Essays on Strategic Advocacy and Influence in Rulemaking

by

Bradley Alexander Hackinen

B.Sc., The University of Victoria, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Economics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2019

© Bradley Alexander Hackinen, 2019
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

**Essays on Strategic Advocacy and Influence in Rulemaking**

submitted by Bradley Alexander Hackinen in partial fulfillment of the requirements for

the degree of Doctorate of Philosophy

in Economics

**Examin ing Committee:**

Matilde Bombardini, Vancouver School of Economics  
Co-supervisor

Francesco Trebbi, Vancouver School of Economics  
Co-supervisor

Marit Rehavi, Vancouver School of Economics  
University Examiner

Wie Cui, Peter A. Allard School of Law  
University Examiner

Decio Coviello, HEC Montreal  
External Examiner

**Additional Supervisory Committee Members:**

Kevin Milligan, Vancouver School of Economics  
Supervisory Committee Member
Abstract

Politicians and regulators rely on external expertise when setting policies, providing an opportunity for interest groups to influence government policy with persuasive information. This process forms the foundation of many models of lobbying and special interest politics, but is very difficult to study empirically. I use data from U.S. Federal rulemaking process to create new measures of influence, revealing that U.S. corporations use charitable donations to non-profits to influence regulators, and opening up possibilities for future research.

In Chapters 1 and 2, I present co-authored research on the relationship between strategic advocacy in rulemaking and corporate philanthropy. For-profit corporations and non-profit entities are active in the rulemaking process and are arguably expected to provide independent viewpoints. However, non-profits also receive grants from firms with an interest in the outcomes of the regulatory process. Chapter 1 shows that corporations give large amounts of money to non-profits that participate in the rulemaking process and that, shortly after a firm donates to a non-profit, the non-profit is more likely to comment on rules for which the firm has also provided a comment. Chapter 2 extends the study of firms and non-profits to include the text content of regulatory comments, and the regulators response in the final rule. We find that, when a firm comments on a rule, the comments by non-profits that recently received grants from the firm’s foundation are systematically closer in content similarity to the firm’s own comments than to those submitted by other non-profits commenting on that rule. And, when a firm comments on a new rule, the discussion of the final rule is more similar to the firm’s comments when the firm’s recent grantees also comment on that rule. These patterns suggest that corporations strategically deploy charitable grants to induce non-profit grantees to make comments that favor their benefactors.

In Chapter 3, I present a new technique for estimating influence of comments on final rules that improves on the approach used in Chapter 2. These new estimators can also control for each author’s writing style with fixed effects.
Lay Summary

This dissertation explores how corporations and other interest groups influence public policy by communicating with policymakers during the design process. It focuses on U.S. federal regulators, who are legally required to accept public comments on proposed regulations and respond to the comments in the final version. The first two chapters present co-authored work. We discovered that large U.S. corporations use charitable donations to non-profits to influence regulators. Non-profits are more likely to comment on the same regulation as their donors in the year after receiving a grant, and the comments they submit are unusually similar in content to their donor’s comments. We find suggestive evidence that the strategy is effective by comparing the text of the resulting regulation with the comments themselves. In the third chapter, I present new and improved techniques for estimating how comments influence regulations that will be useful for future projects.
Preface

This dissertation is original, unpublished work.

Chapters 1 and 2 are based on a working paper titled “Hall of Mirrors: Corporate Philanthropy and Strategic Advocacy”, co-authored with Marianne Bertrand, Matilde Bombardini, Raymond Fisman, and Francesco Trebbi. The paper was the result of collaborative discussions between all the authors. My coauthors provided data on corporate donations to non-profits and assisted with implementing the co-commenting regressions and writing the working paper. I provided data on the rule-making process, handled all the data linking, and took the lead on developing and implementing the text analysis. I contributed to writing the working paper in the areas where we describe institutional features of the regulatory system, associated data collection, and the text analysis algorithm.

Chapter 3 is my own independent work.
Table of Contents

Abstract .................................................................................................................. iii
Lay Summary ........................................................................................................ iv
Preface .................................................................................................................... v
Table of Contents .................................................................................................... vi
List of Tables .......................................................................................................... viii
List of Figures ....................................................................................................... ix
Acknowledgements ............................................................................................... xii
Dedication ............................................................................................................... xiii
Introduction ........................................................................................................... 1

1 Corporate Philanthropy and Strategic Advocacy ..................................................... 7
   1.1 Introduction ................................................................................................... 7
   1.2 Institutional context: Rulemaking process ....................................................... 12
   1.3 Data .............................................................................................................. 12
      1.3.1 Charitable giving by foundations ......................................................... 13
      1.3.2 Comments and rulemaking ................................................................. 13
      1.3.3 Basic data facts .................................................................................. 14
   1.4 Evidence based on charitable giving and non-profit commenting on regulations 15
      1.4.1 Co-commenting specifications ............................................................. 16
      1.4.2 Regulation specification .................................................................... 18
   1.5 Getting paid not to comment: The role of hush money .................................. 18
   1.6 Concluding remarks .................................................................................... 20
   1.7 Figures and Tables ....................................................................................... 21

2 Text Similarity and Regulatory Influence ............................................................. 27
   2.1 Introduction .................................................................................................. 27
   2.2 Quantifying the similarity in content across regulatory comments ................ 27
      2.2.1 Sample construction ......................................................................... 31
List of Tables

1.1 Co-commenting - Time-in varying specification ........................................ 23
1.2 Co-commenting - Recent donation ............................................................ 24
1.3 Commenting on regulations ................................................................. 25
1.4 Hush money ......................................................................................... 26

2.1 Similarity of comments - Any donation ................................................... 37
2.2 Similarity of comments - Recent donation ............................................... 38
2.3 Sentiment - Any donation ...................................................................... 39
2.4 Sentiment - Recent donation ................................................................. 40
2.5 Rule-comment similarity - Any donation ............................................... 41
2.6 Rule-comment similarity - Recent donation .......................................... 42

3.1 Summary of Estimators .......................................................................... 79

A.1 Organization name extraction accuracy .................................................... 90
A.1 Top Agencies by Number of Comments ............................................... 91
A.2 Co-commenting in time-varying sample - Contemporaneous and lagged donations ..................................................... 92
A.3 Co-commenting in time-varying sample - Future donations ............... 93
A.4 Similarity - Contemporaneous and lagged donations ..................................... 94
A.5 Similarity - Sector Control ............................................................ 95
A.6 Similarity - Future Donation ............................................................. 96
A.7 Rule-comment similarity - Robustness ................................................... 97
A.8 Citations of Firm in Rule Discussion .................................................. 98
A.9 List of Agencies on regulations.gov (A-F) ........................................ 99
A.10 List of Agencies on regulations.gov (F-Z) ...................................... 100
## List of Figures

1.1 Rulemaking process. ................................. 21
1.2 Co-commenting and charitable donations. ................................. 22

3.1 Single-sample estimation: This grid of plots summarizes the performance of the main linear estimators when the same input documents are used for generating the rule and estimating influence. Each dot represents the mean performance on one docket across 100 samples with random $\beta$s. The solid horizontal line shows the average performance across all dockets. The horizontal axis is the number of comments in the docket. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, Log Mean Squared Error (MSE), Bias (mean $\beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \epsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). Vectors are generated from document-level LSA ($k = 200$). ................................. 58

3.2 Split-sample estimation: This grid of plots summarizes the performance of the main linear estimators when the tokens of each input document are split into separate “gen” and “est” samples. The “gen” samples are used to construct the final rule, while the “est” samples are used to estimate influence. Each dot represents the mean performance on one docket across 100 samples with random $\beta$s (with the same gen/est split). The solid horizontal line shows the average performance across all dockets. The horizontal axis is the number of comments in the docket. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, Log Mean Squared Error (MSE), Bias (mean $\beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \epsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). Vectors are generated from document-level LSA ($k = 200$). ................................. 59

3.3 Split-sample vector performance comparison: This grid of plots summarizes the performance of three different types of vectors in the VFGLS estimator when the tokens of each input document are split into separate “gen” and “est” samples. The “gen” samples are used to construct the final rule, while the “est” samples are used to estimate influence. Each dot represents the mean performance on one docket across 100 samples with random $\beta$s (with the same gen/est split). The solid horizontal line shows the average performance across all dockets. The horizontal axis is the number of comments in the docket. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, Log Mean Squared Error (MSE), Bias (mean $\beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \epsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). ................................. 60
3.4 Split-sample vector performance comparison by $k$: This grid of plots summarizes the performance of three different types of vectors in the VFGLS estimator when the tokens of each input document are split into separate “gen” and “est” samples. The “gen” samples are used to construct the final rule, while the “est” samples are used to estimate influence. Each dot represents the mean performance average performance across all dockets for vectors truncated to length $k$. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, Log Mean Squared Error (MSE), Bias ($\text{mean } \beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \epsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$).

3.5 Split-sample weighting performance comparison: This grid of plots summarizes the performance of four token weighting schemes in the VFGLS estimator when the tokens of each input document are split into separate “gen” and “est” samples. The “gen” samples are used to construct the final rule, while the “est” samples are used to estimate influence. Each dot represents the mean performance on one docket across 100 samples with random $\beta$s (with the same gen/est split). The solid horizontal line shows the average performance across all dockets. The horizontal axis is the number of comments in the docket. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, Log Mean Squared Error (MSE), Bias ($\text{mean } \beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \epsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). Vectors are generated from document-level LSA ($k = 200$).

3.6 This grid of plots summarizes the performance of the panel estimators when the rules are generated with no style effects. Each dot represents the mean estimate for one commenter across 1000 samples with random $\beta$s. The solid horizontal line shows the average performance across all commenters. The horizontal axis is the similarity between the commenter’s style and the regulator style, as measured by the correlation between $\bar{x}$ and $\bar{y}$. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, log Mean Squared Error (MSE), Bias ($\text{mean } \beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \epsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). Vectors are generated from document-level LSA ($k = 100$).

3.7 This grid of plots summarizes the performance of the panel estimators when the rules are generated with style effects. Each dot represents the mean estimate for one commenter across 1000 samples with random $\beta$s. The solid horizontal line shows the average performance across all commenters. The horizontal axis is the similarity between the commenter’s style and the regulator style, as measured by the correlation between $\bar{x}$ and $\bar{y}$. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, log Mean Squared Error (MSE), Bias ($\text{mean } \beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \epsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). Vectors are generated from document-level LSA ($k = 100$).
3.8 This grid of plots shows the point estimates for all 1000 samples for three randomly selected commenters (indicated by colour) using panel VOLS. Each horizontal axis indicates the true value being estimated, while the vertical axis indicates the estimate. Points along the 45 degree line are correctly estimated. In the top row, rules are generated with no style effects, while in the bottom row the rules are generated with style effects. In the left column the estimator has no fixed effects, while in the right column the estimator uses style fixed effects.
Acknowledgments

I would like to thank to the Izaak Walton Killam Memorial Fund, Social Sciences and Humanities Research Council, and Vancouver School of Economics PhD program for providing funding that supported this research. I am grateful to my two excellent supervisors Matilde Bombardini and Francesco Trebbi for helping to set me on this research path, and for the support, guidance, and motivation they provided along the way. Kevin Milligan was a valuable committee member, and provided both useful perspective and detailed feedback at each stage of developing this research. I was lucky to meet many wonderful people in the PhD program who made the good times better and the hard times much more bearable. Most of all, I would like to thank Amelja Zoehner for her patience and support throughout the process, and Arthur and Fiona, who arrived part-way through the PhD to remind me that there are more important things than research.
To Amelja, Arthur, and Fiona
Introduction

In the U.S., corporations and other special interest groups spend billions of dollars annually on campaign donations, lobbying, and other more secretive activities in the hope of shaping public policy at federal, state and local levels (Drutman, 2015; Bertrand et al., 2018, 2014). There are also strong “revolving door” patterns in many industries (where people move back and forth between private and public-sector employment), and becoming a lobbyist after public service is the expected career path for members of congress and their staff (Vidal et al., 2012). If we want to understand why we see the policies we do, or how to improve them, we need to understand how these activities influence government decision-making. Many researchers have explored these questions of political influence, but the literature is biased toward theoretical analysis and restricted by the scarcity of good data.

In this dissertation I explore what can be learned about political influence using large-scale data analysis of U.S. federal rulemaking. The U.S. federal rulemaking process provides an unusually transparent view of interactions between special interest groups and government. Under the Administrative Procedures Act of 1946, regulators who make new rules (regulations, interpretive guidance, and other policies with legal weight) must provide a mechanism for public input and justify their decisions in relation to that input. The most important format for feedback is the notice-and-comment process for new rules. In this process, regulators publish a proposed version of the new rule, provide a way for the public to submit written comments, and then summarize those comments and respond the main arguments in the text of the final rule. This process is now fully digital, and both comments and regulatory documents can be downloaded in bulk. U.S. federal agencies have published about 2,500 rules per year over the last two decades, and received millions of public comments. I have collected and organized a near-universal sample of comments and regulations for U.S. federal rulemaking since 2003 containing 6 million public comments, and hundreds of thousands of documents published by regulators. The scale, detail, and breadth of this data across agencies, industries, and commenters could transform our understanding of political influence.

To capitalize on the potential of rulemaking data we must overcome two challenges. First we must link documents to each other to form a usable database, identify the authors of comments, and link comments to rules. Second, we must construct a measure of influence from the text of the documents. My work in this dissertation addresses both challenges.

I have largely overcome the challenge of linking the documents by using several custom machine-learning based algorithms. In my data, comments are linked to regulatory documents, proposals are linked to rules, and I have identified about 1 million comments submitted by specific organizations. Chapter 1 shows how useful this linked data can be. With Marianne Bertrand, Matilde Bombardini, Raymond Fisman, and Francesco Trebbi, I study how firms use corporate philanthropy to influence regulators. We show that non-profit organizations often comment on the same regulation as their donor.
firm, and that the probability of this type of co-commenting is much higher in the year following a
donation.

I have also made progress on the challenge of extracting useful information about influence from
the text of the documents. In Chapter 2 I use Latent Semantic Analysis (an existing topic modeling
algorithm) to measure similarities between text in different documents. I show that comments between
firms and non-profit organizations are more similar when the firm has donated to the non-profit. And
by comparing the text of comments to the final rule, we find patterns that suggest the strategy of
donating to non-profit organizations may indeed help firms shape public policy.

In Chapter 3 I show how influence can be measured more directly and precisely with new estimators
that combine ideas from topic modeling and econometrics. The estimators overcome several challenges
that make large-scale analysis of rulemaking data difficult or impossible with existing techniques, and
improve on the LSA similarity measure developed in chapter 2 in terms of precision and interpretability.
Another benefit is that the new estimators naturally extend to a panel data setting and allow for the
separation of fixed writing style from regulation-specific comment content.

Together, these chapters represent a significant step forward in the study of political influence using
large-scale data from the rulemaking process.

Related Literature

The research in this dissertation contributes to a growing part of the political economy literature that
uses empirical techniques to study how private firms and other special interest groups influence public
policy in the U.S. and other advanced economies. The defining idea behind Political Economy is
that there is no benevolent and omniscient government agent to turn to for policy design. Instead,
government policy is the product of imperfect institutions and actors, who are often motivated by self
interest. When we study influence, we hope to understand how actors outside the government affect
decisions made inside the government.

Theoretical foundations

The theoretical foundation of the literature on influence is very general, particularly early theoretical
research that explores lobbying and special interest politics. Part of this literature evolved out of Public
Choice theory, which aims to understand how large-scale empirical patterns (like industry-specific trade
barriers or economic subsidies) can be explained by the outcome of interactions between rational agents
who represent special interest groups and government. The workhorse model that developed out of
this literature is the common-agency model of lobbying (Rodrik, 1995; Grossman and Helpman, 1994;
Persson and Tabellini, 2000). This model takes the existence of different organized groups as given.
Each group has a corresponding local public good which benefits only the group, and the government
taxes all players to provide public goods to each group. Groups can be either organized to lobby or not.
Lobbying influence is modeled as a quid-pro-quo exchange, where interest groups bid to provide rents
to the government in return for increased provision of their preferred local public good. The model
is simple to work with, and generates the result that public goods are over-provided to organized
(lobbying) groups, and under-provided to the disorganized groups. Because it is simple and agnostic about institutional details, the model is also easy to apply in many different contexts. The common agency model gives us the intuition that industries that are better organized politically (or have more concentrated ownership) will obtain more preferable outcomes from the rulemaking process.

The common-agency model is closely related to the idea of “regulatory capture” (Stigler, 1971), where a particular firm or industry becomes so effective at influencing regulators that policies are designed primarily for industry benefit. Theories of regulatory capture emphasize the principle-agent relationship between the main government agent and the agency tasked with regulating a particular industry, and how the agency has an incentive to collude with industry instead of maximizing the government’s objective. Like the common-agency model, regulatory capture does not depend on a particular mechanism of influence, and the literature in this area considers a variety of mechanisms, from simple bribery (Dal Bó, 2006), to career incentives that accompany revolving door relationships (Cohen, 1986; Makkai and Braithwaite, 1992; Lucca et al., 2014), to more subtle forms where agency views about optimal policies are shifted through repeated interactions with industry or industry propaganda (Shapiro, 2012).

An important mechanism of influence that appears in many models is strategic information provision. Bribery and simple quid-pro-quo exchanges are theoretically simple, but do not provide satisfying explanations for many of the activities that special interest groups engage in such as lobbying, commenting on rulemaking, or media campaigns that appear to be primarily about shaping the information available to policymakers. There is a large literature that seeks to understand this type of communication from a game-theoretic perspective (Grossman and Helpman, 2001). The general assumption is that special interest groups are also domain experts with private information that is relevant for policy design. The government can improve policy if special interest groups truthfully report their information to the policymaker. But special interest groups have an incentive to mislead the government into designing policies in a way that maximizes the interest group’s private benefit instead of the government objective. When communication is costly, we call it a “signaling” model, and the cost of communicating makes the choice to send a message meaningful in a way that can convey information in itself. When communication is free, we refer to the setting as a “cheap talk” model, and successful transmission of information is more complex.

The literature on cheap talk is enormous, and has some surprising twists and turns. In the canonical model by Crawford and Sobel (1982), the interest group (sender) communicates to the government (receiver) about a state of the world that is represented by a point on the one-dimensional unit interval. The authors show that the amount of information that can be transmitted depends on the alignment of preferences between the interest group and government. When the interest group has similar preferences to the government they can transmit high quality information in equilibrium, but as their preferences diverge, the precision of the information that they can transmit decreases. As intuitive as it is, this result turns out to be quite sensitive to the setup of the model. Battaglini (2002) shows that in a model where multiple senders communicate with a receiver in a multi-dimensional, unbounded state-space, it is possible to convey information perfectly regardless of the preferences of senders or receiver. As a rejoinder, Takahashi and Ambrus (2008) recover some of the intuition of the Crawford and Sobel
(1982) with an extension that adds bounds back into the multi-dimensional state-space. When the state-space is bounded, some messages are no longer credible, and the relative preferences of interest groups and government matter once more. Similar sensitivity to modeling assumptions can be seen in models with different timing, presence of reputation, or other variations (for example, see Aumann and Hart (2003); Ottaviani and Sørensen (2006); Chakraborty and Harbaugh (2010)). Cheap talk models also always have multiple Nash equilibria, adding another degree of freedom in the form of equilibrium selection. One takeaway from this literature is that the details matter in these strategic information transmission games, and that it is hard to predict what types of communication will be possible and who will benefit from them without extremely precise information about the institutional structure and payoffs to all players.

**Empirical lobbying studies**

The research in this dissertation is closely related to empirical studies of lobbying. These studies aim to understand how interest groups use lobbying to influence policy, who lobbies who, how effective lobbying is at achieving interest group objectives, and the role of lobbyists as intermediaries between interest groups and government. De Figueiredo and Richter (2014) provide an excellent overview of established results.

Ideally, empirical research can answer questions that theoretical analysis cannot, especially in areas where it is not clear what assumptions are reasonable. In practice, empirical research faces its own limitations. One constraint is that there are only a small number of good data sources. The most popular data comes from the U.S. federal *Lobbying Disclosure Act*, which requires that lobbyists record how much clients pay them and what issues they lobby on. Some states and cities release similar data. Lobbying expenditure is most useful when examining the market for lobbying services. For example, Vidal et al. (2012) use the retirement of politicians to show that lobbyists are valued primarily for their political connections and they lose business when their contacts in government leave office. And, Bertrand et al. (2014) find that lobbyists move from topic to topic according to who they know in government, not what they know about policy issues. However, lobbying expenditures are hard to connect to an outcome. Lobbyists only need to report what branch of government they are contacting and a general topic (sometimes a bill). The data does not reveal anything about the specific messages conveyed from interest groups to politicians, and it is hard to attribute changes in policy to specific lobbying interventions.

Some researchers have used structural modeling as a substitute for more complete data. In trade policy there is a strand of research which investigates the relationship between political influence and trade protections (Goldberg and Maggi, 1999; Bombardini, 2008). In this context, trade models clearly suggest that protected industries benefit at the expense of unprotected industries. More recently, Huneeus and Kim (2018) look at the proximity of politicians to corporate headquarters and feed this instrument into a structural monopolistic competition model to estimate how lobbying affects the distribution of firm sizes. In this model, the shift in firm sizes results in a large decrease in aggregate productivity. These structural approaches are valuable, but depend heavily on modeling assumptions that are difficult to verify. For example, Huneeus and Kim (2018) assume that the effects of lobbying are
zero-sum, and only reallocate benefits between firms. But perhaps the information conveyed through lobbying also has benefits (as suggested by some cheap talk models), such as reducing the probability of the government implementing economically disastrous policies. If the benefits of lobbying are included, this might substantially change the implied effects.

Another approach is to code approximate data about the goals of interest groups and policy outcomes. Baumgartner et al. (2009) undertake a large study of congressional lobbying where they manually assign groups as either seeking to maintain or seeking to change the status quo on issues, and also determine whether the policy did in fact change. They find little correlation between lobbying effort and whether the status quo changes, and argue that this is actually an intuitive outcome when considering that the status quo already reflects the relative power of different groups.

The empirical literature on lobbying makes it clear that more detailed data on interactions between interest groups and government would be very valuable. Rulemaking offers an under-exploited source of data of this type.

Empirical rulemaking studies

The literature on rulemaking itself is quite small, and most of the most of the research comes from law and political science. Golden (1998) provides an early analysis of public comments, focusing on eleven randomly selected rules from the EPA, NHTSA and HUD. Golden finds that most comments come from business interests, and that nine out of the eleven rules are changed at least a small amount in response to content in comments. Wagner et al. (2011) perform a remarkably thorough analysis of the rulemaking process for the EPA’s Air Toxic Emission Standards. By tracking contacts between the EPA and interest groups over each rulemaking cycle (from pre-proposal to final rule), the authors show that the process is dominated by input by industry groups, but that public interests are more active early in the process. Comments come primarily from industry, and the EPA is more likely to reduce the stringency of rules in response to comments than to increase stringency.

Susan Webb Yackee and co-authors advanced the study of rulemaking with a series of related studies. Yackee (2005) provides the another analysis of whether interest groups are able to affect policy through the notice-and-comment process, showing that regulators often adapt rules to address commenter recommendations, particularly when there is consensus among commenters. Yackee and Yackee (2006) shows that most comments come from business interests, and that these interests are more successful in affecting policy. McKay and Yackee (2007) explore the role of interest group competition in rulemaking, showing that agencies respond most the the group that dominates the commenting process. Naughton et al. (2009) emphasizes the importance of interest groups that engage early in the process when the rule is less well defined. Haeder and Yackee (2015) show how commenters can also be influential by communicating with the Office of Management and Budget which oversees parts of the rulemaking process on behalf of the president.

Daniel Carpenter and Brian Libgober also have several working papers that explore rulemaking. Libgober and Carpenter (2018a,b); Libgober (2018) use stock price event studies of U.S. banks around the time of rule announcements to estimate the financial effects of rules promulgated under the Dodd-Frank Act, and argue that the observed patterns imply the effect of comments is large (“worth billions”).
In related work, Libgober and Rashin (2018) analyze the content of comments to determine what strategies commenters are using to influence regulators. They find that most commenters provide pure information, with a smaller fraction communicating legal threats.

Together, these existing empirical studies of rulemaking represent a path-breaking exploration of influence in the rulemaking process. The main limitation is that they are all based on relatively small sets of rules and comments that were manually coded by the authors. My dissertation focuses on expanding the analysis to the universe of comments and rules. Expanding the data has several advantages: First, it allows us to more fully map the territory - it is not clear whether each result above holds in general, or only for specific industries and time periods. Second, having a large sample with a panel structure allows more flexible specifications including using high-dimensional fixed effects to control for confounders, and enough precision to extract information from those estimates. Finally, broader data allows us to ask basic questions like how the relationship between comments and rules differs across industry, agency, and political cycle, as well as painting a more complete picture of how each firm or interest group interacts with multiple agencies.

**A path forward**

Ultimately, it would be nice to be able to explore questions of policy design. For example, how can the institutions through which interest groups interact with government (such as congressional lobbying or the notice-and-comment process for rulemaking) be designed to maximize the information that special interest groups provide to government and minimize zero-sum rent-seeking? Or, given the institutions that exist, how will other economic policies feed back into the political system? Some researchers and public intellectuals are concerned that lax U.S. anti-trust policy has allowed large firms to grow in political as well as economic power, at the expense of more equal democratic representation (Zingales, 2012). Should we be concerned about feedback from economic power to political influence? Finally, global challenges like climate change require rapid and large-scale social changes to address. Corporate interests appear to have played a large part in downplaying climate risk and delaying government action (Oreskes and Conway, 2011; Mayer, 2017). Are there institutional changes that would help improve how environmental policy is determined? Answering any of these questions requires a better understanding of how interest groups affect government policy – not just in broad strokes, but details about who gains and loses, and how these outcomes change under different economic and institutional conditions. Large-scale analysis of rulemaking data provides a promising path forward.

This dissertation is structured as follows. Chapter 1 presents the results of a co-authored study using rulemaking data to examine how non-profits advocate on behalf of their corporate donors. Chapter 2 extends this analysis to the text content of comments and rules using existing topic-modelling technology. Chapter 3 proposes novel estimators for measuring the effect of comments on rules from text data.
Chapter 1

Corporate Philanthropy and Strategic Advocacy

1.1 Introduction

Economists and political scientists have long studied – both theoretically and empirically – the role interest groups play in the formation of laws and regulations. In the U.S., as in many democracies, there are well-established channels through which interest groups can try to influence the laws and rules that may impact their communities, their businesses, or society at large. Through means such as lobbying, grassroots campaigns, testimonies, or public advocacy, interested parties inform politicians and bureaucrats of the costs and benefits of government action.

While interest groups may have expertise on topics of direct relevance to them, they may also be tempted to present information that is tainted by their self-interest. This logic is at the core of the literature on informational lobbying (Grossman and Helpman, 2001). For example, oil company representatives may have expertise in drilling, but also a strong incentive to minimize, say, the predicted environmental costs of Arctic oil exploration. Government officials must thus weigh both the quality of information and its impartiality, based in part on its source. As such, lawmakers and rulemakers may view information provided by for-profit corporations as less credible if that information is not corroborated by other groups with non-aligned (i.e., neutral or opposing) interests.

Non-profit organizations often fall into the role of interests that are non-aligned with business. Some non-profits – such as research groups, universities, and think tanks – are providers of non-partisan, technical expertise and are commonly expected to offer more neutral input into the lawmaking and rulemaking process, with a focus on cost-benefit analysis and broader societal interests. Other non-profits – such as human services organizations, environmental protection groups, social welfare organizations, and advocacy groups – may have opposing interests to business, to the extent that laws or regulations that benefit their members (or those on whose behalf they advocate) adversely constrain business profits. Non-profit organizations are therefore expected to play an important balancing role in the informational lobbying process.

This role may be subverted, however, by the financial links between corporations and non-profits: in exchange for donations, a non-profit may (consciously or otherwise) take a perspective that is favorable to its benefactor’s bottom line. If politicians and bureaucrats are more likely to implement a proposal when it is supported by interest-group diverse coalitions (as suggested theoretically in the strategic

---

1 By informational lobbying we refer to the broad literature on information transmission which encompasses cheap talk and costly signalling models in the context of lobbying, for example Potters and Van Winden (1992), Austen-Smith (1993), Austen-Smith (1995) and Lohmann (1995).
advocacy literature, e.g., Krishna and Morgan (2001), Dewatripont and Tirole (1999), Dewatripont and Tirole (2005), and empirically in Lorenz, 2017) and if such ties are undisclosed, such “coalition building via corporate giving” may distort the outcome of the political process away from the public good and towards private interests.2

The goal of this chapter is to provide systematic evidence establishing this to be an empirically relevant phenomenon. The context of U.S. Federal Regulation, with its far-reaching economic implications and its carefully documented record of communications between organizations and government agencies, offers an ideal setting to establish such evidence.

There exists anecdotal evidence that these concerns are well-founded. Across a range of issues and regulatory agencies, researchers and journalists have documented cases of companies using charitable contributions to co-opt ostensibly neutral and even non-aligned non-profits. Notably, Peng (2016) describes the efforts of telecommunications firms to win merger approvals in front of the Federal Communication Commission (FCC), in part by assembling diverse and vocal coalitions of supporters. Peng quotes Crawford (2013) on the Comcast-NBCU merger, in which “[t]he company encouraged letters to the FCC from more than one thousand non-profits...including community centers, rehabilitation centers, civil rights groups, community colleges, sports programs, [and] senior citizen groups.” For the AT&T/T-Mobile merger, Peng similarly documents letters of support addressed to the FCC from non-profits that, at first glance, would appear to have little interest or expertise in telecommunications policy, including a homeless shelter in Louisiana, a special needs employment agency in Michigan, and the Gay & Lesbian Alliance Against Defamation (GLAAD). The non-profits were all AT&T Foundation grantees (in the case of the homeless shelter, the donation had come in just five months before the merger was announced). In no case did the non-profit disclose its AT&T funding in its letter to the FCC. In at least one case, the comments did not appear to represent the views of the non-profit membership. According to Peng, “GLAAD’s president and six board members resigned when its merger endorsement made headlines and revealed that the organization had received AT&T funds.”

Journalists and medical experts have documented similar persuasion-via-donation in public health debates. Jacobson (2005), for example, describes a (“no-strings attached”) $1 million donation from Coca-Cola Foundation to the American Association of Pediatric Dentistry (AAPD), accompanied by a shift in the tone of AAPD statements on sugary beverages, from describing soft drinks as “a significant factor” in tooth decay, to describing the scientific evidence of the relationship as “unclear.” 3 Similar concerns have been raised with respect to the role of donations from corporations to university research

---

2Implicitly we are presuming that Coasian bargaining in the political sphere does not already lead to efficient policies. To the extent that, for example, it is difficult to contract across multiple regulatory agencies and/or pieces of legislation (let alone make outright side payments), one may think of the government as aiming to set optimal policy on a rule-by-rule basis, assigning winners and losers in each instance. See, e.g., Acemoglu (2003), for a discussion.

3A more direct link to policy can be found in the soda industry’s efforts against New York City’s ban on large sugary drinks in the 2010s. In his decision to strike down the Bloomberg administration policy, the presiding judge cited amicus briefs filed by two New York non-profits (the local chapter of the NAACP and the Hispanic Federation), which argued that the ban would disproportionately affect ethnic and racial minority groups. Both non-profits were recipients of funds from Coca-Cola and PepsiCo. See “Minority Groups and Bottlers Team Up in Battles Over Soda.” The New York Times March 12, 2013. Aaron and Siegel (2017) show that 95 national public health organizations received funding from Coca-Cola and PepsiCo during 2011-2015; the study does not look, however, at the effect on organizations’ publicly stated positions.
hospitals.4

Investigative journalists have also documented many instances of companies influencing the policy statements of “neutral” non-profits that ostensibly provide evidence-based analysis on matters of public interest. Confidential memos and documents suggest that some think tank reports are discussed with corporate donors before the research is complete, with donors potentially shaping the final reports, so that the resulting “scholarship” can be used to corroborate their separate lobbying efforts. In her 2017 book Dark Money, journalist Jane Mayer, provides one prominent example, documenting how the philanthropic activities of the billionaire industrialist brothers Charles and David Koch furthered their efforts to influence political discourse: “[The Koch brothers] subsidized networks of seemingly unconnected think tanks and academic programs and spawned advocacy groups to make their arguments in the national political debate. [...] Much of this activism was cloaked in secrecy and presented as philanthropy, leaving almost no money trail that the public could trace. But cumulatively it formed, as one of their operatives boasted in 2015, a ‘fully integrated network.’” Raising concerns about such practices in general, Senator Elizabeth Warren, also a commercial law professor, observed that, “[t]his is about giant corporations who figured out that by spending, hey, a few tens of millions of dollars, if they can influence outcomes here in Washington, they can make billions of dollars.”5

In this chapter we show that the patterns discussed in these anecdotes hold more broadly in a setting in which we can plausibly draw a strong circumstantial connection between corporate donations and the participation of non-profits in the political and regulatory process.

We focus on the formation of federal rules and regulations. Federal agencies in the U.S. are legally required to publish proposed rules in the Federal Register and accept public comments on those proposals before rules are finalized and comments discussed.7 While there is no legal requirement for agencies to act on feedback received in comments, the agencies themselves often attribute changes between proposed and final rules to arguments made via rulemaking. As emphasized by Sunstein (2012),

---


6 Warren also commented on the use of these practices in the rulemaking context, which we focus on in the empirical analysis below: “Unlike congressional action, agency rules are constrained by well-established judicial review standards that seek to determine whether the agency’s action is supported by the evidentiary record and the authority delegated to it by Congress. Rules must be supported by ‘substantial evidence’; agency actions must not be “arbitrary and capricious.” But corporate players are savvy. They have learned that those same judicial review standards can be used to suffocate new rules. They play a sophisticated game—leverage their own expertise and paying outside experts with purportedly independent credentials to produce long, detailed comments filled with data and analyses, all selectively produced to serve their own interests.” Discussing fixes, she also writes: “Another [principle] would be to help agencies and courts distinguish between legitimate, high-quality data and research, on the one hand, and bought-and-paid-for studies on the other, by requiring disclosure of financial arrangements and editorial relationships associated with regulatory comments.” See https://www.theregreview.org/2016/06/14/warren-corporate-capture-of-the-rulemaking-process/ (accessed October 31, 2018).

7 The Administrative Procedures Act of 1946, 5 U.S.C. 553(c) states: “...the agency shall give interested persons an opportunity to participate in the rule making through submission of written data, views, or arguments with or without opportunity for oral presentation. After consideration of the relevant matter presented, the agency shall incorporate in the rules adopted a concise general statement of their basis and purpose.” https://www.law.cornell.edu/uscode/text/5/553. Accessed October 31, 2018.

8 There are some exceptions for urgent actions or cases in which the change is so trivial that the agency does not expect comments, but in general, agencies which fail to publish a sufficiently informative proposal or fail to follow the commenting procedure can have their regulations vacated in court.
public commentary is also a valuable source of feedback to preempt regulatory mistakes “when the stakes are high and the issues novel.” Regulations.gov provides the largest single source for comment information on proposed rules, and was rolled out in 2003 when most agencies started a systematic effort to digitize the commenting process. By 2008, 80% of all proposed rules provided a regulations.gov link for commenting, and the fraction is about 90% as of 2018.

For the purpose of this chapter, we use regulations.gov to build a comprehensive dataset including the majority of the comments submitted in the rulemaking process since 2003. For each comment, we know the specific proposed rule the comment is in response to, as well as the content (text file) of the comment and the identity of the commenter. We may thus connect specific organizations to commentary on the same proposed regulation and its final discussion (we refer to a sequence of rule postings from proposal to final version as a “regulatory stream” or docket).

We complement the commentary data with information on corporate foundations and their beneficiaries, using data on charitable donations by foundations linked to large corporations through tax forms filed to Internal Revenue Service (IRS). The combination of these datasets allows us to explore whether (i) non-profits that benefit from corporate philanthropy are more likely to comment on the same rule as their benefactors; (ii) conditional on both providing feedback on the same regulation, the non-profits’ comments are unusually similar to that of their benefactors; and (iii) co-comments by a corporate foundation’s grantees lead to discussions of the rule by the regulator that use language that is more similar to the language contained in the company’s comments. By exploiting the particular timing of corporate donations and comments, as well as the inclusion of firm-grantee pair fixed effects, we argue that we can plausibly draw a compelling link from funding to co-commentary and comment overlap.9

Our sample of firms is comprised of the companies that have appeared at any point in the 1995 to 2016 lists of Fortune 500 or S&P 500 (or both) for which we identify a corporate foundation, and our sample of non-profits is the set of all grantees that received at least one donation from these foundations over the period 1998-2015. Organizations (firms or non-profits) are linked by name via a fuzzy match to 981,232 rulemaking comments made on all proposed regulations on regulations.gov during the years 2003-2017. The main sample for our analysis is comprised of the 414 corporations with charitable foundations and 11,746 grantees that commented at least once during this period.

We show that non-profits are more likely to comment on the same regulation as their benefactors, and that this “co-commentary” is most strongly associated with donations in the year preceding the comments. We consider several specifications: first, a cross-sectional regression that reveals strong correlations between donations and commenting on the same regulation independent of any particular timing. Second, we break the commenting information into annual bins and show that there non-profits are particularly likely to comment on the same regulation as their donor when they have received a donation in the year prior to the comment. This result is robust to the inclusion of pair-wise firm-grantee fixed effects which control for any fixed relationship between the donating and commenting of

---

9 Of course, this does not obviate the possibility that non-profits have time-varying policy preferences, and corporate gifts coincide with shifts in these preferences. While we cannot rule out this possibility (a critique that applies even to the Coca-Cola/AAPD example mentioned above), our approach does help to rule out the possibility that latent, time-invariant shared interests drive both donations and comment overlap.
firms and non-profits. Our analysis implies that a donation in the preceding year is associated with nearly a doubling in the likelihood of co-commentary. We also run a similar regression at the regulation level and show that non-profits are more likely to comment on a particular regulation when they have received money from one of the firms that also comment. The regulation specification allows us to include regulation fixed effects which control for the fact that some regulations attract much more commenting interest than others.

Finally, we explore whether corporations use charitable donations to encourage otherwise opposing voices to remain silent (rather than encouraging non-profits to provide supportive commentary). While it is challenging to devise a decisive test to detect the omission of comments that might otherwise have been made, we provide suggestive evidence, based on an extension of our main results on donations and co-comment frequency, that “hush money” may not be of first-order importance in our setting. More specifically, we show that the link between co-commentary and donations is strongest in areas in which a non-profit most commonly provides comments, the opposite of what one might expect if hush money played a dominant role.

Our findings first and foremost provide a contribution to the literature on the mechanisms by which interest groups seek to influence government policy (for canonical early contributions see, for example, Grossman and Helpman, 1994, 2001 and for a more recent discussion Baumgartner et al., 2009; Bertrand et al., 2014; Drutman, 2015). We differ from much of this prior work in our focus on influence via expert commentary rather than financial contributions and, much more importantly, in documenting one mechanism by which private interests may cloak biased advice by inducing its provision by a non-obviously aligned party. This has implications for how we model the process of governmental information acquisition (Austen-Smith, 1993; Laffont and Tirole, 1993), and is also of direct policy relevance. Our results suggest that calls for restrictions on financial relationships among those aiming to influence government policy may be well-founded, and that at a minimum potential conflicts-of-interest statements should be required for any organization providing input on government regulations (Peng, 2016). Our work is also related to prior research that has shown the value of coalitions of diverse interest groups in the adoption of government policy. In particular, studying bills introduced in Congress between 2005 and 2014, Lorenz (2017) shows that bills supported by interest-diverse coalitions are more likely to receive committee consideration; in contrast, Lorenz (2017) finds no association between committee consideration and lobbying coalitions’ size, or their interests’ PAC contributions. Generalizing beyond the lawmaking process, this work complements our findings in that it suggests that corporations can expect some return for the type of charitable “investments” we uncover in this chapter. Other papers that have focused on returns to lobbying instead include Bombardini and Trebbi (2011, 2012); Kang (2016); Kang and You (2016). Finally, this chapter expands on earlier work highlighting how corporations may strategically use their corporate philanthropy as an undisclosed tool of political influence. Bertrand et al. (2018) show that corporations allocate more of their charitable giving to congressional districts that are more relevant to the corporations due to the committee assignments in the House of Representatives of their elected representatives. We identify in this chapter another, independent, category of “strategic CSR” (Baron, 2001) in the government arena.
1.2 Institutional context: Rulemaking process

The rulemaking process of U.S. federal agencies provides a context in which we may observe both the presence and the content of communication by different entities with an interest in influencing the policymaker. While informational lobbying at the federal or local level does not come with statutory requirements of disclosure of the content or even the exact target of communication, the rulemaking process consists of a series of codified procedures that regulate the activity of federal agencies in the production of “rules” under the Administrative Procedure Act of 1946 (APA). The subject of policy deliberation is a rule “designed to implement, interpret, or prescribe law or policy,” according to the APA. The process of rulemaking may be set in motion by Congress passing a new law requiring implementation or by an agency itself, upon regularly surveying its area of legal responsibility and identifying areas that need new regulations.

Figure 1.1 sketches the process of informal rulemaking. It starts with a Notice of Proposed Rulemaking (NPRM) including the objective of the rule and how it would modify the current Code of Federal Regulations. The NPRM is published in the Federal Register, at which point the agency specifies a period of 30 to 60 days during which the public can submit comments on the proposed rule. After comments have been received and additional information collected, the agency may proceed to publish a Final Rule in the Federal Register or issue a Supplemental Notice of Proposed Rulemaking if the initial rule was modified substantially, in which case further comments are invited. This notice-and-comment procedure is meant to include the general public and all interested parties in the crafting of the new rule. Importantly, the agency also publishes in the Federal Register a discussion of the goals and rationale of the policy, and how the comments were incorporated into the final rule in the Supplementary Information section of the final rule.

Occasionally, the process of rulemaking requires merging or splitting specific elements of a rule, issuing interim versions of the rule if the process is delayed, and more generally adapting to other external factors, including further direction from the legislative branch, and so forth. These various additional documents are typically filed in dockets maintained by the regulatory agency.

Upon finalization of the rule, comments represent part of the official record, and rules can be challenged judicially on procedural or substantive grounds based on comments filed by entities that participated in the rulemaking. Judicial review is an important constraint to rulemaking activity in the United States in that it effectively forces regulators to attend to opinions expressed via commentary.

1.3 Data

This section introduces our sources and provides a brief overview of the data. For further details we refer to Appendices A and B. We begin by describing the data on charitable giving by corporate foundations, followed by the data on public comments on rulemaking. The starting point for our

---

10 Under the Lobbying Disclosure Act of 1995, lobbying registration and reporting forms only require lobbyists to list the topic and the agency lobbied (e.g., Trade, the Senate of the United States), in addition to clients and payments. See Vidal et al. (2012); Bertrand et al. (2014).

11 Agencies may decide to engage in rulemaking under the recommendation of congressional committees, other agencies, or following a petition from the general public.
sample is the set of corporations that have appeared at any point during the period 1995 to 2016 in the Fortune 500 and/or S&P 500 lists, which counts 1398 firms.\textsuperscript{12}

\subsection*{1.3.1 Charitable giving by foundations}

Data on charitable donations by corporate foundations come from FoundationSearch, which digitizes publicly available Internal Revenue Service (IRS) data on the 120,000 largest active foundations in the U.S. We find 629 active foundations that can be matched by name to 474 of the initial list of 1398 firms.\textsuperscript{13} As noted in Brown et al. (2006), larger and older companies are more likely to have corporate foundations, which results naturally from the fixed cost of establishing a foundation.\textsuperscript{14}

Each charitable foundation must submit Form 990/990-P-F “Return of Organization Exempt From Income Tax” to the IRS annually, and this form is open to public inspection. Form 990 includes contact information for the foundation, as well as yearly total assets and total grants paid to other organizations. Schedule I of Form 990, entitled “Grants and Other Assistance to Organizations, Governments, and Individuals in the United States,” specifically requires the foundation to report all grants greater than $5,000. For each grant, FoundationSearch reports the amount, the recipient’s name, city and state, and a giving category created by the database.\textsuperscript{15}

While the IRS assigns a unique identifier (Employer Identification Number, EIN) to each non-profit organization, FoundationSearch does not report this code, so we rely on the name, city and state information to match a grantee to a master list of all non-profits. This list, called the Business Master File (BMF) of Exempt Organizations, is put together by the National Center for Charitable Statistics (NCCS) primarily from IRS Forms 1023 and 1024 (the applications for IRS recognition of tax-exempt status). The BMF file reports many other characteristics of the recipient organization, including address, assets and a non-profit sector called the National Taxonomy of Exempt Entities (NTEE). The results of the matching between all public charities, private foundations or private operating foundations (designated as 501(c)3 organizations for tax purposes) in the BMF and the recipients of charitable giving by 2014 Fortune 500 and S&P 500 companies is reported in Bertrand et al. (2018).

\subsection*{1.3.2 Comments and rulemaking}

The source of data on comments, proposed, final and interim rules, as well as discussion of final rules is regulations.gov, a website through which the majority of U.S. federal agencies collect public comments in the notice-and-comment phase of rulemaking.\textsuperscript{16} The website regulations.gov API provides a search function for document metadata.

\textsuperscript{12}The initial number of firms is 1434, but we combine firms that merge during the sample, hence obtaining a smaller total number.

\textsuperscript{13}The 629 foundations we find are linked to 474 corporations, since there are instances of multiple foundations associated with the same corporation.

\textsuperscript{14}They also find that state-level statutes – in particular laws relating to shareholder primary and the ability of firms to consider broader interests in business decisions – predict establishment of a foundation. Various endogenous financial variables are also predictive of foundation establishment. The analysis in Brown et al. (2006) is cross-sectional, so their variables are absorbed by the various fixed effects in many of our analyses.

\textsuperscript{15}The 10 broad categories are: Arts & Culture, Community Development, Education, Environment, Health, International Giving, Religion, Social & Human Services, Sports & Recreation, Misc Philanthropy.

\textsuperscript{16}For the complete list, see Appendix tables A.9 and A.10.
Our research sample consists of all comments posted to regulations.gov in the years 2003-2017. We use a custom machine learning tool to extract organization names from the comment metadata. The algorithm identified 981,232 comments that appear to be authored by organizations (as opposed to private individuals) and downloaded the full text of the comments. We are particularly interested in comments submitted by non-profits and by corporations that we observe in our FoundationSearch sample. The comments are linked to corporations’ and grantees’ names through a custom name matching tool that implements multiple types of fuzzy matching and manual corrections.\footnote{Available from the authors upon request.}

The unit of observation is what regulations.gov refers to as a docket. This is a way for agencies to organize comments that relate to a particular topic. Most straightforwardly, one may think of a one-to-one correspondence between a rule and a docket. Conceived in this way, as mentioned above, a docket will contain all comments that pertain to all versions of that rule. An example of a simple docket is FNS-2006-0044 from the Food and Nutrition Service (FNS) which contains a proposed rule (06-09136) and its corresponding final rule (E8-21293) on “Fluid Milk Substitutions in the School Nutrition Programs.” All comments in this docket therefore are easily linked to this regulatory stream. There are more complex cases in which a docket contains multiple proposed rules and notices (see, for example, docket EPA-HQ-OAR-2008-0699, the Environmental Protection Agency’s review of the National Ambient Air Quality Standards for Ozone). We associate all comments to the same docket given the homogeneity of the topic. The only exception is when we turn to examine the wording of the discussion of final rules as a function of corporate and non-profit comments. There, we will consider each rule within the docket separately to ensure a finer connection between comments by corporations and the exact wording of the final rule in the docket under discussion, as the multiple rules associated within a docket are discussed and published separately in the Federal Register. We will elaborate on this distinction in Section 2.3, which discusses those results.

1.3.3 Basic data facts

Recall that our sample starts with the set of companies that appeared at least once in the Fortune 500 or S&P 500 lists between 1996 and 2015. Of the 1398 firms in that sample we find 909 that have commented at least once in the period 2003-2016.\footnote{We only consider comments starting in 2003 because this is when the comments database is complete.} This is the sample of firms that forms the basis of our regressions. We have a total of 22,654 firm comments over 5,792 docket. Of these 909 firms 414 have a foundation.

In terms of non-profits we start from the 225,180 entities that received at least one grant from any foundation in our sample over the period 1998-2015. Our sample consists of the 11,531 of these grantees that comment at least once at any point during the period starting in 2003. We have a total of 318,841 comments in 8,729 docket from those grantees.

There is vast heterogeneity among firms in their activity in the commenting phase. The most actively commenting firm, Boeing, provided comments on 1284 docket. On average each firm comments on 18 docket, but the distribution is skewed: the median firm comments on 6 docket, while the firms at the first and third quartile comment on 2 and 17 docket, respectively. The distribution of comments
among grantees is even more skewed. On average each grantee comments on almost 5 dockets, but the median is 1 and the third quartile is 3 dockets. The most active grantee (Center for Biological Diversity) comments on 905 dockets.

Tables A.1 reports the agencies that receive the highest number of comments from grantees and firms. At the top of the list for grantees are the EPA (Environmental Protection Agency), the FAA (Federal Aviation Administration) and the FDA (Food and Drug Administration). The top three agencies as recipients of grantees’ comments are the FWS (Fish and Wildlife Service), the NOAA (National Oceanic and Atmospheric Administration), and the HHS (Health and Human Services Department). It is worth noticing that the EPA, the FAA and the FDA feature in the top 10 agencies for grantees as well.

Finally, we provide some information on the prevalence of commenting behavior of grantees and their co-commenting with firms in our sample. In our regressions we will often focus on “recent” donations, defined as donations from a firm to a grantee that occur in the same year, or one year prior to a public comment on a rule. Consider the set of all firms-years where the firm has commented at least once and donated recently. We can break the recipients of these recent donations in to a set of nested groups with increasingly close ties to the firm.

Firms donate to an average of 327 non-profit grantees. Of these, an average of 54 grantees ever submit a comment in our sample. Within these “commenters”, 28 non-profits ever comment to one of the same agencies as the firm (not necessarily at the same time or in the same year), 8 ever co-comment on a regulation with the firm, and 1.4 co-comment with the firm that year.

In terms of expenditures, the average total amount spent on donations over a two year period is $26 million dollars, 26% of which go to grantees that ever comment. Within commenting grantees, expenditures are biased towards the grantees with closer commenting ties to the firm. Thus, grantees that never comment to the same agency as the firm receive an average of $103,412 each, and grantees that comment to one of the same agencies as the firm, but never on the same regulation, receive $156,348 each, while those that ever co-comment with the firm receive an average of $240,515 each and grantees that co-comment with the firm that specific year receive $206,994 each.

1.4 Evidence based on charitable giving and non-profit commenting on regulations

This section focuses on the link between firms and non-profits through charitable grants, and establishes a relationship between firm-grantee financial ties and their tendency to comment on the same regulations.

We denote firms/foundations by \( f \in F \) and grant-receiving non-profits (“grantees”) by \( g \in G \). Let \( D_{fgt} \) be an indicator function that takes a value of 1 if we observe a donation from firm \( f \) to grantee \( g \) in year \( t \), and 0 otherwise. The indicator function \( C_{frt} \) is equal to 1 if firm \( f \) comments on regulation \( r \) in year \( t \), and 0 otherwise (throughout this section and the following one, “regulation” or “rule” will refer to a docket). The indicator function \( C_{grt} \) is defined similarly and is equal to 1 if grantee \( g \) comments

\(^{19}\)Agency acronyms are listed in Appendix tables A.9 and A.10.
on regulation \( r \) in year \( t \), and 0 otherwise. A graphical representation of this configuration is described in Figure 1.2.

We adopt two types of specifications: co-commenting specifications and a regulation specification.

### 1.4.1 Co-commenting specifications

We begin by relating the event of a firm and a grantee commenting on the same regulation to a financial tie between the two in the form of a charitable donation. The indicator function \( CC_{fgrt} \) is equal to 1 when donor \( f \) and grantee \( g \) comment on the same regulation \( r \) at time \( t \), so that \( CC_{fgrt} = C_{frt} \times C_{grt} \), and 0 otherwise. Our first specification explores a time-invariant link between co-commenting and donations, aggregating co-commenting to the firm-grantee pair, so that we define a new indicator \( CC_{fg} \) which is equal to 1 if we observe any co-commenting from firm \( f \) to grantee \( g \) in our sample, and 0 otherwise. That is, \( CC_{fg} = I(\sum_r \sum_t CC_{fgrt} > 0) \). Similarly, the indicator variable \( D_{fg} \) indicates whether we observe any donation from \( f \) to \( g \) in our sample.

We first consider the following time-independent specification that relates the presence of co-commenting by firm \( f \) and grantee \( g \) to the presence of a donation within the same pair:

\[
CC_{fg} = \beta_0 + \beta_1 D_{fg} + \delta_f + \delta_g + \epsilon_{fg}.
\]  (1.1)

The specification includes firm fixed effects \( \delta_f \) to capture the potential bias resulting, for example, from the higher probability that large and profitable firms both donate to charities and comment on multiple regulations. Similarly, we include grantee fixed effects \( \delta_g \), to control, for example, for the fact that charities that are more successful at fundraising may on average have more resources to devote to commenting on various regulations. A positive coefficient \( \beta_1 \) would indicate that firm-grantee pairs that are connected by donations are also more likely to comment on the same regulations.

The results are reported in Table 1.1. The different columns of Table 1.1 include different sets of fixed effects (and clustering dimensions) of increasing levels of stringency. Of particular interest is column (4), the most conservative specification, that includes both grantee and firm fixed effects. The firm fixed effects may account for the average propensity of firms to comment and to donate, which may depend on size and sector. The grantee fixed effects can capture the average level of commenting activity of the grantee, which may in turn be related to its size and overall resource endowment. Across all specifications in Table 1.1, we can see that a grantee and a firm are more likely to comment on the same regulation when we observe any donation from the firm to the grantee. The magnitude of this effect is large. The baseline probability of co-commenting for a firm and a grantee is 2.16\%, meaning that of all the possible pairs of grantees and firms only 2.16\% comment on the same rule at any point in time. This probability increases by 4 to 8 percentage points when we observe a donation connecting firm and grantee. Put differently, the presence of a donation is associated with a two- to four-fold increase in the probability of co-commenting.

Of course, this cross-sectional pattern of co-commenting may stem from the fact that firms contribute to non-profits sharing similar objectives and views, or that, more simply, operate in similar sectors. For instance, the Bayer Science & Education Foundation associated with Bayer US, a phar-
maceutical company, may be more likely to donate to healthcare-related research non-profits, and both Bayer and healthcare-related non-profits may be more likely to comment on healthcare-related regulation than an average organization.

Our second specification addresses this concern, and further allows us to control for the general tendency by some firms to comment on certain issues and to contribute to non-profits that operate in related areas. It does so by focusing on the timing of donations. In particular, we examine whether co-commenting is more likely in the year immediately following the presence of a donation. For this, we turn to the following panel specification, which exploits time variation in both co-commenting and donations:

$$CC_{fgt} = \beta_0 + \beta_1 D_{fgt-1} + \delta_{fg} + \delta_t + \varepsilon_{fgt}$$

(1.2)

where $CC_{fgt} = I(\sum_r CC_{fgrt} > 0)$ indicates whether firm $f$ and grantee $g$ comment on the same regulation at time $t$, and $D_{fgt-1}$ is equal to 1 if we observe a donation from $f$ to $g$ in the concurrent ($t$) or preceding ($t - 1$) year of the comments, 0 otherwise. This specification includes firm-grantee fixed effects $\delta_{fg}$ and time fixed effects $\delta_t$. Therefore, $\beta_1$ is estimated only employing within-pair variation over time in donations and co-commenting. In particular $\beta_1$ will detect whether, controlling for the average tendency of a certain firm $f$ to co-comment with and donate to a specific non-profit $g$, we observe co-comments occurring immediately after a donation from $f$ to $g$ has been made.

Given the coarseness of the data along the time dimension (we only observe year of comment), it is possible for a comment to be made in, say, January of 2006 and a donation in June 2006; hence we can only be certain that the lagged-year donation took place prior to co-commenting. In Table 1.2, we report results in which we create a dummy that is equal to 1 if we observe a donation at either $t$ or $t - 1$, and 0 otherwise.\(^2^0\) Our preferred specification in Table 1.2 is column (5), where we include firm-grantee pair fixed effects. This specification exclusively exploits variation within a firm-grantee pair in donations and in co-commenting. The $\delta_{fg}$ pair fixed effects control not only for the higher probability of donation and co-commenting for firms and grantees in the same sector, but also for the general ideological alignment of firm and grantee that may result in both donations and co-commenting on similar topics.

We find a robust association between donations in year $t - 1$ and the likelihood of co-commenting in year $t$. The magnitude of effects is large in this panel specification. Co-commenting is obviously more sparse in equation (1.2) than equation (1.1): of all firm-grantee-year triples only 0.163% feature co-commenting. In column (4) of Table 1.2, the presence of a recent donation is associated with a quadrupling of the probability of co-commenting. In column (5) of Table 1.2, the presence of a recent donation is associated with an 81% increase in the likelihood of co-commenting, even after controlling for the general propensity of a specific firm to give to and as well as co-comment with a specific grantee. The example provided in the introduction, which described AT&T Foundation grantees such as GLAAD or a homeless shelter commenting on the AT&T/T-Telecom merger close on the heals of receiving donations, provide an illustration of the behavior implied by this statistical evidence (see

\(^{20}\)In Appendix Table A.2 we separate contemporaneous and lagged donations and find that lagged donations strongly predict co-commenting, while contemporaneous donations are a weak predictor of co-commenting.
As a further robustness exercise, in Appendix Table A.3 we augment our preferred specification with a dummy for whether firm $f$ donated to $g$ in year $t+1$. In column (5) of that table, with the most restrictive set of fixed effects (i.e. pair fixed effects), we find that donations made immediately after the commenting period are not associated with co-commenting, whereas only immediately preceding donations are. This pattern further confirms that co-commenting seems to be more prevalent after we observe a donation from firm to grantee.

### 1.4.2 Regulation specification

In the specifications we have considered thus far, we have aggregated co-commenting across different rules within a $fg$ pair or $fgt$ pair-year. We now present an alternative approach that links commenting by a grantee to donations received by a firm that also comments on the same rule $r$. The following “regulation specification” relates the probability of commenting by a grantee on a regulation $r$ to donations received:

$$C_{gr} = \beta_0 + \beta_1 I \left( \sum_f D_{fg} \times C_{fr} > 0 \right) + \delta_g + \delta_r + \eta_{gr}$$

where $C_{rg}$ is equal to 1 if $g$ comments on regulation $r$ (0 otherwise) and $DonorComment_{gr} = I \left( \sum_f D_{fg} \times C_{fr} > 0 \right)$ is equal to 1 if $g$ receives a donation from any firm that comments on $r$, and 0 otherwise. This specification includes regulation fixed effects $\delta_r$, which capture how certain rules are subject to more intense commenting, and grantee fixed effects $\delta_g$, that account for factors like resources and size of the non-profit, which may make $g$ both more visible and more likely to comment on any regulation.

Table 1.3 reports estimates of $\beta_1$ under different fixed effects and clustering options. Our preferred specification in column (4) has docket and grantee fixed effects, as well as two-way clustering on these attributes. When considering all the possible combinations of grantees and rules, we find a comment in 0.039 percent of the cases. It is not surprising that this number is small, since the universe of all possible grantee-rule pairings involve non-profits, like the Red Cross, that we would not expect to comment on, say, financial regulation. Starting from this baseline probability of commenting on a specific rule, we find that the probability that the non-profit comments is three to five times higher when a donor firm commented on the same rule, a result that accords with our previous results under specification (1.2).

### 1.5 Getting paid not to comment: The role of hush money

Sections 1.4-2.3 focused on the role of donations from corporations to non-profits in generating additional messages that are more similar to the donor’s position. In our final set of results, we examine whether corporations also use donations for a distinct strategic purpose: to silence opposing opinions. It is plausible to envision an informational lobbying environment in which agents supporting a
specific action opposed by a counterparty may be motivated to suppress these opposing voices (and compensate the counterparty for its silence). For example, in a discussion of the strategies employed in the multi-year campaign of the tobacco industry Lando (1991) writes: “The tobacco industry has been effective in purchasing what has been described as ‘innocence by association’. Tobacco industry sponsorship of sports events is notorious. The industry has also contributed substantially to the arts, to women’s groups, and to organizations representing minorities. These types of pernicious industry activities have been successful in buying the silence or the tacit support of some groups that have suffered a disproportionate share of the tobacco burden.” Payment in exchange for inaction and silence is commonplace in the market (e.g. noncompete, nondisclosure agreements, non-disparagement clauses, etc.) and such private agreements or clauses do not represent per se invalid contracts or violations of free speech. They may be, however, private agreements that are undisclosed to regulators, who may interpret the silence of some parties to the regulatory process as informative.\textsuperscript{21}

The role of such “negative” strategies is thought to be crucial to the success of special interest groups in politics. Blocking unfavorable bills from ever seeing the light of day (or committee discharge) in the U.S. Congress is as much a part of lobbying as facilitating the passage of bills favorable to an industry. Similarly, interest group comments in rule making often involve aim to kill unfavorable provisions or stalling the implementation of rules. (“Nothing happening” is almost always the desirable policy outcome for incumbent industry, see Baumgartner et al., 2009.)

To test for the presence of “hush money” in rule making, we propose an extension of our empirical framework in Section 1.4. In particular, we modify the regulation specification in Section 1.4.2 as follows:

\[ C_{gr} = \beta_0 + \beta_1 \text{DonorComment}_{gr} + \beta_2 \text{DonorComment}_{gr} \times \text{ShareComments}_{gR} + \delta_g + \delta_r + \eta_{gr} \quad (1.3) \]

where \text{DonorComment}_{gr} is equal to 1 if grantee \( g \) received a donation from a firm that also commented on the same regulation, and 0 otherwise. \text{ShareComments}_{gR} is the number (or share) of comments from \( g \) that are directed at rules under agency \( R \) over the entire sample. This new variable captures how common it is for grantee \( g \) to comment on rules from agency \( R \).

To understand the intuition behind this test, observe that certain non-profits may have specific expertise or focus in a specific area of regulation, which we approximate by the identity of the agency overseeing the rule (e.g., the Sierra Club commenting on rules proposed by the EPA).\textsuperscript{22} Interacting \text{ShareAgency}_{gR} with the donation from a commenting firm, \text{DonorComment}_{gr}, aims to establish whether such donations have a differential effect on the likelihood of commenting for grantees that typically comment on rule considered by agency \( R \), versus grantees that normally do not comment on rules by \( R \). We argue that this interaction is useful for assessing the potential role of hush money, as within the set of issue experts (high \text{ShareComments}_{gR}), it more likely that donations are made with the aim of inducing silence and muting commentary. A plausible null hypothesis supporting the presence of hush money is therefore \( \beta_2 < 0 \), as charitable donations may be more likely to be hush

\textsuperscript{21} Absence of a signal is in fact informative in games of incomplete information in which Bayesian rationality is assumed. For an applications to elections see Kendall et al. (2015).

\textsuperscript{22} A similar approach was followed to define issue expertise of individual lobbyists from federal lobbying reports in Bertrand et al. (2014).
money for grantees that routinely comment on rules from $R$.

Our results based on this specification and reasoning suggest that hush money is not a common strategy in our setting. In Table 1.4 we present several specifications accounting for the nonlinearity in equation (1.3), adding increasingly conservative sets of fixed effects across the six columns. The evidence points clearly in the direction of donations increasing co-commenting from grantees that routinely comment on rules from the regulator proposing $r$. The coefficient $\beta_2 > 0$ is systematically positive and highly statistically significant, indicating that firms are more likely to induce – rather than stifle – comments from such grantees. While this does not completely rule out the existence of hush money, it suggests that it is at a minimum less prevalent than the co-commenting behavior documented in Sections 1.4-2.3.

1.6 Concluding remarks

Politicians (and voters) are frequent targets of messages aimed at persuading them of the merits of specific policy positions. While in most cases the identity of senders is disclosed, allowing an assessment of the bias and interests of the originators of the message, in other cases their identity may be obscured, and deliberately so. These situations range from the use of dark money in U.S. electoral politics in the aftermath of the Supreme Court’s decisions of *Citizens United v. Federal Election Commission* and *McCutcheon v. Federal Election Commission* to the circulation of white papers by think tanks and non-profits.

In such circumstances, a common trait identified by the qualitative literature reviewed in this article is the reliance on independent arms-length organizations to extend the credibility of the positions held by special interest groups. While in most cases such overlap of intent and opinion is genuine, one has to be careful in assessing those cases where such support is offered in close proximity to monetary donations from corporations to advocate non-profits. Such transfers, often in the form of charitable grants, are virtually undetectable by private citizens and civil servants without access to detailed tax forms. Thus, these transfers represent potential forms of distortion that cannot be weighted and assessed in decision making.

In order to provide a quantitative and systematic perspective on this issue, this chapter studies the interaction of non-profit organizations and large corporations within the United States federal regulatory environment. We offer systematic empirical evidence which underscore new findings in the literature on corporate philanthropy and special interest politics. The chapter presents evidence that the charitable grants of corporate foundations reach targeted non-profits just before those same non-profits engage in public commentary. In the next chapter, we extend this analysis to the text of the comments and rules.
1.7 Figures and Tables

Figure 1.1: Rulemaking process
Figure 1.2: Co-commenting and charitable donations
### Table 1.1: Co-commenting - Time-invarying specification

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Grantee $g$ and firm $f$ comment on same regulation $\times 100$</th>
<th>Mean</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grantee $g$ received donation from firm $f$</td>
<td>8.156*** (0.393)</td>
<td>6.044*** (0.506)</td>
<td>6.276*** (0.215)</td>
<td>4.033*** (0.484)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed Effects</td>
<td>Firm Y, Grantee Y, Firms + Grantee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>11,111,716 11,111,716 11,111,716 11,111,716</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** The dependent variable is equal to 100 if grantees $g$ and $f$ comment on the same regulation in any year between 2003 and 2016. The independent variable is equal to one if grantees $g$ received a donation from firm $f$ in any year between 2003 and 2016. Standard errors are clustered at the level indicated in each column under “SE Clusters.” *** $p<0.01$, ** $p<0.05$, * $p<0.1$
Table 1.2: Co-commenting - Recent donation

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Firm $f$ and grantee $g$ commented on the same regulation in year $t$</th>
<th>Mean 0.163</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Firm $f$ contributed to grantee $g$ in year $t$ or $t-1$</td>
<td>1.153***</td>
<td>0.960***</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.065)</td>
</tr>
<tr>
<td>Fixed effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Grantee</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Donor</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Grantee-Firm Pair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE Clusters</td>
<td></td>
<td>Grantee</td>
</tr>
<tr>
<td>Observations</td>
<td>136,400,199</td>
<td>136,400,199</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is equal to 100 if grantee $g$ and firm $f$ comment on the same regulation in year $t$. The independent variable is equal to one if grantee $g$ received a donation from firm $f$ at year $t$ or $t-1$. Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** $p<0.01$, ** $p<0.05$, * $p<0.1$
Table 1.3: Commenting on regulations

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Grantee $g$ commented on regulation $r \times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>(1) 0.210*** (2) 0.157*** (3) 0.181*** (4) 0.122***</td>
</tr>
<tr>
<td></td>
<td>(0.012) (0.011) (0.010) (0.014)</td>
</tr>
<tr>
<td>Grantee $g$ received donation from any firm commenting on $r$</td>
<td></td>
</tr>
<tr>
<td>Fixed effects</td>
<td></td>
</tr>
<tr>
<td>Grantee</td>
<td>Y</td>
</tr>
<tr>
<td>Regulation</td>
<td>Y</td>
</tr>
<tr>
<td>SE Clusters</td>
<td>Grantee Grantee Regulation +Grantee Regulation</td>
</tr>
<tr>
<td>Observations</td>
<td>144,628,498 144,628,498 144,628,498 144,628,498</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is equal to 100 if grantee $g$ comments on regulation $r$. The independent variable is equal to one if grantee $g$ received in any year 2003-2016 a donation from a firm that commented on $r$. Standard errors are clustered at the level indicated in each column under “SE Clusters”: *** $p<0.01$, ** $p<0.05$, * $p<0.1$
Table 1.4: Hush money

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Grantee $g$ commented on regulation $r \times 100$ Mean</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DonorComment_{gr}$</td>
<td>0.086***</td>
<td>0.058***</td>
<td>0.042***</td>
<td>0.035***</td>
<td>0.024</td>
<td>-0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.007)</td>
<td>(0.009)</td>
<td>(0.008)</td>
<td>(0.017)</td>
<td>(0.010)</td>
<td></td>
</tr>
<tr>
<td>$DonorComment_{gr}$</td>
<td>0.150***</td>
<td>0.167***</td>
<td>0.150***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\times NumberComments_{gr}$</td>
<td></td>
<td>(0.027)</td>
<td>(0.009)</td>
<td>(0.028)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$DonorComment_{gr}$</td>
<td>2.560***</td>
<td>2.540***</td>
<td>2.517***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\times ShareComments_{gr}$</td>
<td></td>
<td>(0.149)</td>
<td>(0.185)</td>
<td>(0.232)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fixed effects
Grantee | Y | Y | Y | Y | Y |
Regulation | Y | Y | Y | Y |

SE Clusters
Grantee | Grantee | Regulation | Regulation | Grantee | + Regulation | Grantee | + Regulation |
Observations | 144,628,498 | 144,628,498 | 144,628,498 | 144,628,498 | 144,628,498 | 144,628,498 |

Notes: The dependent variable is equal to 100 if grantee $r$ comments on regulation $r$. The $DonorComment_{gr}$ is equal to one if grantee $g$ received in any year 2003-2016 a donation from firm that commented on rule $r$. Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** $p<0.01$, ** $p<0.05$, * $p<0.1$
Chapter 2

Text Similarity and Regulatory Influence

2.1 Introduction

Our findings on the link between corporate donations and co-commentary frequency by advocate non-profits point to potential corporate influence over non-profits in their regulatory feedback. In this chapter we extend our analysis of the relationship between comments and donations to the content of the comments and rule documents.

Incorporating text data allows us to examine whether, conditional on co-commentary frequency, the content of comment-pairs from firms and non-profits linked via charitable donations tend to be more similar, relative to other comments on the same proposed rules. Using established methods of natural language processing, we generate pairwise measures of textual similarity between any two firm-non-profit comments on a given rule. Co-comments by non-profits contain textual material that is more similar to comments by their corporate benefactors relative to other co-comment pairs and, importantly, the timing of this relationship parallels that of our first set of findings – co-comments in the year immediately following a donation are most similar. We also investigate the semantic orientation of the comments and show that the co-comment similarity for firm-grantee pairs does not result from comparably-worded comments that express opposing sentiments. Specifically, we find that co-comments by firm and grantees that are connected by an immediately preceding charitable donation do not express adversarial views on the same regulatory matter.

We also show that the co-commenting relationship matters for the final rules. Focusing on all comments made by corporations in our dataset, we show that, if a grantee (and particularly one receiving a recent donation) also commented on the proposed regulation, the language of the discussion of the final rule is more closely aligned with that of the corporation’s comments. This result survives the inclusion of both firm fixed effects and rule (docket) fixed effects, and also holds when we measure a firm’s influence based on whether it is cited by the regulators in their discussion of the final rule.

2.2 Quantifying the similarity in content across regulatory comments

In chapter 1, we explored non-profit’s propensity to comment on regulation. However, a crucial implication of our thesis that non-profits may act as strategic advocates for their corporate donors is that the content of the message delivered by non-profits to regulators may be affected by financial connections. In particular, upon receipt of (a) charitable grant(s), comments targeted to federal regulators by non-profits should be closer in content to the messages sent by their corporate benefactors (relative to the
counterfactual of no corporate donations). To provide evidence in this direction, we build a portfolio of circumstantial findings with the intent of discriminating among alternative theoretical mechanisms based on how well they match the empirical regularities that we present.

To build intuition (and without intent to claim any deliberate deception by the parties involved in this particular instance), consider the example of Bank of America’s donation of $150,000 to the Greenlining Institute in 2010. While Bank of America is the second largest bank in the United States by total assets and is a central player in housing finance in the country, the Greenlining Institute is a non-profit focused on improving access to affordable housing and credit to low-income families and minorities (African American, Asian American, and Latino, in particular). In 2011 both organizations commented on the Office of the Comptroller of the Currency’s Credit Risk Retention (CCR) docket, as part of one of the regulatory rulemaking streams initiated under the Dodd-Frank Act of 2010 (Title IX, Subtitle D, Section 941). CCR, also known as the “skin in the game” rule, imposed a 5 percent retention requirement on all mortgage loans originated by lenders in the United States to moderate “originate-to-distribute” moral hazard problems pervasive in the build-up to the 2008 financial crisis.

The main comment submitted by Bank of America remarks that, in relation to relaxing the definition of qualified mortgages exempted from retention requirements on the issuing bank’s balance sheet (i.e. of mortgages deemed safe enough not to warrant the restriction): “…the PCCRA provision will cause some borrowers to be unable to obtain a loan at all. In the currently tight private residential mortgage market, borrowers already must provide significant down payments.” The Greenlining Institute provides a similar assessment in its comment, suggesting that “by raising the barrier to affordable home ownership with an unreasonable 20% down payment requirement, we will not only keep families from rebuilding after foreclosure, but we will prohibit an entire generation of first time borrowers from owning a home, despite lower home prices across the country.” In sum, both organizations appear to advocate openly for laxer definitions of the CCR exemptions, limiting the rule’s bite, and allowing assets with substantially lower quality and higher risk to be exempt – an effort that ultimately succeeded in entirely defanging the rule.

In this section, we provide a framework for examining the content and textual similarity of comments filed by non-profits and firms, and show that, upon receipt of a donation from a firm’s foundation, comments by a non-profit are more similar to those of its donor, suggesting that the Bank of America-Greenlining example holds more broadly in the data.

We compute approximate measures of semantic similarity of pairs of public comments using Latent Semantic Analysis (LSA) with bag-of-words features. LSA is an established technique borrowed from the natural language processing (NLP) literature, and it has been shown to perform well on a variety of different document classification and retrieval tasks. LSA requires the conversion of text documents into vectors of word counts and applying term frequency–inverse document frequency feature extraction.

23 Docket ID OCC-2011-0002
24 Document ID OCC-2011-0002-0141
25 Document ID OCC-2011-0002-0353
27 See Dumais et al. (1988) and Deerwester et al. (1990). For a more recent discussion of latent semantic analysis, see Dumais (2004).
within each regulatory docket \( r \). Following this preparation phase, one can compute document-level singular vectors from a singular value decomposition of the text matrices and take the cosine similarity of any pair of document vectors. This approach provides a similarity score \( S_{fgr} \) normalized by the standard deviation in each docket \( r \) and distributed between -1 and 1 for every pair of texts formed by a comment by firm \( f \) and a comment by grantee \( g \) within a given docket. To further demonstrate the validity of our approach, we show in Appendix B that our measure performs well in a classification task of separating documents from different regulations and in clustering comments from similar organizations.

Using this comment-pair similarity score as the outcome, we consider a specification of the form:

\[
S_{fgr} = \beta_0 + \beta_1 D_{fgr} + \delta_f + \delta_g + \delta_r + \varepsilon_{fgr}
\]

where the coefficient of interest is \( \beta_1 \) and \( D_{fgr} \) is indicator variable that equals 1 if firm \( f \) donates to grantee \( g \), 0 otherwise. As the timing of such donations is a useful discriminant for interpretation of our findings, we will be careful in constructing \( D_{fgr} \) under different time horizons. The dataset we exploit for this analysis includes all possible firm-grantee pairs of comments conditional on commenting on a docket \( r \).

We begin by exploring the sign and magnitude of the estimated coefficient \( \beta_1 \) when the donation indicator variable takes the value of one in the event of any grant from \( f \) to \( g \) over our entire time period. Table 2.1 reports estimates for \( \beta_1 \) across a set of four specifications with an incremental inclusion of firm, grantee, and docket fixed effects. Coefficients are clustered at the docket, firm-grantee, or double clustered at both levels depending on the specification. The estimates of \( \beta_1 \), which capture the increase in units of standard deviations of similarity across comment pairs within each \( r \), range from 0.25 to 0.09 in the most restrictive specification (all significant at least at the 1 percent level). This indicates that pairs of comments made by firms and their grantees are more similar relative to a baseline similarity obtained by pairing comments at random within a docket.

As with our results on comment propensity in Section 1.4, the presence of a donation at any point in our sample period may proxy for some average similarity in the interests and beliefs of a firm and its grantee. Table 2.2 thus focuses on donations that take place in either the year in which the comments are filed (year \( t \)) or in the previous fiscal year (\( t - 1 \)). The point estimates are smaller in magnitude across comparable columns in Tables 2.2 and 2.1, but statistically indistinguishable. In separating explicitly contemporaneous donations and those made in the fiscal year immediately preceding the comments, as reported in Appendix Table A.4, we observe that precision and magnitude of the effect come from the donations made at time \( t - 1 \). The estimates, which capture the increase in units of standard deviations of similarity across comment pairs within each docket, range from 0.17 to 0.08 in the most restrictive specification.

Appendix Table A.5 addresses the concern that the timing of donations may be spuriously related to some underlying tendency of firms and grantees working in related areas of interests, by controlling in our most restrictive specifications also for North American Industry Classification System (NAICS) 6 sector code of the firm interacted with the IRS’s National Taxonomy of Exempt Entities Classification (NTEEC) code of the non-profit. As can be seen in the table, the estimated coefficient \( \beta_1 \) remains
precisely estimated and within the confidence intervals of our baseline estimates across specifications when accounting flexibly for such industry pair controls. Finally, notice that the reduction in sample size for this table results from missing sector information for some firm-grantee pairs, and that this sample shift also does not affect the point estimates relative to the baseline specifications. More precisely, we estimate a $\beta_1$ of 0.073 in column (4) of Table 2.2 and of 0.074 in column (1) of Appendix Table A.5, and a $\beta_1$ of 0.079 for $D_{fg}$ at time $t - 1$ in column (4) of Appendix Table A.4 and of 0.072 in column (3) of Appendix Table A.5.

We also present a placebo exercise that underscores the very specific timing of the link from donation to comment similarity. In particular, we modify our definition of donations to focus on the period immediately after the regulatory commenting phase. Appendix Table A.6 reports these results. As can be seen in the table, across specifications with incremental sets of fixed effects and industry controls, the estimated coefficient $\beta_1$ appears insignificant and smaller in magnitude relative to our base estimates.\(^2\)

This placebo exercise is informative along several dimensions. As the donation is close in time to the commentary activity, but statistically and economically insignificant, these findings further assuage the concern that our results may be spuriously driven by some underlying tendency of firms and grantees operating in related areas. The systematic timing of excess similarity between comments’ texts just following the disbursement of a charitable grant offer intuitive support to the logic of some form of suasion being exerted by the donor over the grantee.

As a final check, we investigate whether firm-grantee co-comments differ in their sentiment. We do so to assess the possibility that firms and grantees may employ a similar terminology while nonetheless delivering adversarial messages to regulators.

Our test is based on an analysis of comment sentiment, which relies on established NLP scholarship. Semantic orientation exercises are common in the NLP literature (e.g., the unsupervised classification of book reviews as positive or negative), including application to economics and finance, for example in the classification of monetary policy announcements as hawkish or dovish, in the study of the tone of financial news, or in partisan speech (Lucca and Trebbi, 2009; Tetlock, 2007; Tetlock et al., 2008; Gentzkow et al., 2016).\(^3\) Using these tools, our goal here is to rule out the possibility that the comments of non-profits receiving grants use similar words which express views that are nonetheless in opposition to their corporate benefactors. This specifically rules out the possibility that donations by firms may reach non-profits intervening on the same issues as the donor (and therefore using similar terminology), but expressing systematically antagonistic views.

Tables 2.3 and 2.4 maintain the same design and structure of fixed effects as Tables 2.1 and 2.2, but replace the similarity score $S_{fg}$ with a semantic orientation concurrence score $W_{fg}$ as our dependent variable. The construction of this variable relies on polarity scores defined for each comment based on the popular AFINN sentiment lexicon, with valence scores ranging between -5 (negative) and 5 (positive) for each labeled word. For each comment we construct the sum of valence scores divided by the number of words with non-zero valence scores. $W_{fg}$ is defined as the negative absolute difference

---

\(^2\)In the last column of the table, we also include donations at $t$ or $t - 1$, and show that only pre-comment donations matter, relative to donations at $t + 1$.

\(^3\)In general, by semantic orientation we refer to the direction (polarity) of words, phrases or longer pieces of text in a semantic space or context (e.g., friendly/adversarial, dovish/hawkish, positive/negative) calculated based on a reference lexicon of words or n-grams over which directionality is carefully labeled by a pool of researchers.
between this measure for the pair of comments from firm $f$ and from grantee $g$ on rule $r$. The interpretation of the coefficient of interest $\beta_1$ on $D_{fg}$ is therefore the effect of a charitable donation on the alignment of sentiment across firm and non-profit (i.e. the excess comovement of sentiment in the two comments relative to any randomly generated pair of firm and grantee comments on that rule).\[^{30}\]

The results in the tables do not support the view that donations reach grantees expressing opposing views to the firm providing the grant relative to a random grantee. If anything, the evidence points in the opposite direction: the coefficient $\beta_1$ is consistently positive in sign, indicating that firm-grantee comments are more aligned in sentiment. This relationship is significant in the specifications that link firms and non-profits by the existence of a donation at any point during our sample period (Table 2.3). The coefficient $\beta_1$ is positive, though significant in only one out of four specifications, for $fg$ pairs linked by donations at year $t-1$ or $t$ (Table 2.4). These findings are inconsistent with firm and grantee comments carrying antagonistic messages.

In sections 2.2 and 2.3 of the chapter we compare the content of firm comments with grantee comments and regulator discussion text. In the first case, our goal is to capture similarities between in the policies advocated for (or against) in by different commenters. In the second, it is to measure how much attention the regulator has paid to different comments. Complete solutions to these problems (in the sense of replicating what a literate and informed human could deduce from reading the text) are currently beyond the frontier of natural language processing (NLP) technology. Instead, we approximate these notions with a simple and robust method of text analysis called Latent Semantic Indexing (or sometimes Latent Semantic Analysis) with bag-of-words features. The basic recipe is as follows: After extracting and cleaning the comment text (to remove headers, page numbers, etc), each comment is converted into a vector of word counts. Very rare and very common words are dropped completely, and the remaining counts are weighted by a standard term-frequency-inverse-document-frequency (tf-idf) function to emphasis the words that are most useful in distinguishing between documents in each regulation. These weighted count vectors are then summarized by computing document-level singular vectors from a singular value decomposition of the feature-document matrix (this is the “latent” part of LSA, and generally improves the performance beyond using the raw feature vectors). Finally, the pairwise document similarity is computed as the cosine similarity between the document LSA vectors. The rest of this section explains these steps in greater detail, and describes a docket classification test we conducted to verify that the measure is informative.

### 2.2.1 Sample construction

We perform our analysis at the docket level. For each docket where at least one firm or one grantee comments, we load all organization comment text documents (initially treated as separate even if they are from the same author), and also discussion text from all linked rule documents. If there are at least three documents in total, we process the text and perform LSA to compute similarity measures.

Comment text is “cleaned” in by a Python script that attempts to identify and remove addresses and other header material that appear before the body text, tail sign-off and other material that appear

\[^{30}\text{In addition, a standardization within rule as for the variable } S_{fg} \text{ is employed for } W_{fg}, \text{ which allows to read coefficients in units of standard deviations of sentiment alignment across comment pairs within each } r.\]
after the body text, as well as repeated headers and footers (including page numbers) that appear on multiple pages. The script does not always succeed in removing the desired material (the comments are too varied in format to cover every possible case), but it is intended to remove some noise from the data.

Regulator discussion text is identified in the following way: First we load all rules that follow one or more comments in the docket (see appendix X on Federal Register document linking) and construct a separate discussion text document for each Federal Register rule document. We immediately drop Agency, Action, Dates, Summary, Addresses, Contact sections, as well as all appendices and tables of contents. Then we search for the strings “comment” and “letter” in all paragraphs and footnotes, and count a paragraph or footnote as discussion text if it appears under the same 2-level header as an instance of those strings. In other words, if the word “commenters” appears in the third paragraph under the heading “SUPPLEMENTARY INFORMATION: V. Discussion of Final Rule”, every paragraph and footnote located under that heading will be included.

2.2.2 LSA implementation

LSA is essentially the application of singular-value decomposition (SVD) to a document-feature matrix. We follow a standard approach in constructing this document-feature matrix from word counts, and use the Gensim\(^{31}\) python package for efficient implementation of these steps. First, each document is converted to lower case and words are stemmed (meaning removing common prefixes and suffixes, including pluralization so that “House” and “houses” both become “hous”). This step increasing the probability that closely-related words will be matched across documents. Next we identify every sequence of alphanumeric characters that are unbroken by white-space or other punctuation (except “.”) as a word and count the number of occurrences of each word in each document. We drop all words that appear in more than 70% or less than 20% of documents (this seemingly arbitrary step is important for good results with LSA and the numbers were chosen based on experiment in a docket classification test task). Finally, we re-weight the word counts in each document using term-frequency-inverse-document frequency (tf-idf) weighting with the following formula:

$$w_{ij} = f_{ij} \ln \left( \frac{D}{d_i} \right)$$

where \( f_{ij} \) is the count of word \( i \) in document \( j \), \( d_i \) is the number of documents containing word \( i \), \( D \) is the total number of documents in the docket. The matrix of \( w_{ij} \) entries then form a \((W \times D)\) feature-document matrix \( M \) (where \( W \) is the number of distinct words).

Recall that SVD decomposes the matrix \( M \) into the product of three matrices: \( M = U\Sigma V^* \) where \( U \) is \((W \times W)\) and \( V \) is \((D \times D)\). We use an algorithm\(^{32}\) that can compute the first \( k \) singular values and associated columns of \( U \) and \( V \). If \( k < \min(W,D) \) then the resulting decomposition forms a rank-\( k \) approximation of \( M \). The word “latent” in “Latent Semantic Analysis” refers to the idea that compressing the full feature-document matrix to a lower-dimensional approximation squeezes

\(^{31}\)https://radimrehurek.com/gensim/index.html
\(^{32}\)https://pypi.org/project/sparseevd/
synonyms into the same singular vectors and improves overall quality of the document model. In practice, researchers have found that values of $k$ around 200-400 appear work well in large samples of documents. However, $k$ is bounded above by the number of separate documents $D$, and we have many dockets with fewer than 300 comments. As a general solution, we choose $k$ according to the following formula:

$$k = \min(D - 2, 50)$$

So the LSA vectors have higher rank in large dockets, but we keep the maximum value a bit low so that the approximations are not wildly different in dockets of different sizes. Our object of interest is the resulting $(D \times k)$ matrix $V$. We describe each row as a document LSA vector.

### 2.2.3 Similarity measures

Once the document LSA vectors are computed, estimating the similarity between comments from firms and grantees is straightforward. We compute organization-level vectors by summing the LSA vectors for all documents associated with that organization, and define the pairwise comment similarity as cosine similarity of the organization-level vectors.

### 2.2.4 Rule similarity

Estimating the similarity between the rule discussion and an organization’s comment(s) is only slightly more complicated. In the case that there are multiple rules linked to a docket, we first construct all the comment-rule pairs and keep only those for which the comment was posted before the rule was published. Then we perform the same summing procedure to aggregate document LSA vectors associated with multiple sources of comment text submitted by the same organization, and compute similarity with the rule as the cosine similarity between the rule LSA vector and the organization-level vector.

### 2.3 Comment impact analysis: Evidence from final rule citations

While the preceding sections focus on the frequency and similarity of firm-grantee comments, we now turn to examining whether firms’ comments – and the similar comments made by grantees – comments have an impact on rulemaking. As it is typically very hard to assess the effects of advocacy on policy outcomes (and in general of informational lobbying on government policy choices), we will focus here on a newly devised approximation for such outcomes by asking how the final rule was shaped by the commentary. In particular, we aim to establish that when a firm comments on a rule, the published discussion of the rule by the regulator is closer in content similarity to the firm’s comments when the firm’s grantees also comment on that rule.

It is important to clarify that the final regulatory text itself is written with a terminology and structure that makes it very different from comments submitted or the explanation of the rule itself offered by the regulator in the preamble to the rule. The final regulatory text is designed to formulate,
amend, or repeal sections of the Code of Federal Regulations (5 U.S.C. § 551(5)). The discussion of the rule itself offers a justification and analysis of the regulator’s decision making process and intended scope or interpretation of the regulation. In fact, the discussion of the rule tends to be longer and reveals arguments in favor of or against specific choices that may have been brought forward by, for example, the comments from various entities, firms and grantees, in persuading the regulator. We therefore focus on this part of the final rule.

As an example consider the concern expressed by Wells Fargo, one the U.S. largest depository institutions, on a specific regulatory burden that appeared implied by the proposed rule version of the so called Volcker Rule of the Dodd-Frank Act of 2010. The Volcker Rule aimed at prohibiting depository institutions from engaging in the use of part of its depository funding for speculative trading (proprietary trading). Wells Fargo expresses concern that the proposal requires transaction-by-transaction oversight: “We also do not believe that the Proposed Rule’s transaction-by-transaction approach, which would require analyzing permitted customer trading, market making, underwriting and hedging activities on a transaction-by-transaction basis, is the best way for the Agencies to implement the Proposed Rule...” The OCC addresses this concern directly and concedes some changes to the rule: “A number of commenters expressed general concern that the proposed underwriting exemption’s references to a ’purchase or sale of a covered financial position’ could be interpreted to require compliance with the proposed rule on a transaction-by-transaction basis. These commenters indicated that such an approach would be overly burdensome. .... A general focus on analyzing the overall ’financial exposure’ and ’market-maker inventory’ held by any given trading desk rather than a transaction-by-transaction analysis.” Importantly, also the Black Economic Council, a recent Wells Fargo grantee, is found to express concerns on the same rule on grounds of excessive complexity.

We begin by defining \( S_{fr} \) the similarity score between the discussion of docket \( r \) and firm \( f \)'s comment. In contrast to the score constructed in Section 2.2, \( S_{fr} \) measures the similarity between a comment and the discussion of the rule in a docket, rather than the similarity between the texts of two comments on a rule. \( S_{fr} \) is designed as a proxy for the salience and effectiveness of the firm’s comment in shaping the regulator’s decisions. As with the previous similarity measure \( S_{fgr} \), we normalize \( S_{fr} \) by the standard deviation in each docket \( r \), so that \( S_{fr} \) is distributed between -1 and 1 for every pair of texts.

Dropping time subscripts, let us posit \( S_{fr} \) as function of the commenting effort of the firm and of

---

33The discussion of the rule is found in the Supplementary Information section, which is part of the preamble to the final rule and typically constitutes its most important component. See https://www.federalregister.gov/uploads/2011/01/the_rulemaking_process.pdf
34Docket ID OCC-2011-0014
36Document ID OCC-2011-0014-0024
37As in some cases multiple rules may be included in a docket by regulators (including amendments, notices, etc.) and each regulatory stream can be linked to a final rule, our approach here is to take for each firm and docket the closest in similarity to the firm’s comment vector. This is meant to more accurately represent the dimension of the docket the firm more closely commented about. Our results are similar when removing the lowest similarity score within a docket-firm group and then taking the mean similarity or when keeping only dockets with exactly one rule document. See Online Appendix for these robustness checks.
grantees connected to the firm by donation:

\[
S_{fr} = \beta_1 \sum_g CC_{frg} \times D_{fg} + \beta_2 \sum_g D_{fg} + \beta_3 \sum_g CC_{frg} + \delta_f + \delta_r + \varepsilon_{fr}
\]

Focusing on the extensive margin of commenting behavior, we can replace all sums with indicator functions and also include firm and docket fixed effects:

\[
S_{fr} = \beta_1 I \left( \sum_g CC_{frg} \times D_{fg} > 0 \right) + \delta_f + \beta_2 I \left( \sum_g D_{fg} > 0 \right) + \beta_3 I \left( \sum_g CC_{frg} > 0 \right) + \varepsilon_{fr}
\]  

(2.1)

The variable of interest is \(I \left( \sum_g CC_{frg} \times D_{fg} > 0 \right)\), which is equal to 1 if we observe a donation by the firm to a grantee co-commenting on the same rule, and 0 otherwise. If there is excess similarity between rule discussion and a firm’s comment when grantees connected to the firm by donation also comment on that rule, we expect \(\beta_1\) to be positive. As we established in the previous two sections, such comments by non-profits occur around the time of firm donations and appear to exhibit a systematically higher textual similarity to the comments filed by the grantee’s benefactors. Here, we aim to establish that corporate benefactors appear to gain in terms of \(S_{fr}\), a proxy that at a minimum captures having the attention of the regulator, but could conceivably correlate with influence in shaping the final rule text or keeping certain provisions out.

Let us also clarify that in specification (2.1) the coefficient on the term \(I \left( \sum_g D_{fg} > 0 \right)\) cannot be separately identified from a firm \(f\) fixed effect, since it counts whether the firm ever donates to any grantee. Also the coefficient on the term \(I \left( \sum_g CC_{frg} > 0 \right)\) cannot be separately identified from a docket fixed effect, as it counts the average level of commenting by grantees for that rule (only firms commenting on the rule are included in the estimation and all grantees commenting on \(r\) are, by default, co-commenters of every firm also commenting on \(r\)). As \(\beta_2\) and \(\beta_3\) allow us to measure the direct effects of each element to the main interaction term \(I \left( \sum_g CC_{frg} \times D_{fg} > 0 \right)\), we include firm and docket fixed effects in our key specifications. We also experiment by removing each set (or both) in order to estimate these direct effects.

As in Section 2.2, we begin by exploring the sign and magnitude of coefficient \(\beta_1\) when the donation indicator variable takes value 1 if there is any grant from \(f\) to \(g\) over our entire time period, and 0 otherwise. Table 2.5 reports estimates for \(\beta_1\) across a set of five specifications with an incremental inclusion of firm and docket fixed effects for specification (2.1) in columns (1) to (4) and a specification with the continuous variable \(\sum_g CC_{frg} \times D_{fg}\) in column (5). Coefficients are clustered by firm or docket, or double clustered at both levels depending on the specification. In columns (1) to (4), the increments expressed in terms of increases in units of standard deviation of similarity within each docket range from 4.5 to 23.7 percentage points, indicating that comments made by firms on rules that also received comments from their grantees appear closer in content to the final rule discussion.

As the presence of any donation over time is a less accurate indicator of a direct connection between firms and grantees than recent donations, Table 2.6 looks at donations that take place in either the year in which the comment is filed (year \(t\)) or in the previous fiscal year \((t-1)\). In this specification, the
point estimates of \( \beta_1 \) appear more precise and quantitatively sizable, with 0.173 of a standard deviation higher similarity for comments filed by firms with co-commenting grantees who were recipients of their donations in our preferred column (4). Similar results are obtained focusing on the intensive margin, as reported in column (5).

Appendix Table A.7 further probes our results on rule-comment similarity by adding controls for the log number of pages of commentary filed on \( r \) by \( f \) which, even controlling for firm and docket fixed effects, turns out to be a strong predictor of similarity between rule discussion and comment by the firm. The effect of this control is intuitive, in the sense that carefully articulated comments may capture more of the attention of the regulator and translate in higher \( S_{fr} \). The coefficient on \( I \left( \sum_g CC_{frg} \times D_{fg} > 0 \right) \) based on donations at \( t \) or \( t - 1 \) remains positive and statistically significant in all specifications in Appendix Table A.7. Contrasting these estimates with those based on the same variable constructed with donations at any time, included in columns (2) to (4), shows that the increase in similarity is driven by the co-commenting of a grantee that received a donation in the current or previous year, i.e., recent donations. When both variables (constructed with recent donations versus donations at any point in time) are included in columns (3) and (4), it is evident that recent donations carry the relevant variation.\(^{38}\)

Importantly, the content of the messages simultaneously communicated by non-profits and by corporations appears systematically closer in terms of textual and semantic similarity in presence of a charitable contribution provided immediately before those comments are filed. While circumstantial, the evidence seems to point to potential concerns in the assessment of prima facie independent information on the part of targeted regulators, who may be unaware of the philanthropic grants that realize in the backdrop and may interpret similar comments stemming from different segment of the public spectrum as indicative of merit.

### 2.4 Concluding remarks

In this chapter we use text analysis to flesh out the relationship between charitable donations from corporations, and the commenting behavior of non-profits in federal rulemaking. First, we compare comments submitted by corporations and non-profits according to the similarity of their text, and find that non-profits receiving donations submit comments that are unusually similar in content to those of their benefactors. This result strengthens the interpretation that firms are able to use donations to amplify their messages to the regulators. Second, we use textual similarity between the commenting firm and final rule discussion to gauge influence of comments over policymakers. It appears that the co-commenting patterns of firms and non-profits can offer additional visibility to the messages sent by the firms themselves measured in terms of comment similarity to the final rule or even likelihood.

\(^{38}\)In online Appendix table A.8 we also replaced similarity to the final rule discussion with indicators or log 1+ counts of the number of times that a firm is cited in the final rule discussion. We obtain similar qualitative results as in the analysis in this section. Specifically, when focusing on an indicator variable for being cited or not for a firm, our results indicate a positive but imprecise relationship when controlling for docket and firm fixed effects, but when focusing on number of times the firm is cited, the presence of recent donations to co-commenting non-profits is positive, significant at standard confidence levels, and robust to firm and docket fixed effects, and controlling for log pages of comments submitted and the any donations over time.
of citation of a donor firm. As rates of return for political influence activities are extremely complex
to measure, this is an area of statistical investigation requiring further study. Its exploration remains
open to future empirical research. The availability of a large set of public comments by non-profits
and by corporations on a diverse set of rules and regulations, ranging from banking to environmental
regulation, makes for a rich and virtually untapped empirical environment.

2.5 Tables

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Similarity of comments by grantee $g$ and firm $f$ on same regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Grantee $g$ received donation from firm $f$</td>
<td>0.249***</td>
</tr>
<tr>
<td>Fixed Effects</td>
<td>Docket Y</td>
</tr>
<tr>
<td></td>
<td>Firm Y</td>
</tr>
<tr>
<td></td>
<td>Grantee</td>
</tr>
<tr>
<td>SE Clusters</td>
<td>Docket</td>
</tr>
<tr>
<td>Observations</td>
<td>301,602</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is a similarity index between the comment of firm $f$ and the comment
of grantee $g$ on regulation $r$, divided by the standard deviation of similarity of all comments relative to
$r$. The independent variable is equal to one if grantee $g$ received a donation from firm $f$ between 2003
and 2016. Standard errors are clustered at the level indicated in each column under “SE Clusters”. ***
p<0.01, ** p<0.05, * p<0.1
<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Similarity of comments by grantee $g$ and firm $f$ on same regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Grantee $g$ received donation from firm $f$ at $t$ or $t-1$</td>
<td>0.172***</td>
</tr>
<tr>
<td></td>
<td>(0.058)</td>
</tr>
</tbody>
</table>

Fixed Effects
- Docket: Y
- Firm: Y
- Grantee: Y

SE Clusters
- Docket: Docket
- Firm + Grantee: Docket
- Firm + Grantee + Docket: Docket

Observations
- 301,602
- 301,602
- 300,817
- 300,792

Notes: The dependent variable is a similarity index between the comment of firm $f$ and the comment of grantee $g$ on regulation $r$, divided by the standard deviation of similarity of all comments relative to $r$. The independent variable is equal to one if grantee $g$ received a donation from firm $f$ in the year when the comment appears or the year before. Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** $p<0.01$, ** $p<0.05$, * $p<0.1$
Table 2.3: Sentiment - Any donation

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Sentiment similarity of comments by grantee $g$ and firm $f$ on same regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Grantee $g$ received</td>
<td>0.030*</td>
</tr>
<tr>
<td>donation from firm $f$</td>
<td>(0.017)</td>
</tr>
</tbody>
</table>

Fixed Effects

Docket: Y
Firm: Y
Grantee: Y

SE Clusters

Docket
Docket
Firm + Grantee
Firm + Grantee + Docket

Observations

309,033
308,576
308,184
307,719

Notes: The dependent variable is the difference between the sentiment score assigned to the comment of firm $f$ and the comment of grantee $g$ on regulation $r$ as described in Section 2.2, divided by the standard deviation of this measure within rule $r$. The independent variable is equal to one if grantee $g$ received a donation from firm $f$ between 2003 and 2016. Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** $p<0.01$, ** $p<0.05$, * $p<0.1$
Table 2.4: Sentiment - Recent donation

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Sentiment similarity of comments by grantee $g$ and firm $f$ on same regulation</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grantee $g$ received donation from firm $f$ at $t$ or $t-1$</td>
<td></td>
<td>0.027</td>
<td>0.032***</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>Fixed Effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Docket</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firm</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grantee</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE Clusters</td>
<td>Docket, Docket, Firm + Grantee, Firm + Grantee + Docket</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>309,033, 308,576, 308,184, 307,719</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: The dependent variable is the difference between the sentiment score assigned to the comment of firm $f$ and the comment of grantee $g$ on regulation $r$ as described in Section 2.2, divided by the standard deviation of this measure within rule $r$. The independent variable is equal to one if grantee $g$ received a donation from firm $f$ in the year when the comment appears or the year before. Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** p<0.01, ** p<0.05, * p<0.1
Table 2.5: Rule-comment similarity - Any donation

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Similarity of rule discussion and comment by firm ( f ) on same regulation ( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>At least one grantee</td>
<td>0.237***</td>
</tr>
<tr>
<td>co-commenting and receiving donation from firm ( f ) in any year</td>
<td>(0.065)</td>
</tr>
<tr>
<td>Log number of grantees</td>
<td>0.055</td>
</tr>
<tr>
<td>co-commenting and receiving donation from firm ( f ) in any year</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Fixed Effects</td>
<td></td>
</tr>
<tr>
<td>Docket</td>
<td>Y</td>
</tr>
<tr>
<td>Firm</td>
<td>Y</td>
</tr>
<tr>
<td>SE Clusters</td>
<td>Docket</td>
</tr>
<tr>
<td>Observations</td>
<td>5,538</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is a similarity index between the comment of firm \( f \) and the discussion of regulation \( r \), divided by the standard deviation of similarity of all comments relative to \( r \) and discussion of regulation \( r \). The independent variable is equal to one if there is at least one grantee \( g \) co-commenting on regulation \( r \) and receiving a grant from firm \( f \) in any year. Column 5 reports the coefficient on the logarithm of one plus the number of such grantees. Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** \( p<0.01 \), ** \( p<0.05 \), * \( p<0.1 \)
Table 2.6: Rule-comment similarity - Recent donation

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Similarity of rule discussion and comment by firm $f$ on same regulation $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>At least one grantee $g$ co-commenting and receiving donation from firm $f$ in year $t$ or $t-1$</td>
<td>$0.274^{***}$</td>
</tr>
<tr>
<td>Log number of grantees co-commenting and receiving donation from firm $f$ in year $t$ or $t-1$</td>
<td>$0.112^{**}$</td>
</tr>
</tbody>
</table>

Fixed Effects

<table>
<thead>
<tr>
<th></th>
<th>Docket</th>
<th>Firm</th>
<th>SE Clusters</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Docket</td>
<td>Y</td>
<td>Y</td>
<td>Docket</td>
<td>5,538</td>
</tr>
<tr>
<td>Firm</td>
<td></td>
<td></td>
<td>Firm+Graante</td>
<td>5,145</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Firm+Docket</td>
<td>5,367</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Firm+Docket</td>
<td>4,965</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4,965</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is a similarity index between the comment of firm $f$ and the discussion of regulation $r$, divided by the standard deviation of similarity of all comments relative to $r$ and discussion of regulation $r$. The independent variable is equal to one if there is at least one grantee $g$ co-commenting on regulation $r$ and receiving a grant from firm $f$ in year $t$ or $t-1$. Column 5 reports the coefficient on the logarithm of one plus the number of such grantees. Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** $p<0.01$, ** $p<0.05$, * $p<0.1$
Chapter 3

Estimating Influence from Bag-of-words
Text Features

3.1 Introduction

One of the most important research areas in Political Economy is the study of how special interest
groups influence government policymaking. Quantifying the degree of influence that firms and other or-
ganizations have on policy is a necessary prerequisite for exploring many important research questions.
However, quantifying influence is generally very difficult because data on interactions between interest
groups and government rarely include detailed information about messages conveyed by interest groups
to government, or policy outcomes chosen by the government after these messages are received.

Digitization of text documents is opening up a new frontier in the study of influence. Many actions
by governments, courts, interest groups, firms, and individuals are recorded as text documents, and
are increasingly available for download in bulk. In some cases, such as in the context of U.S. federal
rulemaking explored in chapters 1 and 2, it is possible to obtain text documents describing the advocacy
of interest groups and the actions taken by policymakers. This type of data can be used to produce
new estimates of influence. However, the task is challenging, and no existing approach is completely
suitable in this context. In this chapter I present a novel approach to estimating the influence that
overcomes several key challenges with analyzing influence in the rulemaking.

My new approach bridges a large family of common text analysis techniques with a simple model of
influence that allows me to derive linear estimators of the influence of each comment on the final policy.
It has several advantages. First, the process is completely automated, as opposed to hand-coding some
or all of the data. This makes feasible to analyze the thousands of rules and millions of comments
that can be obtained from rulemaking data. Second, the estimators make no assumptions about the
dimensionality of the topic or policy space, which can be very complicated in regulatory discussions.
Third, the estimators can be extended to operate in panel data settings where the same interest groups
comment on multiple different rules. Panel data offers increased precision by aggregating estimates
from across rules, and more importantly, it allows for the possibility of controlling for the author’s
writing style which could otherwise bias estimates. I demonstrate how this can be done with my new
estimators using the appropriate fixed effects.

These advantages are best understood in the context of the existing techniques for measuring
influence. There are three competing approaches (broadly defined). The first is hand coding. In
principle, it is possible for a researcher to read the text describing each interest group’s advocacy and
code the group’s preferences on each issue they raise. Then, the researcher could also read the text
describing the policymaker’s decision and determine which groups obtained their preferred outcomes. Yackee (2005) performs such an analysis on 1,444 comments submitted to several regulators regarding 40 randomly selected proposed rules. The effort involved in constructing the data for this paper must have been substantial. However, Yackee still faced constraints on how many comments and rules could be reviewed and to what degree of detail. Yackee coded the top five issues mentioned by commenters for each rule, but was unable to cover all issues raised by commenters. When coding the regulator response, she simplified all debates to a single more-or-less regulation dimension. These choices are sensible given the finite budget researchers face when hand-coding data. But the approach rules out estimating commenter-level influence, potentially discards many influential arguments that were only offered by a small number of commenters, and aggregates policy outcomes that may be viewed very differently by the actors involved. Yackee was also only able to analyze a tiny fraction of the proposed rules published during her study period. To the extent that we are interested in rare events or heterogeneity across firms, industries and agencies, manual coding leaves a lot on the table.

A second approach is to automatically classify texts describing interest group positions along one or more policy dimensions, and then relate these positions to policy outcomes. Klüver (2009) experiments with two algorithms for placing comments from interest groups into a space of policy preferences and compares them to manual coding. Unfortunately, this type of coding only solves half the problem. After classifying interest group preferences, it is still necessary to compare these preferences with the final outcome. While one could imagine also automatically classifying outcomes in a compatible way, in all the examples I am aware of this second part is done manually. In contexts where the number of outcomes is small, this part of the manual coding might be relatively easy. But federal rules are complex and multidimensional, and it can be laborious to characterize the outcome for even a single rule, much less thousands. The Wordfish algorithm explored in Klüver (2009) also assumes that preferences are only differentiated along a single dimension. This single-dimension assumption makes sense in contexts like legislative voting in a two-party system, but the interest groups that participate in the notice and comment process are very diverse and I am skeptical that their preferences are well approximated by positions on a line.

A third point of comparison is the pairwise similarity approach introduced in chapter 2. This approach is built on top of Latent Semantic Analysis, a a simple type of topic model. Topic models identify commonalities across documents in terms of topics—groups of words that frequently occur in the same documents—and represent each document as weighted mixture of these topics. This structure makes it straightforward to define and compute measures of similarity between documents that describe how much overlap two documents have in the topics they cover. We use the similarity between comments and the comment discussion as a measure of attention paid to each commenter. The primary difference between our pairwise similarity approach and the other automatic coding algorithms described above is that we do not attempt to place comments and rules in a policy space. Instead we skip this intermediate step and simply compare comments directly to rules. Our argument is that rule discussions that contain many of the same words as a comment are likely to be summarizing or responding to that comment, and that this similarity therefore gives us a measure of regulator attention for each comment. One advantage is that no manual coding is required and we can compute
the similarity between every comment and every rule in the sample with only a few hours of computer processing. Another advantage is that we don’t have to assume that the policy debate around each rule is single-dimensional, or that the dimensions are the same across rules. Pairwise measures therefore require fewer assumptions about the policies and preferences of commenters. The similarity approach therefore overcomes many of the substantial problems with previous approaches. However, it still has drawbacks. One is that the similarity measures produced by topic models are hard to interpret in terms of influence. Another is that LSA provides no notion of precision, so we cannot distinguish between very confident and very uncertain pairwise estimates, or test whether differences between estimates are statistically significant. Finally, the pairwise measures fail to solve the problem of what I think of as “style bias”.

One of the central challenges of working with natural language is that there are many ways to represent the same idea. One way to think about these different texts with similar meanings is to distinguish between the “style” of a text, and the “substance”—that is, the underlying meaning. Ideally, when we measure influence we would like to detect the similarity in the fundamental ideas, arguments, or policies described in different documents. When we accidentally detect similarity in writing style instead, we mis-measure influence. In particular, we are likely to over-estimate the influence of commenters who writing in the same formal, jargon-filled writing style as the regulator, and underestimate the influence of commenters who use more casual language. In section 3.4 I run Monte-Carlo tests which suggest that this type of style bias is a substantial problem when estimating influence. Unfortunately, topic modeling does not provide any tools to differentiate between similarity in writing style and more fundamental similarity of meaning. However, with panel data (that is, multiple comments from each author), it is possible to differentiate between the average writing style of each commenter and the content specific to each rule. The logic behind this decomposition is similar to that used in other economic settings where we want to distinguish between fixed and varying components of panel data: the linearity of the estimators allows us to use high-dimensional fixed effects to control for each author’s average comment. This is another important justification for focusing on linear models beyond their computational tractability, as the same procedure would likely lead to biased estimates in non-linear models due to the incidental parameters problem.

All estimators described above, including my new estimators, use a “bag-of-words” approach to representing text. “Bag-of-words” refers to the way that documents are represented entirely by the number of times each word (or other discrete feature) occurs in the text. All document structure and contextual information is stripped away, and each document is reduced to a simple vector of word counts. The advantage of this approach is that many operations on text can be implemented using linear algebra. It is also easier to write down a process for generating new text documents, since since the only thing that needs to be defined is the distribution of word frequencies.

The alternative approach is to look at sequences of text. Cutting-edge natural language processing uses deep neural networks to model sequences of text and combine information from many words and the way words relate to each other. For example, Devlin et al. (2018) construct a model that reads complete sentences and can be used to improve performance on a suite of natural language processing tasks. There are two main reasons why I do not attempt to use these approaches. First, the models are
very computationally expensive to train and run. Second, the outputs are much harder to interpret. Sequence modeling techniques generate context dependent representations of text that come from a very complex algorithm. The data generating process is an unpredictable black box, and we end up back in the position of having measures that don’t have a clearly defined scale or meaning. I’m optimistic that there are long-term solutions to this problem, but developing them is beyond the scope of this chapter.

I take advantage of a technology known as “word vectors” or “word embeddings” which exists somewhere between bag-of-words and full sequence modeling. Word vectors aim to overcome some of the limitations of using words alone to represent text. Even in two long documents discussing the same topic, it will often be the case that the most important words are only used a small number of times, and the set of key words used in each document may not overlap. For example, one document may use the words “house”, “housing”, and “homes”, while another document discusses “houses”, “dwellings”, and “apartments”. Clearly both documents discuss houses, but the token frequency vectors would be orthogonal. This problem is referred to as “synonymy” in the text analysis literature. Synonymy can be alleviated to some degree by simplifying or “stemming” words when tokenizing. For example, “house”, “housing”, and “houses” can all be stemmed to “hous”. But the problem of linking “house” to “apartment” cannot be solved this way. A popular solution to this approach is to instead represent each word by a real valued vector, where the vectors are generated by an algorithm that uses the way words co-occur in text to infer which words are similar and give them similar vectors (Mikolov et al., 2013; Pennington et al., 2014; Joulin et al., 2016; Bojanowski et al., 2016). Each word has a context independent representation, but that representation summarizes contextual information learned from the text. Thus, “house” and “apartment” are located nearby in vector space because they often appear near the same words. The estimators I propose use a general set of word vectors that could come from any algorithm. I show in Monte Carlo tests that vectors generated with LSA improve estimator performance relative to random vectors.

The chapter is structured as follows: First, I present a very simple model of textual influence. The purpose of this model is to provide a clear interpretation for the coefficients representing influence, and to provide an explicit data generating process to motivate the estimators. Next, I derive estimators based on cross-sectional data (one comment per author and a single rule) and present evidence of their effectiveness from Monte Carlo tests where synthetic rules are generated from real comments. Finally, I extend the model of influence to account for commenter and regulator writing style and derive panel versions of each estimator that can control for writing style. I run Monte Carlo tests to evaluate how well each estimator performs when writing style is an important factor in the data generating process and find that including fixed effects is very important when rules are generated using commenter and author writing styles.
3.2 Modeling textual influence

3.2.1 Bag-of-words influence

The model takes a common approach to describing text. There are $D$ input documents $i = 1, 2, ..., D$. These documents have been broken down into unique “tokens” (which could be n-grams, words, short phrases, or any other discrete features). There are $T$ unique tokens across all the documents, denoted $t = 1, 2, ..., T$. Each document is represented as a column vector of token counts $c_d = \begin{bmatrix} c_{d1} & c_{d2} & c_{d3} & \cdots & c_{dT} \end{bmatrix}^T$ where $c_{dt}$ is the number of times token $t$ occurs in document $d$. The total number of tokens in each document is denoted $n_d$. Each document has a corresponding token-frequency vector that is equal to the token count vector divided by the number of tokens in the document,

$$f_d := \frac{c_d}{n_d}$$

I construct the matrix $F$ by stacking the document token-frequency vectors horizontally:

$$F_{(T \times D)} := \begin{bmatrix} f_1 & f_2 & f_3 & \cdots & f_D \end{bmatrix}$$

I model influence in the following way. After receiving the comments, the regulator generates a rule $r$ from a weighted mixture of the input documents. Specifically, the regulator draws $n_r$ tokens from the input documents according to the following sampling procedure: For each token, the regulator independently selects an input document $d$ with probability $\beta_d$, and then copies one random token from that document into the rule. I interpret the vector $\beta_d$ as the degree of influence document $d$ had on the final rule.

Under the procedure described above, the token counts in the rule are a random variable with multinomial distribution

$$c_r \sim \text{mult}(n_r, \pi)$$

where,

$$\pi = \sum_{d=1}^{D} \beta_d c_d / n_d = F \beta$$

Overall, this is a very simplistic model of both text generation and influence. The regulator simply accepts whatever comments are submitted and then mixes them together to generate the rule. However, I will address several of the most obvious weaknesses throughout the chapter. First, I will show that we can relax the assumption that the regulator copies words exactly from the comments to the rule, and instead allow for a regulator that paraphrases text from the comments with their own words. Second, I will show that we can extend the model to represent a regulator who mixes in their own writing style and filters out the writing style of other authors. These two modifications make the model much more realistic in terms of how we might actually expect a regulator to write a rule using comments as input, while maintaining the simplicity and robustness of the linear estimator framework. Nevertheless, the
model leaves several questions unanswered. One loose end is how the regulator chooses the number of tokens in the rule. I see this as a minor point, since the estimates are not directly affected by this parameter once everything is converted to token frequencies (though longer documents do result in more precise estimates because of reduced sampling error). A more important consideration is where the influence parameters come from and to what extent we can think of the regulator as a rational decision-maker. Ideally we would be able to map the influence parameters to the equilibrium strategies of some game theoretic communication game played by the commenters and regulator and then interpret the estimators as a reduced form. At the moment however, I am agnostic about what game is actually being played, and I think there is value in first building some descriptive statistics that might (if interpreted carefully) point towards certain communication models to focus on. The simple $\pi = F\beta$ model therefore serves as a foundation to build on, both in terms of adding linguistic richness, and also game theoretic considerations in the future.

### 3.3 Estimating influence

#### 3.3.1 Frequency-based estimators

**Set up**

Given the setting described above it is possible to estimate the measures of influence $\beta$ from the observed token-frequency vectors $f_1, f_2, ..., f_D$ and the observed final rule document token frequencies using a linear estimator.

First, note that from the properties of the multinomial distribution,

$$E[c_r|\beta] = n_r\pi = n_r F\beta$$

$$\text{Var}(c_r|\beta) = n_r(diag(\pi) - \pi\pi^T) = n_r(diag(F\beta) - (F\beta)(F\beta)^T)$$

If we define the empirically observed vector of token frequencies in the rule as

$$g := c_r/n_r$$

and the difference between the realized and expected value of $g$ as

$$\varepsilon := g - E[f_r|\beta] = g - F\beta$$

then we can write the observed rule token frequencies $g$ in the familiar form:

$$g = F\beta + \varepsilon$$
OLS estimator

The last equation suggests that we can estimate $\beta$ with a least-squares estimator, the simplest being OLS:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \|g - F\beta\|^2$$

However, $\varepsilon$ is not i.i.d. and instead has a correlation matrix that depends on $\beta$. OLS will be consistent but not efficient.

FGLS estimator

The alternative is to construct an iterative FGLS estimator. Let $\Sigma(\beta)$ be the variance-covariance matrix of $\varepsilon$. Then,

$$\hat{\beta}_{FGLS} = \arg\min_{\beta} \left\| \Sigma^{-1/2}(g - F\beta) \right\|^2$$

$$\Sigma(\beta) := Var(g|F,\beta) = \frac{1}{n_r}(\text{diag}(F\beta) - (F\beta)(F\beta)^T) $$

where we can use OLS to obtain the first estimate of $\beta$. FGLS sometimes performs poorly in finite samples, so it is not clear which estimator will perform better in practice.

Identification conditions

The rank condition for OLS is $\text{Rank}(F) = D$, implying two identification conditions:

1. (a) The input documents must have unique token frequencies
   (b) The number of unique tokens must be larger than the number of documents ($T \geq D$)

3.3.2 Token and Document vectors

Motivation

Estimating influence directly from token frequencies is probably not a good idea in practice, for two reasons. First, the FGLS estimator is difficult to compute. $\Sigma(\beta)$ has dimensions $(T \times T)$, where $T$ is often on the order of hundreds of thousands of unique tokens. This puts the memory requirements for storing this matrix somewhere on the order of hundreds of gigabytes.

The second reason relates to the problem of synonymy raised in the introduction. Suppose that when the regulator is copying tokens from the comments into the rule, they replace each word with a synonym or closely related word. Then the similarity of the token frequencies between comment and rule could be very low even when influence is high. The frequency-based estimators will be biased downwards. This seems like a realistic departure from the original model and something worthwhile to address.

Employing pre-trained token vectors can solve both problems simultaneously. Suppose we have a set of pre-trained token vectors so that each token $t$ has a vector representation $v_t \in \mathbb{R}^k$. Then we
can represent documents as weighted sums of the token vectors rather than the raw token frequencies. Best practices in the text analysis literature generally recommend token vector sizes with $k$ in the low hundreds. Thus, the first benefit of using token vectors is to reduce the size of $\Sigma(\beta)$ from $(T \times T)$ to a much more smaller $(k \times k)$ matrix. The second benefit is difficult to show analytically, but at least as important. Each token vector algorithm will generate different token vectors, but correlations between the occurrences of different tokens will be captured in the values of the vectors, so that tokens that occur in similar contexts (i.e., “house” and “apartment”) have similar vectors. It follows that documents will also be assigned similar vectors when they discuss similar topics, even if they do not use exactly the same words. The resulting estimator will be much more robust to changes in specific word choice while still detecting similarity of meaning and topic.

**Set up**

I assume that each token has an associated column vector $v_t \in \mathbb{R}^k$. The full set of token vectors is represented by the matrix $V$:

$$V_{(k \times T)} := \begin{bmatrix} v_1 & v_2 & v_3 & \cdots & v_T \end{bmatrix}$$

Each document $d$ is represented by a document vector $x_d$, constructed by taking the sum of all token vectors, weighted by the frequency of their occurrence in the document.

$$x_d := \sum_{t=1}^{T} f_{dt} v_{dt} = Vf_d$$

When we stack the input document vectors to form

$$X_{(k \times D)} := \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_D \end{bmatrix}$$

and we can write the vector transformation succinctly as $X := VF$.

I define the final rule document vector similarly, as

$$y := Vg$$

**Token Vector Estimators**

We can derive vector OLS (VOLS) and vector FGLS (VFGLS) estimators from those in the previous section by pre-multiplying $g = F\beta + \varepsilon$ by $V$. This gives the base equation,

$$y = X\beta + \omega$$

where $\omega := V\varepsilon$ is the difference between expected and observed document vectors. Then the VOLS estimator is:
\( \hat{\beta}_{VOLS} = \arg \min_{\beta} \| y - X\beta \|^2 \)

And the VFGLS estimator is defined by the equations:

\[
\hat{\beta}_{VFGLS} = \arg \min_{\beta} \left\| \Omega^{-1/2} (y - X\beta) \right\|^2
\]

\[\Omega = V\Sigma V^T\]

The new rank condition \( \text{Rank}(X) = D \) requires \( \text{Rank}(V) \geq D \), and implies two additional identification conditions:

2. (a) There must be at least \( D \) unique token vectors
   
   (b) The token vectors must have dimension \( k \geq D \)

These estimators are computationally feasible because it is possible to compute \( \Omega \) in parts without ever computing the full \( \Sigma \) matrix.

**Token weights**

Another lesson that can be learned from best-practices in text analysis is that not all tokens are equal. Most tokens are not very informative about the topic. For example, the results should not be driven by how often input documents use the words “the”, “and”, or “this”. In our simplistic model of influence, these are equally informative about which documents have been mixed into the rule. But in practice, we would like to de-emphasize uninformative words and emphasize the importance of matching important key words.

LSA addresses this problem by computing a separate token weight for each token that reflects how informative it is in distinguishing between documents. Document vectors are then constructed as *weighted* means of token vectors, which are often further normalized to unit length before comparison. By analogy, it seems intuitive to apply a similar weighting scheme when constructing document vectors. For example, one could scale each vector by a pre-determined weight that relates to the frequency of the token. Scaling the vectors does not affect the derivation of the estimators, so at first glance this appears to be a tidy solution. However, this type of weighting is actually canceled out by the FGLS estimator since the optimal weighting matrix re-scales all tokens to have equal variance. Instead, we can use token weights to scale the loss function as in weighted least squares.

Suppose we have a desired token weight \( w_t \) for each token \( t \), and diagonal weighting matrix.

\[
W_{(T \times T)} := \text{diag}(\begin{bmatrix} w_1 & w_2 & w_3 & \cdots & w_D \end{bmatrix})
\]

Because \( W \) is \((T \times T)\) we need to apply the weights to the loss before reducing the estimators to their document vector form. This suggests the following weight matrix for the VFGLS estimator:

\[
\Omega(\beta) := \text{Var}(y|F, \beta) = VW^{-1/2}\Sigma W^{-1/2}V^T
\]
Which is equivalent to scaling the variance of each token by $1/w_t$. In the final estimate, high weight/low variance tokens are given more importance when estimating influence.

### 3.3.3 Monte Carlo experiments

The previous sections describe several estimators and consider how to implement those estimators with general sets of token vectors and token weights. However, it is not clear how the choice of estimator, vectors, and weights affect the quality of estimates. In this section I present the results of several Monte Carlo experiments which are designed to a) test the effectiveness of different estimator types, and b) provide guidance on choices of vectors and weights.

**Comment sample and setup**

All of the results presented in this section are based on a common sample of real comments to the Office of the Comptroller of the Currency (OCC) from regulations.gov. I constructed the comment sample as follows: First, I started with all comments to the OCC on regulations.gov (9311 comments in 101 dockets). Then, I sampled a smaller number of comments from each docket, where the number of comments samples was equal to the total number of comments in the docket raised to the $2/3$ power. For example, if a docket had 39 comments in it, I took $39^{2/3} \approx 12$ random comments to represent this docket. The purpose of this sampling procedure was to scale down the size of dockets with thousands of comments while keeping most of the comments in small dockets. The final sample has 1170 comments in 80 dockets. The number of unique tokens is $T = 19,044$.

The comments were used to generate a series of synthetic rules according to the sampling procedure described in 3.2: Comments were first broken down into tokens (in this case single words with standardized capitalization) and converted to vectors of token frequencies. Then, for each docket I generated 100 random $\beta$ vectors. For each $\beta$, I constructed a synthetic rule from the real comments using $\beta$ as the sampling weights.

Each vector $\beta$ was generated independently according to the following procedure:

1. Draw element $\beta_0$ value from uniform(0,1)
2. Draw each element $\beta_d$ for $d = 2, ..., (D - 1)$ in sequence from uniform(0,1 $- \sum_{i=1}^{d-1} \beta_i$)
3. Set elements $\beta_D = 1 - \sum_{d=1}^{D-1} \beta_d$
4. Randomly shuffle the order of the elements of $\beta$

Thus, each $\beta$ is guaranteed to sum to one and will tend to have at least a handful of influential comments, even in very large dockets where the average influence of each comment must necessarily be tiny.

**Single-sample and Split-sample estimates**

As I discussed in section 3.3.2, one of the main advantages of using appropriate token vectors is that they should help to improve an estimator’s performance when the regulator is not using exactly the
same words as the commenter. But in order to observe this benefit, we need to generate synthetic rules with words that are related to, but not exactly the same, as the comments used to estimate influence. One way to achieve this property is to split every comment’s tokens into two separate samples and use one sample for generating the rule, and the other sample for estimating influence. Then the estimator never “sees” the true generating tokens, just a set of documents that are correlated with the generating documents because they came from the same comment. The correlations come in two types. First, words that are more common than average in a particular comment will also be more common in both the generating and estimating samples. Second, the comment is likely to use a variety of synonyms and closely-related topic words that co-occur with higher frequency than in the general corpus. When these are split between the generating and estimating samples, it provides an opportunity for token vectors to improve the quality of the estimates. When the vectors do a good job of capturing the semantic similarity between tokens, the differences between the vectors in the two samples should be smaller and the performance of the estimators should be higher. This is a much more realistic test of the estimator’s performance. I refer to this approach as “split-sample” (as opposed to “single-sample”), and I use it for every Monte Carlo test except the first where I am establishing baseline estimator performance.

Estimators

Definitions  I compare the performance of four estimators: LSA, Cosine, VGLS, VOLS, VFGLS:

1. LSA: This is a reproduction of the docket-level LSA similarity measure used in Chapter 2. The measure is computed from the raw token counts as follows: First, the comment token counts and rule token counts are concatenated into a single term-document matrix. Tokens that appear in less than 20% of the documents or more than 70% of the documents are dropped. Then the counts are transformed with TF-IDF weighting such that

\[ \text{weightedcount}_{dt} = \frac{\log(1 + c_{dt})}{\log(1 + \text{documentcount}_t)} \]

Where \( \text{documentcount}_t \) is the number of unique documents that contain token \( t \). Finally, the matrix of weighted counts is decomposed into token and document vectors using SVD. The estimates are computed as the cosine similarity between each comment document vector and the rule document vector. One major limitation of this measure is that when LSA is performed at the docket level, the maximum dimension of the vectors is constrained by the number of comments, which is often quite small.

2. Cosine: This is the simple cosine estimator often used for comparing the similarity of document vectors in text analysis, where

\[ \hat{\beta}_d = \frac{y \cdot x_d}{\|y\| \|x\|} \]

I include it as an intermediate step between LSA and the new linear estimators. The cosine measure is very similar to LSA except that it constructs document vectors from weighted means of the token vectors within the documents. This means that the cosine measure can use token
vectors that were pre-trained on a much larger set of documents than those in the docket (in this case all the comments in the sample). On the other hand, it might turn out to be the case that vectors trained on the docket alone are better at capturing the relationships between tokens that are distinct to a particular rule, in which case docket-level LSA may perform better.

3. VGLS: Vector Generalized Least Squares with the optimal weighting matrix derived in section 3.3.2. This estimator is not feasible in practice because the optimal weighting matrix is a function of $\beta$. However, it is a useful reference point when evaluating the FGLS estimator because it shows an upper limit of how well we can expect FGLS to perform when the procedure for estimating $\Omega$ works very well.

4. VOLS: Vector Ordinary least squares with no weighting matrix.

5. VFGLS: Vector Feasible Generalized Least Squares. The OLS estimate for $\beta$ is used to estimate the optimal weighting matrix using the formula derived in section 3.3.2. There is a small additional complication because $\beta$ is strictly positive and $\hat{\beta}_{VOLS}$ may have negative elements because of estimation error. These negative elements can cause the estimated weighting matrix to be non-positive semi-definite. My solution in these cases is to use a linear combination of $\hat{\beta}_{VOLS}$ and $1/D$ to construct the weighting matrix,

$$\hat{\beta} = (1 - \sqrt{\lambda})\hat{\beta}_{VOLS} + \sqrt{\lambda}\frac{1}{D}$$

where $\lambda$ solves $(1 - \lambda)min(\hat{\beta}_{VOLS}) + \lambda\frac{1}{D} = 0$. This ensures $min(\hat{\beta}) > 0$.

**Single-sample estimation results** The first Monte Carlo test establishes the performance of the four estimators when the rule is generated from the same comments that are observed (“single-sample”). Figure 3.1 shows the results. In these plots, each dot represents the mean performance of the estimator for 100 randomly generated rules using one docket’s comments. The 80 docket in the sample result in 80 data points in each plot. The x-axis is the number of comments in the dockets, so that it is possible to see how performance varies across dockets and docket size. Each column is a different estimator, and each row is a different measure of performance. Correlation between $\beta_d$ and $\hat{\beta}_d$ gives a measure of how much information the estimates extract about influence, ignoring issues of bias and whether the coefficients can be interpreted directly. Mean squared error (MSE) captures the overall accuracy. Bias indicates whether the estimates are on average too high or low. Attenuation indicates how estimates are biased towards zero for larger values of $\beta_d$. Coverage shows the fraction of 95% confidence intervals that cover the true $\beta$. To make comparing across plots easier, the mean score among all the dockets in each plot is represented as a solid horizontal line.

The main result is that all versions of the linear estimator perform better than both LSA and cosine similarity. Both LSA and cosine estimates have large biases, high mean-squared error, and of course, no standard errors for which to compute coverage. On important reason for the low performance of these estimators is related to the issue of interpretation raised in the introduction to this chapter. In particular, the relationship between these estimates and the true level of influence (defined by the
weight put on the comment when constructing the rule) is unknown. These results show that these estimates— and cosine estimates in particular—are badly calibrated: an estimate of $\hat{\beta}_d = 0.5$ should not be interpreted directly as the rule putting 50% weight on comment $d$.

However, even when we ignore bias, the correlation between $\beta_d$ and $\hat{\beta}_d$ still much is lower for cosine similarity and LSA than other estimates. This means that the problem with these estimators is not simply that they are biased—they also convey less information about the degree of influence than the new estimators. There are several factors that likely contribute to this result.

First, influence is assumed to be mutually exclusive (the elements of $\beta$ sum to one) and as the number of comments in the docket increases, an increasingly large fraction of the comments have essentially zero influence on the rule. The benefit of using the linear estimators instead of LSA or cosine similarity increases they compute all elements of $\beta$ simultaneously, implicitly incorporating information about the correlations between comments. When two comments are similar, their estimated influence is reduced (even though the estimated $\beta$ is not constrained to sum to one). The LSA and cosine measures just look at pairwise comment-rule relationships and ignore this additional information. This difference is particularly important for large dockets where most comments have negligible influence on the rule, but still have some average degree of similarity with the rule text.

The linear estimators also automatically weight vector components by the inverse of their variance. This places more weight on the informative dimensions of the document vectors. LSA attempts to do something similar by dropping extremely frequent and extremely rare tokens and using TF-IDF weighting for the rest, based on the assumption is that only tokens that occur with moderate frequency are informative about topic, and that the importance of a particular token is inversely related to the number of documents in which it appears. But in this test, the regulator does not differentiate between common and rare tokens when constructing the rule, and all are informative. The number of documents containing the token is probably helpful in guessing which tokens are informative, but it is a fairly ad hoc approach relative to the linear estimators’ inverse variance weighting.

Given the fact that LSA uses this additional weighting scheme it is surprising that the cosine estimates have such similar correlations with the true $\beta$. One possibility is that TF-IDF weighting does improve the LSA estimator, but the cosine estimator benefits instead from using token vectors that have been trained on a larger set of documents. LSA is based on dimensionality reduction of the term-document matrix, which is only beneficial when there is a very large number of documents and in this context there are many dockets with only a handful of comments. This gives very little information to learn about the co-occurrence of different tokens and construct high quality token vectors. More tests would be required to distinguish how different weighting these different factors influence the estimates.

Overall, the test results suggest that the linear estimators are a substantial improvement over LSA, both in terms of interpretability and accuracy. VFGLS is particularly effective, with the low mean-squared error and good coverage—performing essentially as well as the infeasible VGLS estimator that was included for reference. As expected, VOLS is unbiased, but has the higher MSE, and relatively poor coverage because it (incorrectly) assumes that the errors are i.i.d.

**Split-sample estimation results** In the second Monte Carlo test I run the same estimators (with the same betas) on the split-sample comment tokens. The results are presented in figure 3.2. As
expected, all the estimators have higher MSE and the estimates are attenuated. Coverage is also much worse. However, the new linear estimators still produce useful information even in this much harder task, and still perform substantially better than LSA or cosine similarity. In the following Monte Carlo tests I examine whether the choice of vectors and weights can improve performance on this task using the VFGLS estimator.

Vectors

Definitions I consider three types of vectors:

- Random: Completely random vectors for each token, drawn from a standard multivariate normal distribution
- Doc: LSA vectors computed on the whole multi-docket comment sample, using the usual document-level token frequencies.
- Par: LSA vectors computed on the whole multi-docket comment sample, but splitting comments into paragraphs and treating each paragraph as a separate document. This is an experimental approach to see if LSA vectors computed this way are higher quality. It is inspired by Word2Vec and other algorithms that use a small neighborhood around each word to infer the word’s vector.

Additionally, for each type of vector I run the estimator with several truncations at $k = 50, 100, 200, 400$.

Results Figures 3.3 and 3.4 show the results of the vector experiments. The first shows docket level scatter plots as before. The second figure averages docket scores together, but splits the estimates by $k$. The docket sample is restricted to those with fewer than 25 sample comments to avoid violating the $k \geq D$ identification condition. Overall, random vectors perform surprisingly well. We can expect that as $k \to T$, $V$ becomes full rank and the estimator converges to the token frequency-based described in section 3.3. This pattern is apparent in figure 3.4 where MSE decreases with $k$. But it is interesting how good the performance is with relatively low values of $k$ relative to $T$.

At the same time, it is clear that the LSA vectors (particularly the “doc” vectors) improve performance and reduce attenuation as hypothesized. The benefit of using LSA over random vectors appears to peak at a moderate degree of $k$. This suggests that LSA is successfully extracting a low-dimensional “latent” representation of the tokens that successfully removes much of the noise in the raw token counts. Given that the vectors do seem to matter, it will be worthwhile to experiment with a broader set of vector types in the future.

Weighting schemes

Definitions I test several weighting schemes that have been explored in the LSA literature (see Anaya (2011)).

1. None: Vectors are normalized by their standard deviation, but no differential token weights are applied.
2. Idf: Inverse document frequency weights, defined as follows:

\[ w_t = \frac{1}{\log(1 + \sum_{d=1}^{D} 1(c_{dt} > 0))} \]

3. Ent: Log-entropy weights, as defined in (Pincombe, 2004):

\[ w_t = 1 + \frac{\sum_{d=1}^{D} s_{dt} \log(s_{dt})}{\log(D)} \]

\[ s_{dt} = \frac{c_{dt}}{\sum_{j=1}^{D} c_{jt}} \]

4. Norm: Weights are set such that the column vectors of \( F \) have the equal euclidean length.

\[ w_t = \frac{1}{\sqrt{\sum_{d=1}^{D} c_{dt}^2}} \]

In each case, I apply the weights to tokens to the count matrix before computing LSA vectors as well as using them to adjust the estimator as in section 3.3.2.

**Results** Adjusting the token weights has minimal impact on estimator performance in this setting. One might expect that rare words are more informative and that increasing the weight on these words would improve the estimates. However, it seems that the optimal weighting matrix in the VFGLS estimator already accounts for this factor and is difficult to improve upon. I believe this is a small contribution to the NLP literature.
Figure 3.1: Single-sample estimation: This grid of plots summarizes the performance of the main linear estimators when the same input documents are used for generating the rule and estimating influence. Each dot represents the mean performance on one docket across 100 samples with random $\beta$'s. The solid horizontal line shows the average performance across all dockets. The horizontal axis is the number of comments in the docket. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, log Mean Squared Error (MSE), Bias (mean $\beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \epsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). Vectors are generated from document-level LSA ($k = 200$).
Figure 3.2: Split-sample estimation: This grid of plots summarizes the performance of the main linear estimators when the tokens of each input document are split into separate “gen” and “est” samples. The “gen” samples are used to construct the final rule, while the “est” samples are used to estimate influence. Each dot represents the mean performance on one docket across 100 samples with random $\beta$s (with the same gen/est split). The solid horizontal line shows the average performance across all doackets. The horizontal axis is the number of comments in the docket. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, Log Mean Squared Error (MSE), Bias (mean $\beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \epsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). Vectors are generated from document-level LSA ($k = 200$).
Figure 3.3: Split-sample vector performance comparison: This grid of plots summarizes the performance of three different types of vectors in the VFGLS estimator when the tokens of each input document are split into separate “gen” and “est” samples. The “gen” samples are used to construct the final rule, while the “est” samples are used to estimate influence. Each dot represents the mean performance on one docket across 100 samples with random $\beta$s (with the same gen/est split). The solid horizontal line shows the average performance across all dockets. The horizontal axis is the number of comments in the docket. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, Log Mean Squared Error (MSE), Bias (mean $\beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \varepsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$).
Figure 3.4: Split-sample vector performance comparison by $k$: This grid of plots summarizes the performance of three different types of vectors in the VFGLS estimator when the tokens of each input document are split into separate “gen” and “est” samples. The “gen” samples are used to construct the final rule, while the “est” samples are used to estimate influence. Each dot represents the mean performance average performance across all docketts for vectors truncated to length $k$. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, Log Mean Squared Error (MSE), Bias (mean $\beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \varepsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$).
Figure 3.5: Split-sample weighting performance comparison: This grid of plots summarizes the performance of four token weighting schemes in the VFGLS estimator when the tokens of each input document are split into separate “gen” and “est” samples. The “gen” samples are used to construct the final rule, while the “est” samples are used to estimate influence. Each dot represents the mean performance on one docket across 100 samples with random $\beta$s (with the same gen/est split). The solid horizontal line shows the average performance across all dockets. The horizontal axis is the number of comments in the docket. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, Log Mean Squared Error (MSE), Bias (mean $\beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \varepsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). Vectors are generated from document-level LSA ($k = 200$).
3.4 Panel Estimators

3.4.1 Motivation

The main advantage of panel data is that it allows one to distinguish between fixed and transient components of the data. For example, when estimating the effect of a new labour policy on wages, panel data allows researchers to separately control for each worker’s average wages using a fixed effects model. When we observe organizations submit comments on multiple rules, we can pursue an analogous estimation strategy to control for each author’s average writing style. By “style” I mean broadly any aspects of word choice that a) are correlated across comments by the same author, and b) do not substantially inform the meaning of the comment on a particular rule. For example, consider the following two excerpts that come from comments submitted to the Environmental Protection Agency (EPA) regarding the Clean Power Plan, a major regulation that would have limited CO2 emissions from power plants. The first is from an individual “TK” who opposes the proposed rule because it will increase electricity prices for seniors on fixed incomes:

“I am a retired veteran. Over 40 years we have been trying to fight climate change! The EPA has done its share, to both fight and deter thru reglations. Some good, some bad. I will not be alive to see the current results of EPA’s efforts; however, my primary concern is that the rule as proposed will result in significant electricity rate increases and additional energy costs for all users. The costs always seems to fall most heavily on we, the elderly, the poor, and those of us on fixed incomes. The Administration is blowing smoke, to make us believe that in 30 years we can overcome the pollution problem, much less the climate change, they dream of!”

The second is from a comment submitted by the Thomas Jefferson Institute for Public Policy:

“Energy prices and stagnant incomes are straining the budgets of Virginia’s lower- and middle-class families. Virginia households with gross annual incomes below $50,000, representing 41% of Virginia’s households, spend an estimated average of 20% of their after-tax income on energy. Energy bills for the 185,000 poorest households earning less than $10,000 represent 71% of their family incomes, before accounting for any energy assistance programs. Increased energy costs are competing with other necessities for lower- and middle-income family budgets across Virginia.

The impacts of increased energy costs are falling disproportionately on Virginia’s elderly Social Security recipients, representing 27% of the state’s households. In 2012, Social Security recipients in Virginia had average household Social Security incomes of $16,928. Some 21% of Virginia households had retirement income averaging $28,357 before taxes.

Unlike young working families, many fixed income seniors are limited to cost-of-living increases that do not keep pace with energy prices. Maintaining the relative affordability of electricity and natural gas prices, and increasing low-income energy assistance, are essential to the wellbeing of millions of Virginia’s senior and lower-income citizens.

These Virginia citizens cannot afford this rule and, as discussed below, it produces no benefits worth this cost and is not authorized under the Clean Air Act.”
The two comments make similar arguments, but use many different words. The first is written in informal language with non-standard spellings of “thru” and “reglations”, the use of “I” as a pronoun, and the vernacular “blowing smoke”. In contrast, the second comment is more formal in tone, contains no obvious spelling mistakes and uses many numerical quantities. Intuitively the two excerpts seem quite different, and we can expect these differences to be measurable in terms of token frequencies.

Now consider an example of the EPA discussing how the plan will affect electricity prices:

“Under Option 1, average nationwide retail electricity prices are projected to increase by roughly 6 to 7 percent in 2020 relative to the base case, and by roughly 3 percent in 2030 (contiguous U.S.). Average monthly electricity bills are anticipated to increase by roughly 3 percent in 2020, but decline by approximately 9 percent by 2030. This is a result of the increasing penetration of demand-side programs that more than offset increased prices to end users by their expected savings from reduced electricity use.”

Like the Thomas Jefferson Institute, the EPA uses formal language and many numerical quantities. None of the informal spellings and vernacular used in the first comment are present. Thus, it is likely that the second comment will be estimated to be more similar to the EPA’s rule text (and therefore more influential) than the first, even though the two comments make essentially the same argument. Furthermore, these differences are likely to persist across all comments submitted by these authors, leading to biased influence estimates for every rule.

Comments also often contain boilerplate text that are unrelated to the content of the message. Many commenters begin with a paragraph describing their organization such as these examples from comments to the OCC:

“The National Community Reinvestment Coalition (NCRC) is an organization of more than 600 community-based organizations with a mission to promote access to basic banking services, including credit and savings, to create and sustain affordable housing, job development, and vibrant communities for America’s working families. NCRC has been an outspoken advocate on these issues for more than 20 years.”

“The American Bankers Association represents banks of all sizes and charters and is the voice of the nation’s $14 trillion banking industry and its 2 million employees. Learn more at www.aba.com.”

Comments also often contain extraneous text such as email footers like this confidentiality notice from Community National Bank & Trust:

“CONFIDENTIALITY NOTICE: This message contains information from Community National Bank & Trust which may be confidential and privileged. If you are not an intended recipient, please refrain from any disclosure, copying, distribution or use of this information and note that such actions are prohibited. If you have received this transmission in error, please notify by e-mail techsupport@communitynational.net and destroy all copies of the original message. Thank you.”
Ideally these pieces of unrelated text would be removed from comments before estimating the similarity, but in practice it is difficult to reliably detect which text is unrelated to the current rule. If these pieces of extraneous text are correlated with the regulator’s average writing style (either positively or negatively), they will also bias individual influence estimates.

Existing pairwise similarity measures cannot solve this problem. As an illustrative example, suppose the rule document vector $y_r$ can be decomposed into a fixed “style” component $\bar{y}$ and a rule-specific “substance” component $\tilde{y}_r$, and each comment document vector $x_{ir}$ can be similarly decomposed into fixed “style” component $\bar{x}_i$ and a rule-specific “substance” component $\tilde{x}_{ir}$:

$$y_r = \bar{y} + \tilde{y}_r$$
$$x_{ir} = \bar{x}_i + \tilde{x}_{ir}$$

Here the “style” component contains token frequencies that reflect specific word choices that might identify the commenter, such as the use of “I” or “we”, vernacular terms, choices of specific synonyms, as well as portions of extraneous text that appear in every comment. The “substance” component on the other hand, contains all the deviations from the average style, and therefore the information in the comment that is specific to a particular rule.

Suppose we compute a pairwise similarity measure $s(y_r, x_{ir})$ between each rule $r$ and comment submitted by author $i$. For the linear measures discussed in this chapter,

$$s(y_r, x_{ir}) = s(\bar{y}, \bar{x}_i) + s(\tilde{y}_r, \tilde{x}_{ir}) + s(\bar{y}, \tilde{x}_{ir}) + s(\tilde{y}_r, \bar{x}_i)$$

If $s(\bar{y}, \tilde{x}_{ir})$ and $s(\tilde{y}_r, \bar{x}_i)$ are small on average (which seems likely given that the style and substance components will look very different), then we can focus on the similarity between styles and the similarity between substances:

$$s(y_r, x_{ir}) \approx s(\bar{y}, \tilde{x}_i) + s(\bar{y}_r, \tilde{x}_{ir})$$

Now consider two commenters, $a$ and $b$. Commenter $a$ is very influential, but has a very different writing style to the regulator so $s(\bar{y}, \tilde{x}_a) = 0$ and $s(\tilde{y}_r, \tilde{x}_{ar}) = \beta$. Commenter $b$ is not at all influential, but has a similar writing style to the regulator so $s(\bar{y}, \tilde{x}_b) = \beta$ and $s(\tilde{y}_r, \tilde{x}_{br}) = 0$. Then, when the pairwise estimates are computed, $s(y_r, x_{ar}) \approx s(y_r, x_{br}) \approx \beta$, and this relationship holds for every rule. So there is no observable difference between $a$ and $b$ using pairwise measures. Subtracting the mean pairwise similarity, for example, will not help at all.

The solution I propose here is to subtract the means of the document vectors ($\bar{y}$ and $\bar{x}_i$) before computing similarity. Then we isolate $s(\tilde{y}_r, \tilde{x}_{ir})$, the overlap in substance between the rule and comment. This operation is analogous to subtracting each worker’s mean wage before estimating the effect of the policy in the wage example above, and, as in the wage example, a consistent estimator can also be constructed using the appropriate fixed effects.

The remainder of this section takes a systematic approach to deriving a computationally feasible estimator based on this intuition.
3.4.2 Panel estimation

Suppose there are $I$ authors $i = 1, 2, ..., I$ who comment on $R$ rules $r = 1, 2, ..., R$, with a constant degrees of influence captured by the vector $\beta = [\beta_1, \beta_2, ..., \beta_I]^T$. Let $x_{ir} \in \mathbb{R}^k$ be the vector for the document submitted by author $i$ on rule $r$, and $y_r \in \mathbb{R}^k$ be the rule document vector for $r$. Then we can stack the observations as follows:

$$y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_R
\end{bmatrix}, \quad X = \begin{bmatrix}
x_{11} & x_{21} & \cdots & x_{I1} \\
x_{12} & x_{22} & \cdots & x_{I2} \\
\vdots & \vdots & \ddots & \vdots \\
x_{1R} & x_{2R} & \cdots & x_{IR}
\end{bmatrix}, \quad \omega = \begin{bmatrix}
\omega_1 \\
\omega_2 \\
\vdots \\
\omega_R
\end{bmatrix}$$

Such that

$$y = X\beta + \omega$$

We can use the fact that we know the distribution of each rule’s token sampling error $\epsilon_r$ to motivate a feasible generalized least squares estimator,

$$\hat{\beta}_{PFGLS} = \arg\min_\beta \left\| \Omega^{-1/2}(y - X\beta) \right\|^2$$

Because the token sampling error is independent across rules, the variance-covariance matrix of $\omega$ is block diagonal:

$$\Omega = \text{blockdiag}(\Omega_1, \Omega_2, ..., \Omega_R)$$

Where each block $\Omega_r$ is the same variance-covariance matrix as the cross-sectional estimator for that rule:

$$\Omega_r = VW^{-1/2}\Sigma_r W^{-1/2}V^T$$

$$\Sigma_r = \frac{1}{n_r}(\text{diag}(\pi_r) - \pi_r \cdot \pi_r^T)$$

As in the single-rule case, we can construct a feasible version of this GLS estimator with a two-step estimation procedure where $\hat{\Omega}$ is constructed from a first-stage, unweighted estimation of $\hat{\beta}$.

One consequence of moving to a panel estimator is that identification is easier. Recall that for individual dockets, the estimators required that the dimensionality of the document vectors $k$ was larger than the number of documents $D$, and no two documents could have the same text. With panel data this restrictions are relaxed. The rank condition for the panel estimator is

$$\text{Rank}(X) = I$$

This rank condition is much easier to meet, since:

1. Columns of $X$ will be unique even if commenters occasionally submit the same text for one rule (the rank condition only requires that they do not submit the same text for every rule)
2. The token vectors only need to have dimension \( k \) such that \( R \ast k \geq I \). So it is possible to estimate the influence of 100 commenters using 50-dimensional document vectors as long as there are at least two rules in the panel.

3.4.3 Modeling author writing style

Suppose each comment token frequency vector \( f_{ir} \) is generated as the sum of a style component \( \bar{f}_i \) and a substance component \( \tilde{f}_{ir} \) with the following properties:

\[
f_{ir} = \bar{f}_i + \tilde{f}_{ir}
\]

\[
E_r[\tilde{f}_{ir}] = 0
\]

In other words, each commenter \( i \) has a baseline token frequency \( \bar{f}_i \) representing their average commenting style, and a rule-specific deviation \( \tilde{f}_{ir} \) from that baseline frequency that accounts for the “substance” of each comment. Over enough rules, the rule-specific component averages to zero. I treat this property as an assumption, but it can also be thought of as a definition - if a commenter repeats the same content in every rule, the regulator is likely to discount it as repetitive and uninformative.

The regulator generates the token counts \( c_r \) for each rule \( r \) using only the substance component \( \tilde{f}_{ir} \) of each comment, plus an additional rule style component \( \bar{\pi} \):

\[
c_r \sim \text{mult}(n_r, \bar{\pi})
\]

Where I redefine \( \pi \) as

\[
\pi = \theta + \sum_{i=1}^{I} \beta_i \tilde{f}_{ir} = \bar{\pi} + \bar{F}_r \beta
\]

As before, let us define \( g_r = c_r/n_r \) as the observed token frequencies in the rule document. Then:

\[
E[g_r|\bar{F}_r, \beta] = \bar{\pi} + \bar{F}_r \beta
\]

\[
\text{Var}(g_r|\bar{\pi}, \bar{F}_r, \beta) = \frac{1}{n_r} (\text{diag}(\pi) - \pi \pi^T) = \frac{1}{n_r} (\text{diag}(\bar{\pi} + \hat{F}_r \beta) - (\bar{\pi} + \hat{F}_r \beta)(\bar{\pi} + \hat{F}_r \beta)^T)
\]

\[
\varepsilon := g_r - E[g_r|\hat{F}_r, \beta] = g_r - \bar{\pi} - \hat{F}_r \beta
\]

and we can write:

\[
g_r = \bar{\pi} + \hat{F}_r \beta + \varepsilon_r
\]

For simplicity I assume that the matrix of token vectors \( V \) is constant for all rules. Then we can pre-multiply all terms by \( V \) and write the document-vector version as
\[ y_r = \bar{y} + \bar{X}_r \beta + \omega_r \]

And if we stack the observations for each rule, we can write

\[ y = \bar{Y} + \bar{X} \beta + \omega \]

where

\[
y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_R \end{bmatrix}, \quad \bar{Y} = \begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix}, \quad \bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_R \end{bmatrix}, \quad \omega = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_R \end{bmatrix}
\]

We have now a data-generating process that captures the intuitions about writing style explored at the beginning of this section. Each commenter \( i \) has a fixed writing style represented by their average comment token frequency vector \( \bar{f}_i \). The regulator ignores this component of the comment text, and only uses deviations from the average comment to construct the rule. Similarly, the regulator has a fixed writing style represented by \( \bar{\pi} \) which they add to the text of every rule. The challenge is to estimate the influence vector \( \beta \) given that neither \( \bar{f} \) nor \( \bar{\pi} \) are directly observed.

### 3.4.4 Fixed effects estimation

There are two ways to implement a fixed effects estimator that will estimate \( \beta \). One is to estimate \( \bar{Y} \) and \( \bar{X} \) separately by taking the mean of \( y \) and subtracting the mean of \( X \), and then run the regression. This demeaning is a linear transformation that can be represented by a centering projection matrix \( P \) which removes the means across rules. The second is to add commenter-by-vector-component fixed dummy variables and estimate \( \beta \) in one step. The two approaches generate identical estimates, but it turns out that the dummy-variable estimator is more computationally efficient (at least when implementing the full FGLS estimator).

First, suppose we follow the demeaning approach. Then we estimate \( \bar{Y} \) or \( \bar{X} \), as follows:

\[
\bar{Y} = (I - P)y \\
\bar{X} = PX
\]

Then, we can convert the regression equation to one based purely on observables by pre-multiplying with \( P \) to give,

\[
P y = P X \beta + P \omega
\]

I call the FGLS estimator based on this equation the “demeaning” estimator:

\[
\hat{\beta}_{Demean} = \arg \min_{\beta} \left\| \Gamma^{-1/2} P (y - X \beta) \right\|^2
\]
\[ \Gamma = P\Omega P \]

Where \( \Omega \) is the optimal weighting matrix from the panel estimator above.

Alternatively, one can simultaneously estimate \( \bar{Y} \) and \( \tilde{X} \) with dummy variables. Let \( y_j, X_j \) indicate the set of \( R \) observations in the panel data that share a common vector component \( j = 1, 2, \ldots, k \) (It might be helpful to imagine sorting the rows in the data by vector component instead of regulation, so that each block of the data has one observation for each regulation from a common vector component). Then we can replace

\[ \bar{Y}_j = \mathbf{1} \alpha_j \]
\[ \tilde{X}_j = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \text{diag}(\gamma_j) \]

To get

\[ y_j = \mathbf{1} \alpha_j + X_j \beta - \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \text{diag}(\gamma_j) \beta + \delta_j \]

Note however, that \( \mathbf{1} \alpha_j \) and \( \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \text{diag}(\gamma_j) \beta \) are perfectly collinear. So we can actually just simplify the regression to

\[ y_j = X_j \beta + \mathbf{1} \mu_j + \omega_j \]

where \( \mu_j = \alpha_j + \text{diag}(\gamma_j) \beta \).

In other words, commenter and regulator styles cannot be separately identified from the data. Conversely, adding vector component-specific fixed effects is sufficient to control for both commenter and regulator fixed components (“style”). Thus, the “dummy” estimator can be written as,

\[ \hat{\beta}_{\text{Dummy}} = \arg \min_{\beta} \left\| \Omega^{-1/2}(y - X\beta - M\mu) \right\|^2 \]

where \( M = \mathbf{1}_R \oplus I_k \) using Kronecker product notation.

Note that \( P \) projects onto the orthogonal complement of \( M \). Therefore, by the Frisch-Waugh-Lovell theorem (Frisch and Waugh, 1933), the demeaning and dummy estimators will give the same estimates for \( \hat{\beta} \). However, they are not computationally identical for VFGLS estimation because the demeaning estimator requires computing \( (P\Omega P)^{-1/2} \), while the dummy estimator only requires computing \( \Omega^{-1/2} \), which is much easier. To see why, note that \( \Omega \) is block diagonal. Therefore

\[ \Omega^{-1/2} = \text{blockdiagonal}(\Omega_1^{-1/2}, \Omega_2^{-1/2}, \ldots, \Omega_R^{-1/2}) \]

Thus, the weighting transformation can be applied to each set of rule observations independently, just as in the single-rule estimators. In contrast, \( P\Omega P \) is a dense \( kR \times kR \) matrix that may be difficult to hold in memory, much less compute a square root factorization and solve for an inverse transformation.

On the other hand, the dummy estimator requires an additional step of computing \( \Omega^{-1/2}M \) and adding an additional \( k \) variables to the OLS optimization. The block-diagonal structure of \( \Omega \) also helps
when computing $\Omega^{-1/2} M$. Specifically,

$$\Omega^{-1/2} M = \begin{bmatrix}
\Omega_1^{-1/2} \\
\Omega_2^{-1/2} \\
\vdots \\
\Omega_R^{-1/2}
\end{bmatrix}$$

Given a reasonable value of $k$, the additional burden of estimating $\mu$ should not be that high. The dummy estimator is likely to be much more computationally efficient than the demeaning estimator.

### 3.4.5 Monte Carlo Experiments

Here I present the results of Monte Carlo tests with panel estimators. The main goal of these tests is to examine how the existence of commenter and regulator style affect the performance of the basic panel estimators, and to determine whether the fixed-effects estimation approach described above compensates for this more realistic data generating process. To provide a point of comparison, I construct synthetic rules with and without style components, and run both FE and non-FE estimators on both sets of synthetic rules. This allows me to examine how robust each estimator is to misspecification. To make the comparison as realistic possible, I again use real comments submitted to the OCC, and use real rules published by the OCC to choose the regulator’s average writing style. The results of the tests suggest that style can generate huge biases in influence estimates. However, the tests also show that fixed effects are an effective way of removing style bias.

**Practical considerations**

The derivation of the panel data estimators makes two assumptions that are not completely realistic. First, the derivations assume a balanced panel. In contrast, real commenting data almost always forms an unbalanced panel (most commenters do not comment on every rule). I address this issue in one of two ways, depending on the type of estimator. When there are no fixed effects, I treat missing values as zeros. When fixed effects are used, I substituting missing comment token frequencies with the commenter’s average style $\bar{f}_i$. In effect, the commenters are assumed to submit perfectly generic comments (relative to their own commenting style) on every rule. When regulators only pay attention to deviations from average style, submitting a generic comment and not submitting at all become equivalent.

The second problem is that $\bar{\pi} + \tilde{F}_r \beta$ may be negative for some tokens. The negative values represent a fundamental tension between the assumptions that $\bar{\pi} + \tilde{F}_r \beta$ is a valid probability vector and the assumption that $E[\tilde{F}_r] = 0$. The latter implies that $\tilde{F}_r \beta$ will often have negative components, and there is no guarantee they will be smaller in magnitude than the corresponding (positive) components in $\bar{\pi}$. For example, this problem will always occur when a commenter *occasionally* uses a token that the regulator *never* uses. Then the comments lacking this token will have $\tilde{f}_{rt} < 0$ while $\bar{\pi}_t = 0$. Improving the modeling of this style component might be a useful direction for future work. In the meantime, I take the expedient approach of replacing negative values with zero. One limitation of this approach is
that it adds noise to the generated rule and likely decreases the apparent effectiveness of the estimators.

Data

The sample of comments used in these tests consists of all comments submitted to the OCC by organizations that submit at least 5 comments in the years 2003 to 2017. The minimum comment count ensures that commenter style is identified and can be reasonably estimated, and also keeps the size of the regression matrices manageable. The input data consists of 1289 comments from 108 unique commenters in 82 dockets. To estimate the OCC’s average writing style, I use the set of all rules published by the OCC to the federal Register since 2000 (92 documents). For each rule, I extract the portion of the rule text that contains the agency’s discussion of the comments using the same technique as in Chapter 2.

Test procedure

I construct Monte Carlo influence estimates according to the following procedure:

1. I split each comment into “gen” and “est” token samples, as in the split-sample tests described in Section 3.3

2. I compute a style vector \( \tilde{f}_i \) for each commenter by taking the mean token frequency across the “gen” comment token samples

3. I compute the regulator’s style vector \( \tilde{\pi} \) by taking the average token frequency across all discussion text from OCC rules

4. I compute \( V \) with document level LSA using both comment “gen” tokens and the OCC discussion text

5. I draw 1000 random \( \beta \)s using the procedure described in Section 3.3.3

6. For each random \( \beta \):

   (a) I construct two synthetic rules for each docket in the comment sample: One where the rule is a simple mixture of input comments such that \( \pi_r = F_r \beta \), the other where the rule is a mixture of the regulator’s base style and the deviation from commenter’s average styles such that \( \pi_r = \tilde{\pi} + \tilde{F}_r \beta \).

   (b) I run each estimator on each synthetic rule, with and without style fixed effects, to generate a total of four estimates. For two of these estimates, the data generating process matches the estimator’s modeling assumptions, for the other two estimates, the estimators are mis-specified - either using style fixed effects when they are not necessary, or omitting them when they should be included.
Estimators

I compare the performance of four estimators: Cosine, VOLS, VFGLS:

1. Cosine: The cosine estimator, extended to panel data as follows:

\[ \hat{\beta}_d = \frac{y \cdot X_i}{\|y\| \|X_i\|} \]

Where \( X_i \) is the formed by concatenating the document vectors \( x_{i1}, x_{i2}, \ldots, x_{iR} \). The fixed effects version is implemented by subtracting the observed mean token frequencies for each commenter from each of their comments before computing \( \hat{\beta}_d \) (the cosine equivalent of the demeaning estimator derived above).

2. VOLS: Panel Vector Ordinary least squares with no weighting matrix.

3. VFGLS: Panel Vector Feasible Generalized Least Squares. The VOLS estimate for \( \beta \) is used to estimate the optimal weighting matrix using the formula derived in section 3.4. Each \( \Omega_r \) is computed as

\[ \frac{1}{n_r} V(\text{diag}(\hat{\pi}_r) - \hat{\pi}_r \cdot \hat{\pi}_r^T) V^T \]

For the non-fixed effects version of the estimator, \( \hat{\pi} \) is computed the same way as for the cross sectional estimators (\( \hat{\pi} = F\hat{\beta} \)). However, for the fixed effects version, the formula for \( \hat{\pi} \) is,

\[ \hat{\pi} = \bar{\pi} + \tilde{F}\hat{\beta} \]

where \( \bar{\pi} \) and \( \tilde{F} \) are estimated from the observed data. I apply the same correction to the estimated \( \beta \) strictly positive as in the cross-sectional Monte Carlo analysis.

Results

Figures 3.6 and 3.7 summarize the results of the Monte Carlo test. Figure 3.6 shows the results when the rules are generated without style effects. Each column contains estimates from a single estimator, and each row shows a different measure of the quality of the estimates. Each point represents the average performance for a specific commenter.

The x-coordinate of each point is determined by the similarity between the commenter’s average comment vector \( \bar{x}_i \) and the regulator’s average rule vector \( \bar{y}_r \). In other words, commenters with a similar writing styles to the OCC will appear to the right in each plot, while commenters that have very different writing styles from the OCC will appear on the left. When rules are generated with style effects, it seems likely that the this correlation between commenter and regulator style will affect the performance of the estimators, and separating the data along this dimension allows us to inspect for these types of patterns visually. Because this style similarity is fixed across all sample estimates, the x-coordinate for each commenter is constant across all plots. Most commenters have a correlation of between 0.05 to 0.1 with the style of the OCC. Note that this is a highly selected sample, since only organizations that are very active in the regulatory process comment at least five times. Manually
exploring the comments suggests that the sample is dominated by banks and organizations representing banks. Less frequent commenters would probably have writing styles that are, on average, less similar to the OCC.

The main result of these Monte Carlo tests is that generating rules with style effects makes a large difference to the influence estimates and it is very important to include fixed effects to compensate. VFGLS is particularly sensitive to omitting fixed effects when they are required, and this type of misspecification results in extremely high mean-squared error. On the other hand, the performance penalty for including fixed effects when they are unnecessary is negligible. The story for VOLS is similar, though the differences are not as large. For simple cosine estimation, fixed effects actually help even when there are no style effects. Overall, the results suggest that if there is any doubt about whether the true data generating process includes style effects, fixed effects estimators should be used.

Figure 3.8 illustrates how writing style and fixed effects change the estimates. Here I show the individual VOLS point estimates \( \hat{\beta}_i \) plotted against the true values of \( \beta_i \) for three random commenters. In the top-left quadrant the rules are generated without any style effects, and the estimator does not include fixed effects. Here the estimator is correctly specified and the estimates line up along the 45-degree line where \( \hat{\beta}_i = \beta_i \). In the top-right quadrant, the estimator includes unnecessary style fixed effects, but they do not have a large effect on the estimates. In the bottom row, the rules are generated with style effects and the results are quite different. First, when the estimator lacks fixed effects (bottom-left quadrant), each commenter’s estimates are offset by an individual bias that shifts all of their estimates up or down by a constant amount. These biases are very large relative to the estimates themselves. For example, the green commenter is estimated to be substantially more influential than the other two commenters, regardless of the true value of \( \beta_i \). For VFGLS these biases are so large that the estimates often fall far outside of the range of the plot. Given that the actual correlations between commenter and regulator style are only on the order of 0.05-0.1, this pattern suggests that the non-FE VOLS and VFGLS estimators exaggerate similarity in unpredictable ways. However, when fixed effects are included (bottom-right quadrant), all of the commenter-specific biases are removed, leading to accurate estimates.

3.5 Concluding Remarks

The goal of this chapter is to provide improved techniques for measuring political influence from text data. I explore a class of estimators that are based on a novel combination of existing techniques: bag-of-words features for representing documents, word embeddings, and generalized least squares estimators. These estimators can be applied in situations that competing approaches such as manual coding or Wordfish cannot, and improve on the LSA algorithm we applied in 2 in several ways.

I use Monte Carlo tests to demonstrate that the estimators are unbiased and have correct coverage when the rules are generated exactly as assumed. While LSA generates estimates that are correlated with the true degree of influence, the LSA estimates are also biased upwards, less precise, and provide no standard errors. I also test a more realistic variation of the rule-generation process that attempts to mimic a regulator who samples from comments, but replaces many of the words with synonyms and related terms instead of using exactly the same language. I find that the estimators still perform much better.
Figure 3.6: This grid of plots summarizes the performance of the panel estimators when the rules are generated with no style effects. Each dot represents the mean estimate for one commenter across 1000 samples with random $\beta$s. The solid horizontal line shows the average performance across all commenters. The horizontal axis is the similarity between the commenter’s style and the regulator style, as measured by the correlation between $\bar{x}_i$ and $\bar{y}_i$. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, log Mean Squared Error (MSE), Bias (mean $\beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \varepsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). Vectors are generated from document-level LSA ($k = 100$).
Figure 3.7: This grid of plots summarizes the performance of the panel estimators when the rules are generated with style effects. Each dot represents the mean estimate for one commenter across 1000 samples with random $\beta$. The solid horizontal line shows the average performance across all commenters. The horizontal axis is the similarity between the commenter’s style and the regulator style, as measured by the correlation between $\bar{x}_i$ and $\bar{y}$. The vertical axes show four measures of estimator quality: Correlation is the Pearson correlation coefficient for $\beta_d$ and $\hat{\beta}_d$, log Mean Squared Error (MSE), Bias (mean $\beta_d - \hat{\beta}_d$), Attenuation (equal to $1 - \alpha$, where $\alpha$ is the slope coefficient in a regression of the form $\hat{\beta}_d = \alpha \beta_d + \varepsilon$), and Coverage (the fraction of estimates for which the estimated 95% confidence interval for $\hat{\beta}_d$ includes the true $\beta_d$). Vectors are generated from document-level LSA ($k = 100$).
Figure 3.8: This grid of plots shows the point estimates for all 1000 samples for three randomly selected commenters (indicated by colour) using panel VOLS. Each horizontal axis indicates the true value being estimated, while the vertical axis indicates the estimate. Points along the 45 degree line are correctly estimated. In the top row, rules are generated with no style effects, while in the bottom row the rules are generated with style effects. In the left column the estimator has no fixed effects, while in the right column the estimator uses style fixed effects.
better than LSA on this more realistic task.

The proposed estimators can also be extended to control for commenter and regulator writing style when panel data is available with authors submitting comments on multiple rules. Intuitively, the presence of writing style that is separate from the information content of documents is likely to be a large source of bias when estimating influence since commenters with writing style that is similar to the regulator will appear more influential than they actually are. Existing tools for text similarity analysis do not have a good way of dealing with the problem. In the last section of the paper I derive a more realistic model of rule generation that incorporates commenter and regulator writing styles and show that it motivates adding a particular set of fixed effects to the estimators. I then present Monte Carlo tests that use real comments and regulatory text to generate synthetic rules with realistic commenter and regulator writing styles. These tests demonstrate that style biases have a large detrimental impact on the performance of the proposed estimators, but these biases are effectively eliminated by the proposed fixed effects specification. I also show that including fixed effects when there the documents have no consistent writing style is relatively costless, suggesting that the fixed effects estimators should be the default choice when panel data is available.

When thinking about the big picture of how to measure influence from text data, I see two broad paths. The first is to constrain ourselves to traditional econometrics and statistics, applied to discrete text features such as bag-of-words tokens. This approach sacrifices the realism of how we treat text because it ignores much of the nuance and contextual information contained in the sequential structure. It is possible to capture sequential information in the tokens (by using bi-words for example), but longer sequences become increasingly rare and unlikely to appear in multiple documents, even when they have similar or exactly the same meaning as other related sequences. As a thought experiment, consider that we could use the entire document’s text as a single unique feature - but then every document is completely orthogonal and we have no way to think about similarity. On the other hand, we could also represent each document as a count of individual letters, but this would also not be very informative. The optimal token size is somewhere between individual letters and full documents, but there are real limitations to how effectively discrete features can represent text information. Topic modeling and word embeddings improve text analysis measures by capturing some of the relatedness between words or tokens, but this approach still strips away a huge amount of information from the ordering (for example, who “he” refers to in a paragraph).

The second path forward is to embrace new NLP approaches based on neural networks. This is an exciting frontier, and researchers are regularly improving performance on a variety of benchmark tasks (see Radford et al. (2019); Vaswani et al. (2017); Devlin et al. (2018) for a sample of recent research in this area). These approaches use very flexible models that can capture much more information about the context of language features, at the cost of computational complexity and interpretability. The general recipe for applying these models is to train the model to predict omitted tokens on a large corpus of text. The model learns to construct an internal representation of the text that maximizes predictive accuracy. Then supervised learning is then used to train another classifier to solve the main problem (sentiment or topic classification for example) using the model’s internal state as input instead of the raw text. Neural network models are therefore a natural extension of token vectors: the model...
generates a vector for each token in the sequence that captures some lower-dimensional representation of the text and assigns similar vectors to similar tokens. The difference is that the vector assigned to each token is context-specific. This is what makes higher accuracy possible. But it also means that it is much harder to write down a simple model of how a regulator might generate text, and it is also harder to see how to model the covariance structure of token vectors. These are open problems for future research. In the meantime, more traditional approaches, like those outlined in this chapter, remain useful.
<table>
<thead>
<tr>
<th>Estimator</th>
<th>OLS version</th>
<th>GLS version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>$\hat{\beta}<em>{OLS} = \arg \min</em>{\beta} | g_r - F_r \beta |^2$</td>
<td>$\hat{\beta}<em>{FGLS} = \arg \min</em>{\beta} | \Sigma_r^{-1/2} (g_r - F_r \beta) |^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\Sigma_r = \frac{1}{n_r} (\text{diag}(\pi_r) - \pi_r \cdot \pi_r^T)$</td>
</tr>
<tr>
<td>Vector</td>
<td>$\hat{\beta}<em>{VOLS} = \arg \min</em>{\beta} | y_r - X_r \beta |^2$</td>
<td>$\hat{\beta}<em>{VFGLS} = \arg \min</em>{\beta} | \Omega_r^{-1/2} (y_r - X_r \beta) |^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\Omega_r = VW^{-1/2} \Sigma_r W^{-1/2} V^T$</td>
</tr>
<tr>
<td>Panel</td>
<td>$\hat{\beta}<em>{PVOLS} = \arg \min</em>{\beta} | y - X \beta |^2$</td>
<td>$\hat{\beta}<em>{PVFGLS} = \arg \min</em>{\beta} | \Omega^{-1/2} (y - X \beta) |^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\Omega = \text{blockdiag}(\Omega_1, \Omega_2, \ldots, \Omega_R)$</td>
</tr>
<tr>
<td>FE</td>
<td>$\hat{\beta}<em>{VOLSFE} = \arg \min</em>{\beta} | y - X \beta - M \mu |^2$</td>
<td>$\hat{\beta}<em>{VFGLSFE} = \arg \min</em>{\beta} | \Omega^{-1/2} (y - X \beta - M \mu) |^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\Omega = \text{blockdiag}(\Omega_1, \Omega_2, \ldots, \Omega_R)$</td>
</tr>
</tbody>
</table>
Conclusion

The overarching goal of this dissertation is to advance the study of political influence by providing new data from the U.S. federal rulemaking process, new tools to analyze this data, and new empirical results generated by applying the tools to the data. The rulemaking process is an exciting area to explore because agencies record a huge amount of detailed interactions with interest groups during the notice-and-comment process. In federal rulemaking we can observe specific organizations communicating detailed arguments regarding proposed policies, and we can observe how the government agencies respond to these comments, both in terms of how they describe the comments, and how they change the substance of the rule from the proposed to the final version. All this information is available at a vast scale—millions of comments on thousands of regulations. This data has the potential to transform our understanding of influence, particularly the kind of informational lobbying that is central in advanced economies. But it brings challenges too. Linking the documents together into a coherent database is not trivial. And some of the most interesting aspects of the data take the form of complex text which requires new tools and techniques to fully exploit.

The role of chapters 1 and 2 are therefore twofold: First, they reveal systematic evidence of a political strategy employed by large corporations that was previously known only from a few anecdotal accounts. It turns out that large U.S. corporations donate millions of dollars to non-profits who participate in federal rulemaking, and analysis of the timing and content of these comments suggest that firms are able to amplify their messages to regulators with these donations. The patterns we observe are consistent with only a few stories about how the process works: commenting is correlated with donations in the year prior, so we do not see evidence that non-profits are courting donors with their comments, instead, firms and non-profits must be coordinating in advance. Similarly, we see the strongest relationship between commenting and donations in the set of non-profits that comment most often, which is inconsistent with stories of “hush money” where firms use donations to keep active commenters quiet. This new evidence advances our understanding of how firms interact with regulators and the tools they can use to influence policy debates in this domain. This evidence is a direct continuation and expansion of the new interpretation of corporate philanthropy as a tool for political influence first introduced in Bertrand et al. (2018).

Second, chapters 1 and 2 are “proof of concept” for further research in on influence in rulemaking. One of the challenges with exploring new data with new techniques is that it is hard to tell how well the new techniques are working when we don’t know what to expect from the data. Linking the rulemaking data to donations provides an external reference point. The result that the timing of comments is strongly correlated with donations in very stringent regressions with many controls indicates that there is valuable information embedded in the comment submission data. And the fact that these donations are also correlated with our measures of textual similarity suggest that even these
relatively simple algorithms are extracting meaningful information from the content of comments and rules.

The success of the empirical approaches in chapters 1 and 2 are particularly remarkable considering the limitations of the data. On the donations side, we only have access to donations that occur through corporate foundations and not direct donations from firms (or corporate executives and employees). Given the ethically and legally dubious nature of the coordination between firms and non-profits it is somewhat surprising that firms use the more transparent grant process for donating. It may be the case that we are only observing a small fraction of the politically-motivated donations. On the regulatory side, there are many sources of noise that result from the messy nature of the raw data. Linking organizations to comments and regulatory documents to each other (and therefore comments to rules via the preceding proposal documents) required automated, approximate matching techniques.

Similarly, the measures of text similarity and sentiment provide only very rough approximations of what a human reader would infer from the comment and rule text. Before running our analysis, it was not obvious that any meaningful signal could be extracted from the data. The results in chapters 1 and 2 demonstrate that there is a path forward. The dataset is large enough that scale can compensate for precision in each measurement. And the panel structure allows for very stringent regression specifications that can control for many obvious confounders such as patterns within industry, year, or between particular pairs of firms and non-profits.

Nevertheless, some difficult issues remain. The first is the interpretation of topic and sentiment similarity between comments. Because some comments are very long complicated documents, it is possible that some comments might take opposing positions on overlapping issues and still appear to have higher similarity than other comments that talk about unrelated issues. Document-level sentiment scores alone cannot solve this problem since the sentiment needs to be aligned with each topic. There are existing text analysis technologies that attempt to solve the problem of jointly estimating topics and sentiment which may be worth exploring in future work (for example see Lin and He (2009); McAuley and Leskovec (2013); Rahman and Wang (2016); Li et al. (2010)).

The second issue is the interpretation of text similarity between comments and the rule discussions. In the rule discussions, regulators must summarize comments regardless of whether they act on them. From my experience reading the rule text, I believe it is likely that regulators spend more time discussing comments that influenced their thinking than those that did not. And when regulators make changes to the rule, it seems likely that the influential comments will also be more similar to the regulator’s plain language description of the final rule. But we still need to quantify these intuitions more precisely (something I am currently working on with my co-authors).

Even if the text similarity between comments and rules is correctly measuring the way that regulators respond to comments, there is a deeper issue of causal identification. Large firms frequently engage in many channels of influence simultaneously, in the form of coordinated multi-pronged campaigns. For example, in the case of the high-profile Clean Power Plan rulemaking, work by Nouri Najjar and I (unpublished) found that many large energy companies simultaneously commented, lobbied, and sued to pressure the Environmental Protection Agency to delay the rule. Doubtless, these large firms also called and emailed any contacts they had at the agency or White House, attended public and private
meetings with EPA representatives, and asked members of Congress to pressure the agency as well. It is very difficult to disentangle the casual effects of comments from these other correlated activities. On the other hand, when we think about the causal impact of comments specifically we are perhaps focusing on an odd counterfactual. When firms have multiple channels available to communicate the same message, the causal effect of blocking one channel might be close to zero. But this does not mean that the causal impact of the message content is zero. Comments are somewhat special because they are part of the public record and are given consideration in legal challenges to rules. But in general, I think a better way to view the comments is as a sample of the information passing from interest groups to government, and the policy preferences of interest groups on very fine-scale issues. Then the important questions to answer are how the communication visible in comments is biased relative to the full set of messages, and whether it is possible to instrument for an organization’s general ability to communicate with government.

In future work, my first goal is to improve the measure of influence from comments to rules, and explore what factors predict influence. Chapter 3 is a building block for this project, as it proposes better estimators for the influence of comment text on rule text. In particular, the new estimators make it possible to control for fixed aspects of writing style. In Chapter 2, we had time-varying donations which allowed us to include various controls for the average level of donations and commenting by firms. But without this additional time-varying input it will be particularly important to distinguish between commenters that write in a similar style to the regulator, and commenters who are influential because the regulator integrates content from their comments into the rule. To address how comments might be a biased sample of communication, it will also be useful to investigate how interest groups conduct multi-pronged campaigns. With linked lobbying, campaign finance, and commenting data, we can explore how groups allocate resources across these channels. Court cases would be another area worth exploring, both because the threat of lawsuits provides a source of leverage over regulators, and because courts take input from the public in the form of amicus briefs that parallel regulatory comments. With improved understanding of the main empirical patterns in these activities, I hope it will eventually be possible to find good instruments for influence. Promising areas to look include differences in the resources of firms due to size and market structure, different levels of past regulation leading to more or less experience interacting with government, and personal ties between corporate executives and politicians or government personnel. In the long run, I believe this research path will lead to a more nuanced and complete understanding of how policy is shaped by information provided by interest groups.


Ottaviani, M., Sørensen, P.N., 2006. Reputational cheap talk. The Rand journal of economics 37, 155–175.


Appendix A

Regulation comments

A.1 Overview

Our data on regulatory comments comes from regulations.gov. Under the Administrative Procedures Act (APA), federal agencies must provide a means for the public to submit comments on proposed rules and other regulatory changes. Regulations.gov is a shared platform that is now used by most federal agencies to facilitate submission and public review of comments. Information about submitted comments, including the original text and attachments, can be viewed through a web browser. The site also provides an API that allows more efficient data access, particularly for collecting simple comment metadata such as the title of the comment and posted date.

Our sample starts with the complete collection of metadata for all comments posted to regulations.gov in the years 2003-2017 (inclusive). This is a total of 6,871,697 unique documents. From these, we identify 981,232 comments that appear to be authored by organizations rather than private individuals (“org comments”). We download the complete text for all org comments using common file formats, giving us about 90% of comment text for the org comment sample.

A.2 Collecting metadata

The regulations.gov API provides a search function for document metadata. We retrieved the metadata for all public submission documents posted since the site came online in 2003, and include all years up to and including 2017. Some agencies have begun digitizing older comments and posting them to regulations.gov retroactively. But an EPA spokesperson stated (in personal email correspondence) that this work is currently incomplete, and that the text of some older comments will never be released digitally since the submitters were not aware of this possibility at the time. Thus we consider data on pre-2003 comments on regulations.gov unreliable and do not include them.

A.3 Identifying org comments

Authorship information can appear in three different metadata fields: “title”, “organization”, or “submitterName”. Comments appear to fall into two main types: those that contain “organization” and/or “submitterName” information, and those that only contain authorship information in the title. First, we drop all comments that have “submitterName” information, but no organization. These appear to be written by private individuals. For the remaining comments, we look for an organization name in either the organization field or the title (if the organization field is blank). We use a custom neural
network-based classifier to extract organization names from the selected field (classification is necessary for the organization field because it contains many false positives such as “self” or “none”). The classifier converts each title string to ASCII characters and predicts whether each character is part of an organization string. Contiguous chunks of characters with predicted probability greater than 0.5 are counted as organization names. The classifier is multi-layer bi-directional Gated Recurrent Unit (GRU), implemented in PyTorch\(^{39}\). Code is available on the Brad Hackinen’s github page\(^{40}\). The classifier is trained on almost 9000 manually constructed training examples. This training set was constructed iteratively by starting with easy to parse titles, fitting the neural network, estimating the classifier’s uncertainty from the total entropy of the character-level predicted probabilities, reviewing a sample of high-entropy titles, adding them to the training set, and repeating until the error rate was acceptably low. We also manually classified an additional set of 1000 random titles as a test set. The results of the test are shown below. 93% of titles are classified without error. 83% of titles with an organization are extracted exactly correctly, while 98.5% of titles with no org are extracted correctly (in other words, the classifier avoids 98.5% of false positives).

Table A.1: Organization name extraction accuracy

<table>
<thead>
<tr>
<th>Sample</th>
<th>Count</th>
<th>Character Accuracy</th>
<th>String Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>All test titles</td>
<td>1000</td>
<td>0.970</td>
<td>0.928</td>
</tr>
<tr>
<td>Test titles containing org</td>
<td>371</td>
<td>0.935</td>
<td>0.830</td>
</tr>
<tr>
<td>Test titles with no org</td>
<td>629</td>
<td>0.991</td>
<td>0.985</td>
</tr>
</tbody>
</table>

Notes: Character accuracy is the average fraction of characters classified correctly in each title. String accuracy is the fraction of titles with every character correctly classified.

A.4 Collecting comment text

Comments on regulations.gov can have comment text in two locations: a “text” field in the comment metadata, or in one or more attachments. The “text” field contains text that submitters have entered on a web form. It is often as brief as “see attached”. Most substantial text is contained in the comment attachments where submitters can upload PDFs, word documents, other other file formats. We download all attachments of the following formats: PDF, MS Word 8, MS Word 12, and simple .txt files. The majority of attachments are in PDF format.

We use the XpdfReader `pdftotext`\(^{41}\) command-line utility to extract text from most PDFs. Some PDFs contain only images of each page. In this case we must fall back on Optical Character Recognition (OCR), which we implement with a combination of `GhostScript`\(^{42}\) (to render page images) and `Tesseract-OCR`\(^{43}\). We use Apache Tika\(^{44}\) to extract text from MS Word formats, and the `chardet`\(^{45}\) Python package to detect formatting of simple text files. All the tools are open source.

\(^{39}\) https://pytorch.org/
\(^{40}\) https://github.com/bradhackinen/subex
\(^{41}\) https://www.xpdfreader.com/pdftotext-man.html
\(^{42}\) https://www.ghostscript.com/
\(^{43}\) https://github.com/tesseract-ocr
\(^{44}\) http://tika.apache.org/
\(^{45}\) https://pypi.org/project/chardet/
# Appendix B

## Additional tables and figures

<table>
<thead>
<tr>
<th>Top 30 agencies in firms comments</th>
<th>Number of comments</th>
<th>Top 30 agencies in grantees comments</th>
<th>Number of comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPA</td>
<td>8099</td>
<td>FWS</td>
<td>76404</td>
</tr>
<tr>
<td>FAA</td>
<td>3870</td>
<td>NOAA</td>
<td>69171</td>
</tr>
<tr>
<td>FDA</td>
<td>1942</td>
<td>HHS</td>
<td>60969</td>
</tr>
<tr>
<td>OSHA</td>
<td>1245</td>
<td>CMS</td>
<td>47215</td>
</tr>
<tr>
<td>PHMSA</td>
<td>745</td>
<td>EPA</td>
<td>13556</td>
</tr>
<tr>
<td>NHTSA</td>
<td>724</td>
<td>ED</td>
<td>5105</td>
</tr>
<tr>
<td>CMS</td>
<td>721</td>
<td>FDA</td>
<td>4773</td>
</tr>
<tr>
<td>EERE</td>
<td>709</td>
<td>FAA</td>
<td>3485</td>
</tr>
<tr>
<td>DOT</td>
<td>541</td>
<td>FNS</td>
<td>2821</td>
</tr>
<tr>
<td>OCC</td>
<td>466</td>
<td>FSIS</td>
<td>2436</td>
</tr>
<tr>
<td>FMCSA</td>
<td>451</td>
<td>APHIS</td>
<td>2232</td>
</tr>
<tr>
<td>IRS</td>
<td>444</td>
<td>HUD</td>
<td>1910</td>
</tr>
<tr>
<td>NLRB</td>
<td>366</td>
<td>IRS</td>
<td>1733</td>
</tr>
<tr>
<td>USTR</td>
<td>336</td>
<td>CFPB</td>
<td>1361</td>
</tr>
<tr>
<td>CFPB</td>
<td>328</td>
<td>AMS</td>
<td>1310</td>
</tr>
<tr>
<td>EBSA</td>
<td>302</td>
<td>OSHA</td>
<td>1192</td>
</tr>
<tr>
<td>HHS</td>
<td>276</td>
<td>FHWA</td>
<td>1095</td>
</tr>
<tr>
<td>USCG</td>
<td>222</td>
<td>SSA</td>
<td>1064</td>
</tr>
<tr>
<td>FWS</td>
<td>208</td>
<td>NHTSA</td>
<td>1001</td>
</tr>
<tr>
<td>AMS</td>
<td>181</td>
<td>EEERE</td>
<td>936</td>
</tr>
<tr>
<td>HUD</td>
<td>163</td>
<td>DOT</td>
<td>925</td>
</tr>
<tr>
<td>APHIS</td>
<td>152</td>
<td>BOEM</td>
<td>909</td>
</tr>
<tr>
<td>FSIS</td>
<td>144</td>
<td>ICEB</td>
<td>861</td>
</tr>
<tr>
<td>TSA</td>
<td>129</td>
<td>DOJ</td>
<td>824</td>
</tr>
<tr>
<td>FRA</td>
<td>109</td>
<td>USCG</td>
<td>750</td>
</tr>
<tr>
<td>FHWA</td>
<td>108</td>
<td>OMB</td>
<td>748</td>
</tr>
<tr>
<td>LMSO</td>
<td>102</td>
<td>FMCSA</td>
<td>708</td>
</tr>
<tr>
<td>BOEM</td>
<td>95</td>
<td>DOS</td>
<td>667</td>
</tr>
<tr>
<td>BIS</td>
<td>94</td>
<td>OPM</td>
<td>649</td>
</tr>
<tr>
<td>EIB</td>
<td>91</td>
<td>NLRB</td>
<td>616</td>
</tr>
</tbody>
</table>

*Notes:* This table reports the 30 top agencies as ranked by the number of comments they receive by firms (first two columns) or by grantees (last two columns).
Table A.2: Co-commenting in time-varying sample - Contemporaneous and lagged donations

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Firm $f$ and grantee $g$ commented on the same regulation in year $t \times 100$</th>
<th>0.163</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>(1)</td>
</tr>
<tr>
<td>Firm $f$ contributed to grantee $g$ in year $t$</td>
<td>0.746*** (0.040)</td>
<td>0.614*** (0.041)</td>
</tr>
<tr>
<td>Firm $f$ contributed to grantee $g$ in year $t-1$</td>
<td>0.964*** (0.042)</td>
<td>0.819*** (0.044)</td>
</tr>
</tbody>
</table>

Fixed effects

- Year: Y
- Grantee: Y
- Donor: Y
- Grantee-Firm Pair: Y

SE Clusters: Grantee Firm Grantee Firm × Grantee

Observations: 125,918,520 125,918,520 125,918,520 125,918,520 125,860,865

Note: The dependent variable is equal to 100 if grantee $g$ and firm $f$ comment on the same regulation in year $t$. The independent variable is equal to one if grantee $g$ received a donation from firm $f$ either at year $t$ (respectively, $t-1$). Standard errors are clustered at the level indicated in each column under “SE Clusters”: *** $p<0.01$, ** $p<0.05$, * $p<0.1$
Table A.3: Co-commenting in time-varying sample - Future donations

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Firm $f$ and grantee $g$ commented on the same regulation in year $t \times 100%$</th>
<th>Mean</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.163</td>
<td>0.557***</td>
<td>0.452***</td>
<td>0.447***</td>
<td>0.339***</td>
<td>-0.016</td>
</tr>
<tr>
<td>Firm $f$ contributed to grantee $g$ in year $t + 1$</td>
<td></td>
<td></td>
<td>(0.038)</td>
<td>(0.049)</td>
<td>(0.081)</td>
<td>(0.087)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>Firm $f$ contributed to grantee $g$ in year $t$ or $t - 1$</td>
<td></td>
<td></td>
<td>(0.032)</td>
<td>(0.051)</td>
<td>(0.098)</td>
<td>(0.104)</td>
<td>(0.040)</td>
</tr>
</tbody>
</table>

Fixed effects

<table>
<thead>
<tr>
<th></th>
<th>Year</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grantee</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Donor</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grantee-Firm Pair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SE Clusters

<table>
<thead>
<tr>
<th></th>
<th>Grantee</th>
<th>Firm</th>
<th>Grantee + Firm</th>
<th>Firm $\times$ Grantee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>125,918,520</td>
<td>125,918,520</td>
<td>125,918,520</td>
<td>125,918,520</td>
</tr>
</tbody>
</table>

Notes: Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** p<0.01, ** p<0.05, * p<0.1
Table A.4: Similarity - Contemporaneous and lagged donations

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Similarity of comments by grantee $g$ and firm $f$ on same regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Grantee $g$ received donation from firm $f$ at $t$</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
</tr>
<tr>
<td>Grantee $g$ received donation from firm $f$ at $t-1$</td>
<td>0.169***</td>
</tr>
<tr>
<td></td>
<td>(0.054)</td>
</tr>
</tbody>
</table>

Fixed Effects

| Docket | Y |     |
| Firm   | Y | Y   |
| Grantee| Y | Y   |

SE Clusters

<table>
<thead>
<tr>
<th>Docket</th>
<th>Docket</th>
<th>Firm+Grantee</th>
<th>Firm+Grantee+Docket</th>
</tr>
</thead>
</table>

| Observations | 301,602 | 301,602 | 300,817 | 300,792 |

Notes: The dependent variable is a similarity index between the comment of firm $f$ and the comment of grantee $g$ on regulation $r$, divided by the standard deviation of similarity of all comments relative to $r$. The independent variable is equal to one if grantee $g$ received a donation from firm $f$ in the year when the comment appears (respectively, the year before). Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** $p<0.01$, ** $p<0.05$, * $p<0.1$
Table A.5: Similarity - Sector Control

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Similarity of comments by grantee $g$ and firm $f$ on same regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Grantee $g$ received donation from firm $f$ at $t$ or $t-1$</td>
<td>0.074***</td>
</tr>
<tr>
<td>Grantee $g$ received donation from firm $f$ at $t$</td>
<td>0.020</td>
</tr>
<tr>
<td>Grantee $g$ received donation from firm $f$ at $t-1$</td>
<td>0.072**</td>
</tr>
</tbody>
</table>

**Fixed Effects**
- Docket: Y Y Y Y
- Firm: Y Y Y Y
- Grantee: Y Y Y Y
- NAICS code $\times$ NTEEC code: Y Y

**SE Clusters**
- Firm: Y Y Y Y
- Grantee: Y Y Y Y
- Docket: Y Y Y Y
- NAICS code $\times$ NTEEC code: Y Y

Observations: 162,735 162,735 162,735 162,735

Notes: The dependent variable is a similarity index between the comment of firm $f$ and the comment of grantee $g$ on regulation $r$, divided by the standard deviation of similarity of all comments relative to $r$. The independent variable is equal to one if grantee $g$ received a donation from firm $f$ in the year when the comment appears or the year before. Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** p<0.01, ** p<0.05, * p<0.1
Table A.6: Similarity - Future Donation

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Similarity of comments by grantee $g$ and firm $f$ on same regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Grantee $g$ received donation from firm $f$ at $t + 1$</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td>(0.050)</td>
</tr>
<tr>
<td>Grantee $g$ received donation from firm $f$ at $t$ or $t − 1$</td>
<td>0.069***</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
</tr>
</tbody>
</table>

Fixed Effects

- Firm: Y Y Y Y Y Y
- Grantee: Y Y Y Y Y Y
- Docket: Y Y Y Y
- NAICS code $\times$ NTEEC code: Y Y

SE Clustering

- Firm: Y Y Y Y Y Y
- Grantee: Y Y Y Y Y Y
- Docket: Y Y Y Y
- NAICS code $\times$ NTEEC code: Y Y
- Sample with sector codes: Y Y Y Y

Observations: 300,817 300,792 175,660 175,643 162,735 162,735

Notes: Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** p<0.01, ** p<0.05, * p<0.1
Table A.7: Rule-comment similarity - Robustness

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Similarity of rule discussion and comment by firm ( f ) on same regulation ( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>At least one grantee co-commenting and receiving donation from firm ( f ) in year ( t ) or ( t - 1 )</td>
<td>0.118***</td>
</tr>
<tr>
<td>Log number of pages of comments submitted by firm ( f )</td>
<td>0.405***</td>
</tr>
<tr>
<td>At least one grantee co-commenting and receiving donation from firm ( f ) in any year</td>
<td>0.027</td>
</tr>
</tbody>
</table>

Fixed Effects
- Docket: Y Y Y Y
- Firm: Y Y Y Y


Observations | 4,385 | 4,385 | 4,965 | 4,385 |

Notes: The dependent variable is a similarity index between the comment of firm \( f \) and the discussion of regulation \( r \), divided by the standard deviation of similarity of all comments relative to \( r \) and discussion of regulation \( r \). The independent variables are the same as in tables 2.5 and 2.6. Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** \( p < 0.01 \), ** \( p < 0.05 \), * \( p < 0.1 \).
Table A.8: Citations of Firm in Rule Discussion

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Citation of firm f’s name in Discussion of rule r</th>
<th>(1) Cited(Y/N)</th>
<th>(2) Log (1+Citations)</th>
<th>(3) Cited(Y/N)</th>
<th>(4) Log(1+Citations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>At least one grantee g co-commenting and receiving donation from firm f in year t or t – 1</td>
<td></td>
<td>0.017</td>
<td>0.043*</td>
<td>0.021</td>
<td>0.059**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.013)</td>
<td>(0.024)</td>
<td>(0.014)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>At least one grantee g co-commenting and receiving donation from firm f in any year</td>
<td></td>
<td>-0.005</td>
<td>-0.032</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.011)</td>
<td>(0.022)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log number of pages of comments submitted by firm f</td>
<td></td>
<td></td>
<td></td>
<td>0.028***</td>
<td>0.049***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.007)</td>
<td>(0.013)</td>
</tr>
</tbody>
</table>

Fixed Effects

- Docket: Y Y Y Y
- Firm: Y Y Y Y


Observations: 4,965 4,965 4,385 4,385

Note: Standard errors are clustered at the level indicated in each column under “SE Clusters”. *** p<0.01, ** p<0.05, * p<0.1
<table>
<thead>
<tr>
<th>Agency Code</th>
<th>Agency Name</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF</td>
<td>Children and Families Administration</td>
<td>DOI, Interior Department</td>
</tr>
<tr>
<td>AHRQ</td>
<td>Agency for Healthcare Research and Quality</td>
<td>DOJ, Justice Department</td>
</tr>
<tr>
<td>AID</td>
<td>Agency for International Development</td>
<td>DOL, Employment Standards Administration</td>
</tr>
<tr>
<td>AMS</td>
<td>Agricultural Marketing Service</td>
<td>DOS, State Department</td>
</tr>
<tr>
<td>AOA</td>
<td>Aging Administration</td>
<td>DOT, Transportation Department</td>
</tr>
<tr>
<td>APHIS</td>
<td>Animal and Plant Health Inspection Service</td>
<td>EAB, Economic Analysis Bureau</td>
</tr>
<tr>
<td>ARS</td>
<td>Agricultural Research Service</td>
<td>EAC, Election Assistance Commission</td>
</tr>
<tr>
<td>ASC</td>
<td>Appraisal Subcommittee</td>
<td>EBSA, Employee Benefits Security Administration</td>
</tr>
<tr>
<td>ATBCB</td>
<td>Architectural and Transportation Barriers Compliance Board</td>
<td>ED, Education Department</td>
</tr>
<tr>
<td>ATF</td>
<td>Alcohol, Tobacco, Firearms, and Explosives Bureau</td>
<td>EDA, Economic Development Administration</td>
</tr>
<tr>
<td>ATSDR</td>
<td>Agency for Toxic Substances and Disease Registry</td>
<td>EEOC, Equal Employment Opportunity Commission</td>
</tr>
<tr>
<td>BIA</td>
<td>Indian Affairs Bureau</td>
<td>EERE, Off. Energy Efficiency and Renewable Energy</td>
</tr>
<tr>
<td>BIS</td>
<td>Industry and Security Bureau</td>
<td>EIB, Import Export Bank of the United States</td>
</tr>
<tr>
<td>BLM</td>
<td>Land Management Bureau</td>
<td>EOIR, Executive Office for Immigration Review</td>
</tr>
<tr>
<td>BOEM</td>
<td>Ocean Energy Management Bureau</td>
<td>EPA, Environmental Protection Agency</td>
</tr>
<tr>
<td>BOP</td>
<td>Prisons Bureau</td>
<td>ESA, Employment Standards Administration</td>
</tr>
<tr>
<td>BOR</td>
<td>Reclamation Bureau</td>
<td>ETA, Employment and Training Administration</td>
</tr>
<tr>
<td>BPD</td>
<td>Public Debt Bureau</td>
<td>FAA, Federal Aviation Administration</td>
</tr>
<tr>
<td>BSEE</td>
<td>Safety and Environmental Enforcement Bureau</td>
<td>FARS, Federal Acquisition Regulation System</td>
</tr>
<tr>
<td>CCC</td>
<td>Commodity Credit Corporation</td>
<td>FBI, Federal Bureau of Investigation</td>
</tr>
<tr>
<td>CDC</td>
<td>Centers for Disease Control and Prevention</td>
<td>FCIC, Federal Crop Insurance Corporation</td>
</tr>
<tr>
<td>CDFI</td>
<td>Community Development Financial Institutions Fund</td>
<td>FDA, Food and Drug Administration</td>
</tr>
<tr>
<td>CFPB</td>
<td>Consumer Financial Protection Bureau</td>
<td>FEMA, Federal Emergency Management Agency</td>
</tr>
<tr>
<td>CMS</td>
<td>Centers for Medicare Medicaid Services</td>
<td>FFIEC, Federal Financial Institutions Exam. Council</td>
</tr>
<tr>
<td>CNCS</td>
<td>Corporation for National and Security Service</td>
<td>FHWA, Federal Highway Administration</td>
</tr>
<tr>
<td>COE</td>
<td>Engineers Corps</td>
<td>FINCEN, Financial Crimes Enforcement Network</td>
</tr>
<tr>
<td>COLC</td>
<td>U.S. Copyright Office, Library of Congress</td>
<td>FISCAL, Bureau of the Fiscal Service</td>
</tr>
<tr>
<td>CPSC</td>
<td>Consumer Product Safety Commission</td>
<td>FMCSA, Federal Motor Carrier Safety Administration</td>
</tr>
<tr>
<td>CSREES</td>
<td>Coop. State Research, Education, and Extension Service</td>
<td>FNS, Food and Nutrition Service</td>
</tr>
<tr>
<td>DARS</td>
<td>Defense Acquisition Regulations System</td>
<td>FRA, Federal Railroad Administration</td>
</tr>
<tr>
<td>DEA</td>
<td>Drug Enforcement Administration</td>
<td>FS, Fiscal Service</td>
</tr>
<tr>
<td>DHS</td>
<td>Homeland Security Department</td>
<td>FSA, Farm Service Agency</td>
</tr>
<tr>
<td>DOC</td>
<td>Commerce Department</td>
<td>FSIS, Food Safety and Inspection Service</td>
</tr>
<tr>
<td>DOD</td>
<td>Defense Department</td>
<td>FSOC, Financial Stability Oversight Council</td>
</tr>
<tr>
<td>DOE</td>
<td>Energy Department</td>
<td>FTA, Federal Transit Administration</td>
</tr>
</tbody>
</table>
Table A.10: List of Agencies on regulations.gov (F-Z)

<table>
<thead>
<tr>
<th>Agency Name</th>
<th>Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTC</td>
<td>Federal Trade Commission</td>
</tr>
<tr>
<td>FWS</td>
<td>Fish and Wildlife Service</td>
</tr>
<tr>
<td>GIPSA</td>
<td>Grain Inspection, Packers and Stockyards Adm.</td>
</tr>
<tr>
<td>GSA</td>
<td>General Services Administration</td>
</tr>
<tr>
<td>HHS</td>
<td>Health and Human Services Department</td>
</tr>
<tr>
<td>HHSIG</td>
<td>Inspector General, Health and Human Serv Dept</td>
</tr>
<tr>
<td>HRSA</td>
<td>Health Resources and Services Administration</td>
</tr>
<tr>
<td>HUD</td>
<td>Housing and Urban Development Department</td>
</tr>
<tr>
<td>ICEB</td>
<td>Immigration and Customs Enforcement Bureau</td>
</tr>
<tr>
<td>IHS</td>
<td>Indian Health Service</td>
</tr>
<tr>
<td>IRS</td>
<td>Internal Revenue Service</td>
</tr>
<tr>
<td>ITA</td>
<td>International Trade Administration</td>
</tr>
<tr>
<td>LMSO</td>
<td>Labor-Management Standards Office</td>
</tr>
<tr>
<td>MARAD</td>
<td>Maritime Administration</td>
</tr>
<tr>
<td>MMS</td>
<td>Minerals Management Service</td>
</tr>
<tr>
<td>MSHA</td>
<td>Mine Safety and Health Administration</td>
</tr>
<tr>
<td>NHTSA</td>
<td>National Highway Traffic Safety Administration</td>
</tr>
<tr>
<td>NIFA</td>
<td>National Institute of Food and Agriculture</td>
</tr>
<tr>
<td>NIGC</td>
<td>National Indian Gaming Commission</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>NIST</td>
<td>National Institute of Standards and Technology</td>
</tr>
<tr>
<td>NLRB</td>
<td>National Labor Relations Board</td>
</tr>
<tr>
<td>NOAA</td>
<td>National Oceanic and Atmospheric Administration</td>
</tr>
<tr>
<td>NPS</td>
<td>National Park Service</td>
</tr>
<tr>
<td>NRC</td>
<td>Nuclear Regulatory Commission</td>
</tr>
<tr>
<td>NRCS</td>
<td>Natural Resources Conservation Service</td>
</tr>
<tr>
<td>NSF</td>
<td>National Science Foundation</td>
</tr>
<tr>
<td>NTIA</td>
<td>National Telecommunications and Information Adm.</td>
</tr>
<tr>
<td>NTSB</td>
<td>National Transportation Safety Board</td>
</tr>
<tr>
<td>OCC</td>
<td>Comptroller of the Currency</td>
</tr>
<tr>
<td>OFAC</td>
<td>Foreign Assets Control Office</td>
</tr>
<tr>
<td>OFCCP</td>
<td>Federal Contract Compliance Programs Office</td>
</tr>
<tr>
<td>OFPP</td>
<td>Federal Procurement Policy Office</td>
</tr>
<tr>
<td>OJJDP</td>
<td>Juvenile Justice and Delinquency Prevention Office</td>
</tr>
<tr>
<td>OJP</td>
<td>Justice Programs Office</td>
</tr>
<tr>
<td>OMB</td>
<td>Management and Budget Office</td>
</tr>
<tr>
<td>ONRR</td>
<td>Natural Resources Revenue Office</td>
</tr>
<tr>
<td>OPM</td>
<td>Personnel Management Office</td>
</tr>
<tr>
<td>OPPM</td>
<td>Procurement and Property Management, Office of</td>
</tr>
<tr>
<td>OSHA</td>
<td>Occupational Safety and Health Administration</td>
</tr>
<tr>
<td>OSM</td>
<td>Surface Mining Reclamation and Enforcement Office</td>
</tr>
<tr>
<td>OTS</td>
<td>Thrift Supervision Office</td>
</tr>
<tr>
<td>PBGC</td>
<td>Pension Benefit Guaranty Corporation</td>
</tr>
<tr>
<td>PCLOB</td>
<td>Privacy and Civil Liberties Oversight Board</td>
</tr>
<tr>
<td>PHMSA</td>
<td>Pipeline and Hazardous Materials Safety Adm.</td>
</tr>
<tr>
<td>PTO</td>
<td>Patent and Trademark Office</td>
</tr>
<tr>
<td>RBS</td>
<td>Rural Business-Cooperative Service</td>
</tr>
<tr>
<td>RHS</td>
<td>Rural Housing Service</td>
</tr>
<tr>
<td>RITAT</td>
<td>Research and Innovative Technology Administration</td>
</tr>
<tr>
<td>RUS</td>
<td>Rural Utilities Service</td>
</tr>
<tr>
<td>SAMHSA</td>
<td>Substance Abuse and Mental Health Services Adm.</td>
</tr>
<tr>
<td>SBA</td>
<td>Small Business Administration</td>
</tr>
<tr>
<td>SLSDC</td>
<td>Saint Lawrence Seaway Development Corporation</td>
</tr>
<tr>
<td>SSA</td>
<td>Social Security Administration</td>
</tr>
<tr>
<td>TREA</td>
<td>Treasury Department</td>
</tr>
<tr>
<td>TSA</td>
<td>Transportation Security Administration</td>
</tr>
<tr>
<td>TTB</td>
<td>Alcohol and Tobacco Tax and Trade Bureau</td>
</tr>
<tr>
<td>USC</td>
<td>United States Courts</td>
</tr>
<tr>
<td>USCBP</td>
<td>U.S. Customs and Border Protection</td>
</tr>
<tr>
<td>USCPR</td>
<td>Coast Guard</td>
</tr>
<tr>
<td>USCIS</td>
<td>U.S. Citizenship and Immigration Services</td>
</tr>
<tr>
<td>USDA</td>
<td>Agriculture Department</td>
</tr>
<tr>
<td>USPC</td>
<td>Parole Commission</td>
</tr>
<tr>
<td>USTR</td>
<td>Trade Representative, Office of United States</td>
</tr>
<tr>
<td>VA</td>
<td>Veterans Affairs Department</td>
</tr>
<tr>
<td>VETS</td>
<td>Veterans Employment and Training Service</td>
</tr>
<tr>
<td>WCPO</td>
<td>Workers Compensation Programs Office</td>
</tr>
<tr>
<td>WHD</td>
<td>Wage and Hour Division</td>
</tr>
</tbody>
</table>

100