratio.

Figure 2: Delegation and Constraint Ratios.

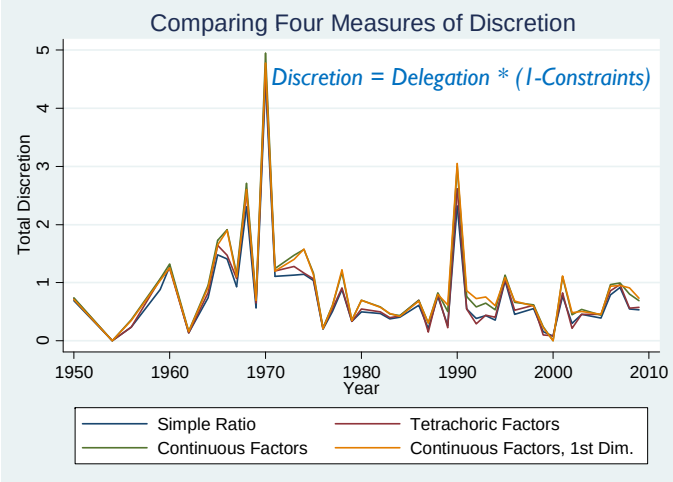


Figure 3: Four measures of executive discretion.

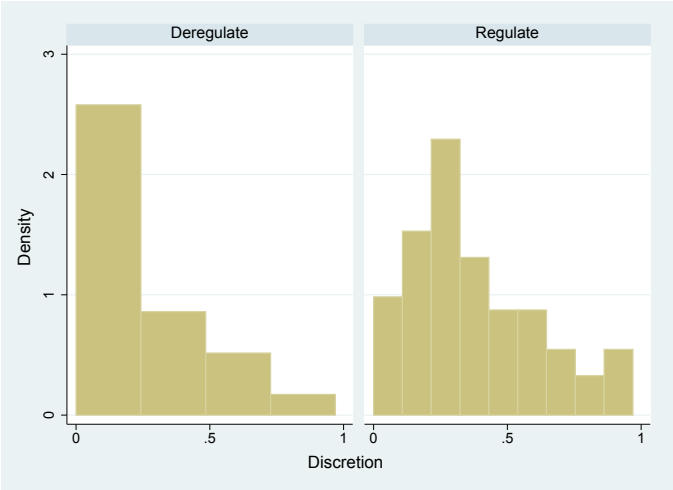
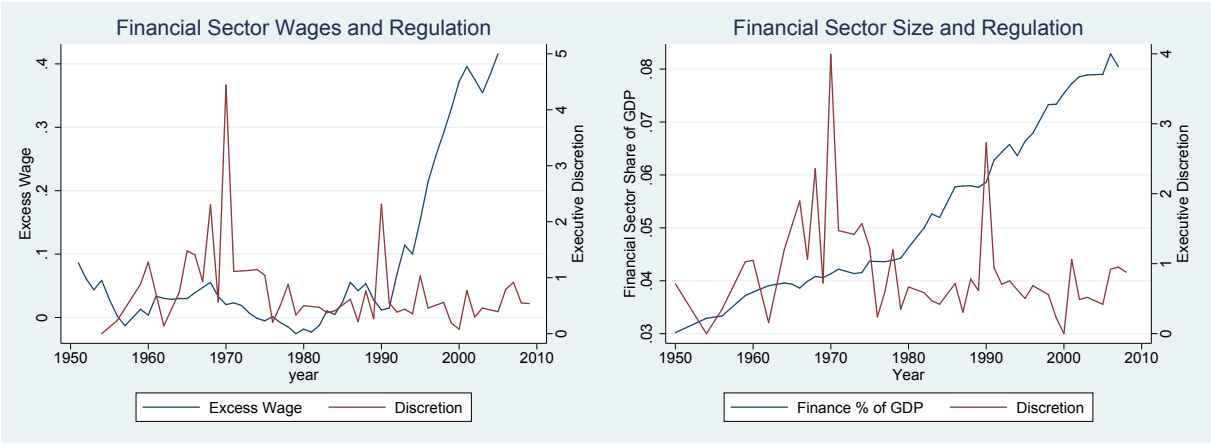


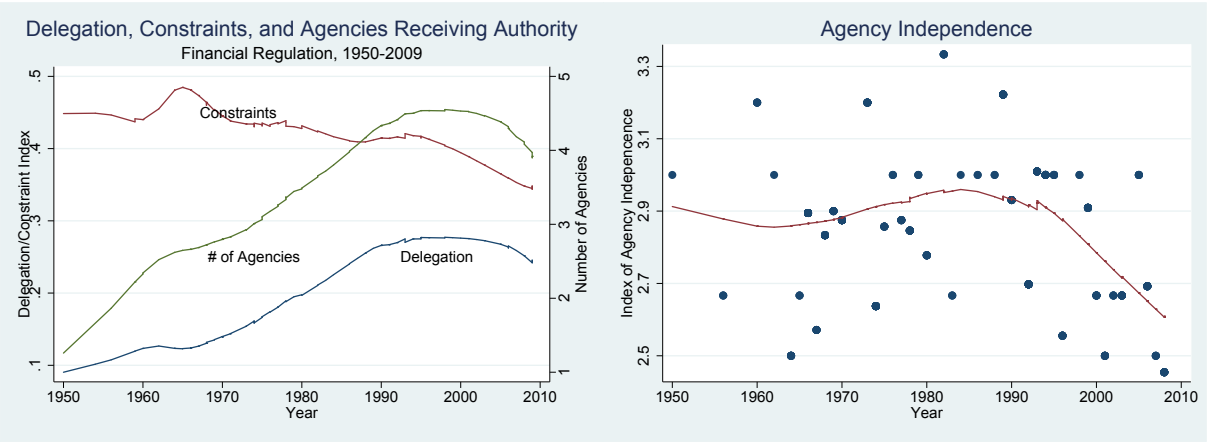
Figure 4: Distribution of Discretion Index for Laws that Deregulate and Laws that Regulate.



(a) Discretion and financial wages.

(b) Discretion and importance of financial sector.

Figure 5: Regulatory Discretion and Financial Sector.



(a) Delegation, Constraints, and Number of Agencies.

(b) Independence of Agencies.

Figure 6: Trends in Constraints and Delegation.

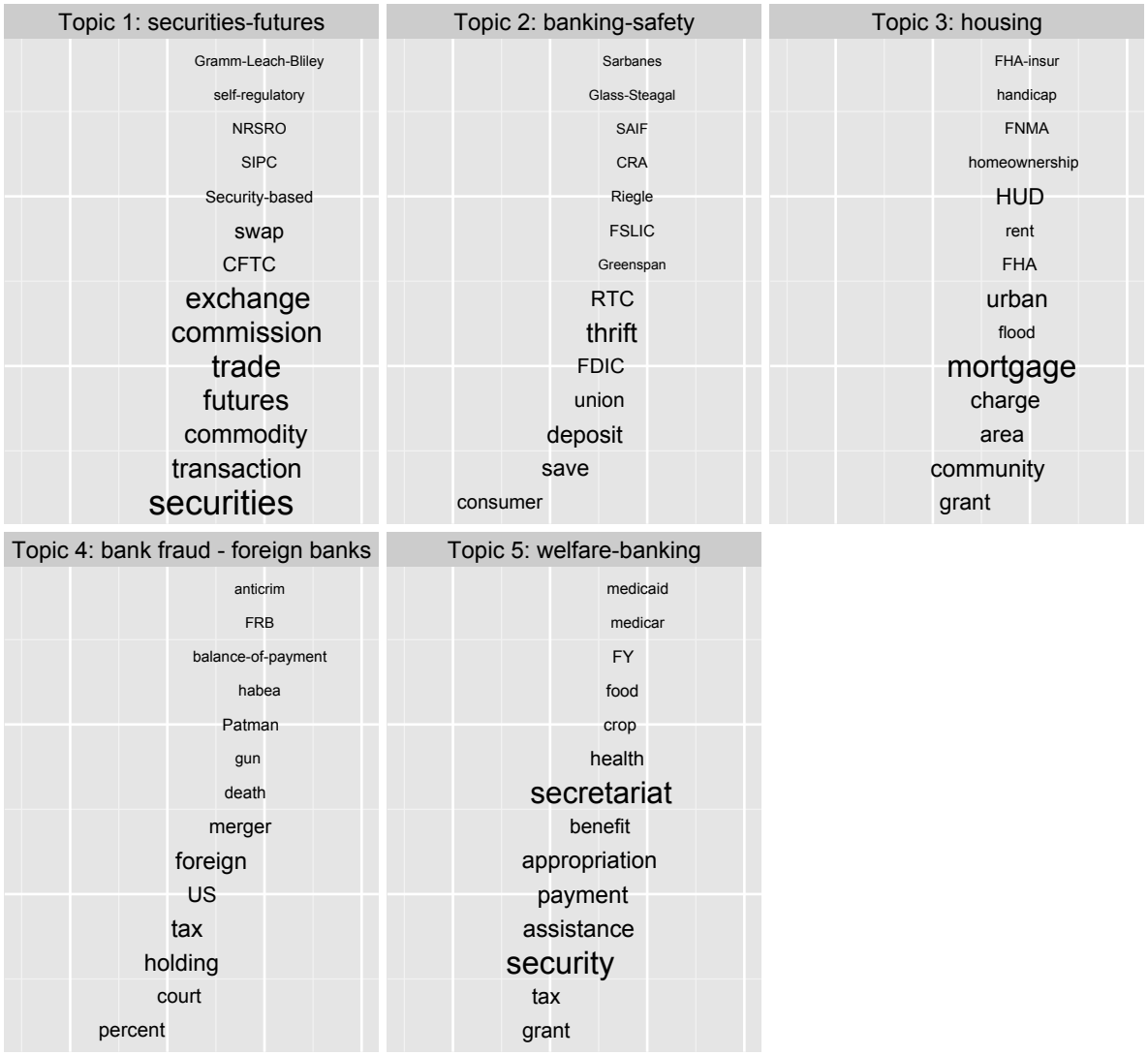


Figure 7: Topic model for the 120 laws with 5 topics. Each topic is summarized by its seven most frequent words and its seven words with the highest FREX score. Larger words are more frequent. Each word's position on the x-axis indicates its exclusivity to the topic.

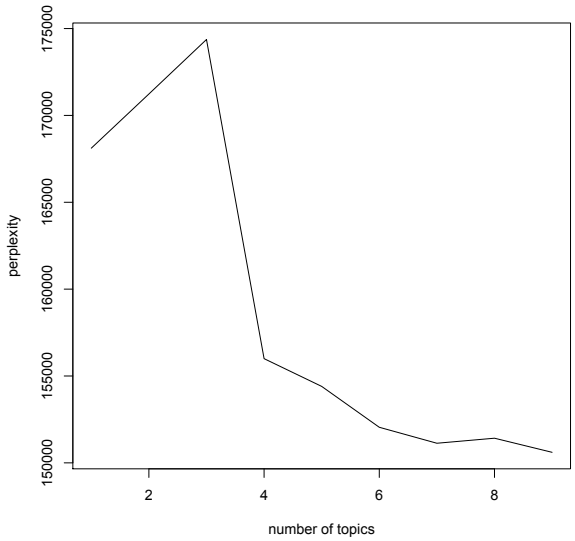
Topic 1 terms and acronyms – Gramm-Leach-Bliley: a 1999 law that deregulated financial markets. NRSRO: Nationally Recognized Statistical Rating Organization, credit rating agency. SIPC: Securities Investor Protection Corporation. CFTC: Commodity Futures Trading Corporation, an agency that regulates futures and options markets.

Topic 2 terms and acronyms – RTC: Resolution Trust Corporation, a federal agency that operated between 1989 and 1996 and administered insolvent federal savings and loan institutions. FDIC: Federal Deposit Insurance Corporation, an agency which provides deposit insurance for banks. FED: Federal Reserve Bank. FSLIC: Federal Savings and Loan Insurance Corporation, an agency that until 1989 administered the insurance of federal savings and loan institutions. CRA: Community Reinvestment Act, a law to incentivize banks to meet the financial needs of lower income communities, particularly regarding mortgage lending. SAIF: Savings Association Insurance Fund.

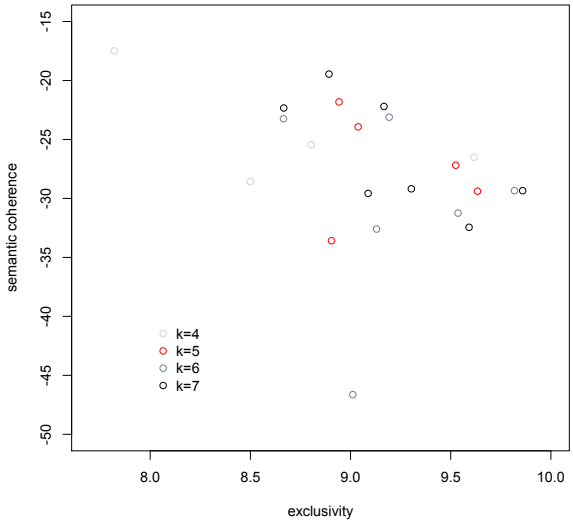
Topic 3 terms and acronyms – HUD: U.S. Department of Housing and Urban Development. FNMA: Federal National Mortgage Association (Fannie Mae), providing mortgage-backed securities. FHA: Federal Housing Administration, a federal agency setting construction standards and insuring loans for home building.

Topic 4 terms: Patman refers to Wright Patman who was the chair of the Senate Committee on Banking and Currency for ten years.

Topic 5 terms and acronyms – FY: fiscal year, a term used in appropriations bill



(a) Perplexity as a function of K



(b) Exclusivity and semantic coherence for different values of K

Figure 8: Different measures for comparing topic models.

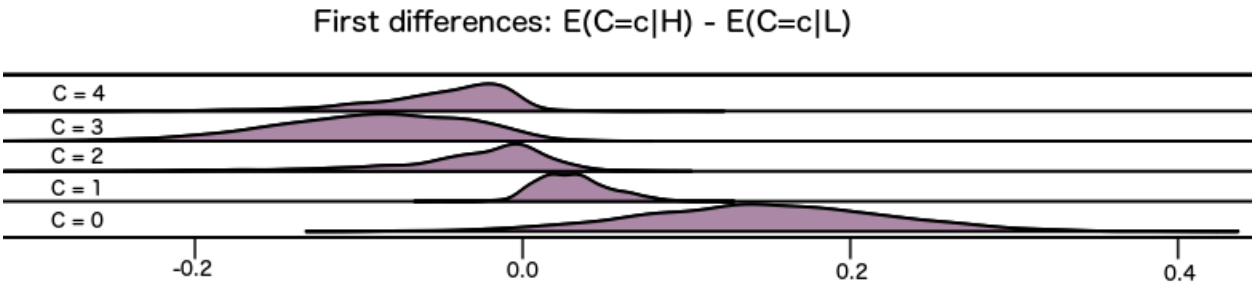


Figure 9: Probability distribution of the first difference of falling in each discretion class between a law covering a high risk policy domain versus a low risk policy domain.