Symposium Article

Social Science Computer Review
2014, Vol. 32(2) 182-194
© The Author(s) 2013
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0894439313506847
ssc.sagepub.com
SAGE

Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry

Karen E. C. Levy[1] and Michael Franklin[1]

Abstract

Public comments submitted during agency rulemakings can provide rich insight into stakeholders' viewpoints around contentious political issues but have been largely untapped as a data source by social scientists. This is in part due to the lack of access to comments in machine-readable formats and in part due to the difficulty in analyzing large corpora of textual data. However, new online repositories and analytic methodologies are beginning to open up this trove of data for researchers. Using data from the online portal *regulations.gov*, we employ probabilistic topic modeling to identify latent themes in a series of regulatory debates about electronic monitoring in the U.S. trucking industry. Our model suggests that different types of commenters use alternative discursive frames in talking about monitoring. Comments submitted by individuals were more likely to place the electronic monitoring debate in the context of broader logistical problems plaguing the industry, such as long wait times at shippers' terminals, while organizational stakeholders were more likely than individuals to frame their comments in terms of technological standards and language suggesting cost / benefit quantification.

Keywords

topic modeling, latent Dirichlet allocation, e-rulemaking, public comments, trucking

Introduction

Agency rulemaking provides a unique arena in which any member of the public can submit substantive feedback to an agency while it is actively formulating new policies. Public comments provide rich insight into stakeholders' viewpoints about contentious political issues, but have been largely untapped as a data source by social scientists, for two reasons. First, until the last decade, public comments were not accessible in easily analyzable formats, as comments were typically submitted to agencies by mail and only viewable via a trip to the National Archives or an agency docket room. Second, the same characteristics that make these comments a rich arena for political participation create challenges for

[1] Princeton University, Princeton, NJ, USA

Corresponding Author:
Karen E. C. Levy, Department of Sociology, Princeton University, 106 Wallace Hall, Princeton, NJ 08544, USA.
Email: kelevy@princeton.edu

researchers who wish to analyze them. Document corpora can be extremely large (in some cases, thousands of documents for a particularly controversial rulemaking), making them unwieldy for qualitative analysis. And because comments consist of unstructured textual data, they are more difficult to measure and quantify than other forms of political participation, like voting in elections.

In light of these difficulties, few previous studies of regulatory comments have used quantitative methodologies to analyze this political forum. But new online repositories and analytic methods are beginning to open up this trove of data for researchers. The online comment portal *regulations.gov* facilitates participation in rulemaking processes by allowing comments to be submitted and viewed online; in combination with the aggregation efforts of pro-transparency organizations, these data are newly available to researchers in largely machine-readable formats. And new methods that facilitate the analysis of large collections of documents hold much promise—in particular, probabilistic topic modeling (Blei, 2011; Blei, Ng, & Jordan, 2003) allows researchers to systematically and inductively identify key latent themes in a text corpus.

We use online comment data and topic modeling strategies to investigate 25 years of regulatory debates around the use of electronic monitoring systems in the U.S. long-haul trucking industry. Electronic monitoring is hugely contentious within the trucking community and has engendered vigorous debate among stakeholders around issues like safety and privacy. We use topic models to uncover thematic patterns in public comments on the proposed regulations. In addition, by supplementing the model with covariates labeling commenter identity, we identify systematic differences among the interests and evaluative principles that different groups of stakeholders emphasize.

Public Participation in Rulemaking

Federal agencies are required to seek and consider public input on most new proposed regulations via "notice-and-comment" procedures, in accordance with the Administrative Procedure Act (5 U.S.C. § 553). An agency officially notifies the public of a new proposal by publishing it in the Federal Register; this publication triggers a period of at least 30 days during which anyone—including individuals, interest groups, corporations, or other government entities—can submit any comment to the agency regarding the proposal. (In some cases, rules go through multiple rounds of notice-and-comment.) The agency is then required to consider these comments in formulating a final rule.

Traditionally, a party interested in submitting a comment needed to find the published notice in the Federal Register and mail the comment to the agency—a high bar for action that effectively dissuaded most everyday citizens from participating in a meaningful way (Coglianese, 2006; Mendelson, 2011). The federal government began efforts to make commenting available online in the 1990s and launched *regulations.gov* in 2003. The site allows for the submission and viewing of comments online and is intended to open up the regulatory process by facilitating greater public access. (Whether these "e-rulemaking" efforts have actually increased meaningful participation, however, is a matter of considerable debate [Coglianese, 2006; Cuéllar, 2005; Shulman, Schlosberg, Zavestoski, & Courard-Hauri, 2003].)

The comment process is a unique form of political participation from both agency and public perspectives. To the agency, notice-and-comment is largely a data-gathering mission. The process ostensibly permits agencies to draw on interested parties' expertize for rules that may be highly technical or detailed. Topics of regulation are often very specific and fine-grained, so it is to agencies' advantage to have interested parties provide them with relevant research (but see Wagner, 2010 for a critical take on information exchange in rulemaking). In addition, agencies often have close enforcement relationships with the industries most closely affected by new regulations, so it is advantageous to foresee potential problems and "take the temperature" of a field before a rule is passed.

From a public perspective, the comment process offers a unique forum in which to provide feedback to a governing body while it is formulating new rules. Unlike electoral participation or direct democracy measures, regulatory commenting allows for regular, unstructured, substantive participation in government in a formalized arena.

Substantial critiques exist regarding how effective commenting is as a tool for influencing policy as well as the extent to which comment processes may be "captured" by resource- and expertise-rich corporations and special interests rather than everyday individuals. We do not focus on these evaluations of rulemaking as a political process, which have been ably addressed by other researchers (Coglianese, 2004a; Cuéllar, 2005; Golden, 1998; Wagner, 2010; Yackee & Yackee, 2006). Rather, we make use of rulemaking as a data site. Despite their debatable efficacy, public comments undeniably provide unique insight into the political concerns, values, and frames articulated by participating stakeholders around contentious political issues—and are newly accessible to researchers, thanks to their recent online availability.

Topic Modeling for Analysis of Textual Political Data

Rich data sources demand rich methodological strategies for their analysis. Previous analyses of public comment data have generally concerned the *process* of rulemaking, and are most commonly critiques of the effectiveness of public comment on substantive regulatory outcomes. We take a different approach, tapping comments as a data source for understanding alternative framings of contentious political issues.

We turn to probabilistic topic modeling in order to analyze a large corpus of comment text. Topic modeling is an exploratory strategy that allows us to detect the latent thematic structure of a set of documents (Blei, 2011). In our analysis, we use latent Dirichlet allocation (LDA), a simple type of probabilistic topic model (Blei et al., 2003).

LDA generates "topics" as lists of words drawn from the vocabulary used in the text corpus; the topic is based on the distributions of those words over the vocabulary. The topics are generated inductively by the model based on the likelihood of words to co-occur within documents. LDA produces the topics through a probabilistic approximation of Bayesian inference. Starting with a set of seed topics (often randomly generated), the algorithm iteratively alters these topics to best match the set of data being learned. LDA also generates proportions for each document for each topic, so that each document can be described as being proportionally composed of (or, interpretively speaking, "about") a number of topics that are expressed by the words used in that document. For details about the algorithmic and computational aspects of LDA, see Blei et al. (2003).

Topic modeling is a good match for a data source like public comments. Because the procedure is automated, it enables analysis of much larger text corpora than would be feasible by hand. Since topics are generated inductively by the model and not predefined by the researcher, the technique protects against implicit coding bias caused by the constraints of researcher knowledge. And because the method assumes that a single document can contain multiple topics (in contrast to some other document clustering methods; see Blei & Lafferty, 2009), it enables researchers to draw insights from interrelations among themes, both within documents and within the dataset as a whole.

As described, public comments are rich data because commenters express all manner of concerns in unstructured ways. Comments run the gamut from technical specifications to personal stories and from thoughtful reflection to threats and name-calling. Many comments defy easy categorization as being for or against the proposal at issue. Thus, hand-coding such documents can be a particularly difficult task; topic modeling appeals because it can uncover hidden patterns in even a varied set of documents. To our knowledge, ours is the first study making use of topic modeling to analyze public comment data. Previous analyses have commonly relied on hand-coding (Cuéllar, 2005; Golden, 1998; Krawiec, 2013; Wagner, Barnes, & Peters, 2011; Yackee & Yackee, 2006); however, see Kwon, Shulman, and Hovy (2006), Cardie, Farina, Aijaz, Rawding, and Purpura (2008), and Lau, Law, and Wiederhold (2005) for notable discussions of other machine-learning approaches to comment data.

Electronic Monitoring of Truckers' Work Hours

We apply topic modeling to public comments stemming from a regulatory debate around the use of electronic monitoring systems in the U. S. long-haul trucking industry. In this section, we provide context for this debate. Our knowledge about the electronic monitoring debate in the trucking industry comes from the first author's intensive qualitative research on the subject, which has included multisited ethnographic fieldwork, extensive review of regulatory and technical documents and media sources, and over 50 interviews with individuals associated with the trucking industry (including drivers, firm representatives, fleet managers, technology vendors, and regulators).

According to federal regulations, long-haul truck drivers in the United States may work a limited number of hours each day; these restrictions are a safety measure intended to reduce the risk of accidents caused by overtired drivers.[1] Drivers have traditionally been required to keep track of their work hours via paper logbooks, which may be spot-checked by law enforcement officers at weigh stations or at roadside.

However, truckers have long evaded the timekeeping rules by keeping inaccurate records of their work time. Paper logbooks are relatively easy to "fudge"—enforcement is inconsistent, and it is difficult and time consuming for an inspector to disprove their contents. Drivers are typically paid by miles driven and thus have incentive to drive more than is permitted (Belzer, 2000). And, because trucking companies and customers emphasize and often incentivize on-time service, drivers face pressure to get goods to their location on time despite the contingencies of highway travel like weather, traffic, and mechanical breakdowns.

In response to this problem, the Federal Motor Carriers Safety Administration (FMCSA), the federal agency responsible for overseeing commercial vehicle safety, has been pursuing the use of electronic monitoring of truckers' work time in place of paper logbooks. The device under consideration, called an electronic on-board recorder (EOBR), attaches directly to the truck and creates an electronic record of when the truck is being driven. Many EOBRs are already in use on the road today, as a number of large trucking firms use them as part of broader fleet management systems to monitor and communicate with their employees while they are on the road.

Since 1987, four major rulemakings concerning EOBRs have been proposed. A 1987 action permitted voluntary use of EOBRs for compliance with the hours-of-service (HOS) rules; in 2003, a mandate for EOBRs in all trucks was proposed but ultimately not adopted (for reasons including significant doubts about the efficacy and costs of a mandate). A 2004 rule mandated EOBRs for carriers with a high rate of noncompliance with HOS rules, and a 2010 proposal (currently under consideration) again seeks to mandate the installation and use of EOBRs in all trucks.

EOBRs are very controversial in the trucking industry, and these rulemakings have generated a great deal of disagreement among stakeholders. Large organizational interests (like insurance groups, public safety coalitions, and many large trucking firms) tend to support their use, based on the assumption that they will increase compliance with HOS rules. But many truck drivers and small companies ardently oppose EOBRs; they object to the costs of the devices and resent the implication that they are not trustworthy. As an occupational culture, truckers value their independence and autonomy, and many consider the so-called "black boxes" an infringement on their privacy.

The electronic monitoring debate is an ideal site for examining political contention, as the proposed regulations attracted comment from a wide variety of parties, motivated by very different concerns. Divergent frames characterize what the EOBR debate is "about" for different stakeholders, and what evaluative principles they use to understand and form opinions about the regulations at issue (Boltanski & Thévenot, 1999). Because the issue is framed in many different ways, the parties tend to "talk past" one another. In addition, because many individuals (primarily truck drivers) submitted comments, the debate allows us to see differences in rhetoric between individuals and the organized interests that often dominate regulatory comment processes.

Research Approach

Obtaining and Processing the Data

Despite the fact that comments are newly available online, a number of processing steps were required to assemble our dataset and prepare it for analysis. As described, most comments are viewable via the *regulations.gov* online portal, which also permits bulk download of comment metadata for a particular rulemaking (e.g., commenter name, date of submission, etc.)—but not, at this time, bulk download of the text of the comments themselves. Fortunately, the Sunlight Foundation—a nonprofit organization devoted to using technology to increase government transparency—produced a web scraper that allowed us to bulk download the text of comments from the 1997, 2004, and 2010 rulemakings. Most documents were already in machine-readable format, either because they were entered as text into the website comment field directly or because they were uploaded in .doc or .pdf formats and digitized by Sunlight using optical character recognition (OCR). Approximately 200 comments, most of which were handwritten submissions mailed to the agency, had not been digitized; these comments were hand-transcribed.

The 1997 proposed regulation was a broad proposal that included provisions about EOBRs as well as other aspects of trucking policy (e.g., proposed changes to the number of hours truck drivers could drive). That proposal generated a good deal of public feedback (about 23,000 written comments), much of which was irrelevant to the EOBR provision we were interested in. In order to filter this docket to relevant comments only, we generated a list of 59 “inclusion words” (“EOBR,” “logbook,” brand names of EOBR manufacturers, etc.) derived from our reading of other comments on the issue; only comments including one or more of these words were included in the analysis. Based on our review of the results, we believe this filter process was over- rather than underinclusive (i.e., included some irrelevant comments in the data set rather than omitting relevant comments).

For completeness, we also wanted to include comments from the 1987 rulemaking, which predated *regulations.gov* and were not available online. These data were obtained by scanning these documents manually at the National Archives in College Park, Maryland; they were then digitized using OCR and hand-transcription.

With our data assembled, we performed two additional preprocessing steps. First, we excluded all documents from the dockets that were not identified as “comments” (e.g., regulatory evaluations performed by the FMCSA). Second, some documents were labeled as being from “multiple submitters”—these were typically multiple discrete comments that had been combined by FMCSA into one (ostensibly for administrative purposes). These were excluded from our analysis, since having multiple unconnected authors of a single “document” would violate the underlying assumptions of the model.

We performed four main processing steps on the remaining data to create the “dictionary” for the topic model. First, we stemmed all of the words (reducing, for instance, *play*, *plays*, and *played* to a single word). Next, we filtered out common “stop words” (*the*, *a*, and so on). We also removed words that were only a single letter or contained only digits. These are common artifacts of errors in the OCR process or the formatting of mailing addresses but are not relevant to the content of the comments. Finally, we restricted the dictionary to the 10,000 most common words. This reduces “noise” in the analysis while eliminating things like unique typos.

As we have described, one advantage of topic modeling is that it inductively exposes latent themes without a researcher having to prespecify what those themes are. However, as is the case with any methodological strategy, topic modeling requires the researcher to make a number of analytical choices in setting up and interpreting the model. For us, these include things like the number of topics in the model, parameters (such as α) affecting the sensitivity of the model, and the point at which the model is considered to have stabilized before it returns results. Because the goal of topic

Table 1. Unsupervised Eight-Topic Model. Table 1 displays the 40 highest-ranked words for each topic. Words were "stemmed" in the model (e.g., *propose*, *proposes*, and *proposal* are treated as the same word, *propos*, for analysis) but have been rewritten as full words here for clarity when applicable. α for this model was set to .01.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
sleep	utility	fatigue	eobr	propose	propose	electronic	propose
work	work	duty	carrier	cost	construction	company	day
shift	regulate	study	require	carrier	industry	propose	work
fatigue	operate	safety	system	operate	duty	eobr	rest
day	propose	period	data	safety	safety	address	make
perform	vehicle	vehicle	motor	regulate	period	make	park
schedule	safety	crash	vehicle	require	day	safety	road
night	exempt	accident	compliance	industry	attach	work	home
study	emergency	carrier	hos	increase	work	log	company
operate	power	rest	duty	addition	transport	industry	load
effect	electric	motor	operate	motor	concrete	request	year
period	employee	fhwa	device	dot	limit	support	stop
de	day	report	cost	transport	delivery	pay	week
circadian	require	research	electronic	duty	maximum	september	duty
report	company	highway	safety	fatigue	company	law	attach
test	line	day	status	agency	product	problem	problem
worker	duty	data	log	company	washington	road	sleep
alert	state	operate	technology	benefit	clerk	load	regulate
safety	period	sleep	enforce	impact	road	owner	force
fhwa	worker	work	support	type	december	regulate	run
subject	crew	transport	violate	number	dot	operate	mile
accident	type	factor	regulate	reduce	operate	shipper	require
manage	motor	commercial	standard	result	deliver	hos	accident
transport	hos	effect	location	rest	nation	enforce	box
duty	transport	limit	specific	year	dc	govern(ment)	family
number	restore	perform	unit	percent	nprm	trucker	industry
data	respond	require	issue	estimate	highway	track	area
al	response	risk	perform	effect	street	put	safe
increase	customer	percent	report	state	seventh	cost	people
regulate	commercial	schedule	include	small	sw	people	break
research	include	flight	fleet	work	mix	year	december
factor	industry	measure	rods	accident	increase	box	put
system	call	propose	agency	business	flexible	device	pay
cmv	employ	involve	mandate	day	require	issue	tire(d)
thc	system	regulate	question	administration	room	include	unload
result	accident	passenger	state	nation	regulate	carrier	live
human	federal	bus	manage	highway	project	speed	find
task	part	increase	paper	economic	improve	business	nprm
health	outage	state	manufacture	crash	profession	paid	long
railroad	limit	hos	review	include	business	violate	back

modeling is to explore themes in a corpus, the interpretability of the topics is of paramount importance; thus, we made these choices in order to maximize topic interpretability and, ultimately, to reveal substantively interesting information (Grimmer & Stewart, 2013: 20). This approach prioritizes the utility of the model as a means for discovery in place of statistical measures of model fit, which may be less semantically meaningful (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009; Grimmer & Stewart, 2013).

With our data processed, we performed the LDA analysis, fitting the model to our text corpus of 1.7 million words across 3,569 documents. We ran multiple instances of the model with different numbers of topics, starting from randomly selected seeds, and obtained quite consistent thematic results across runs; for interpretability, we selected a representative 8-topic model reproduced in Table 1.

(The order of topics is not meaningful.) In light of our prior domain knowledge about the trucking industry, the topics are readily recognizable as relating to major issues in the EOBR debate. We provide some interpretations below.

Results

Exploring Themes of the Debate

Topic 4 contains words like *eobr*, *system*, *device*, *electronic*, and *technology* and thus clearly pertains to the monitoring devices at issue in the proposed regulations. Some words reflect discussion about technical specifications for the devices (*data*, *standard*, *manufacture*, *location*), while others suggest the relationship between technology and compliance with HOS rules (*enforce*, *compliance*, *hos*, *log*).

Topic 7 also includes words that refer to the monitoring devices, like *eobr*, *electronic*, *log*, and *device*—but also contains terms that suggest an alternative frame around the issue. Topic 7 contains the telling words *problem*, *track*, and *box*—the last a common term used by truckers to refer to EOBR devices, as in "black box." (For a sense of how these words appear in context, consider this excerpt from a representative comment, "I resent the need for a black box on my truck to track what I am doing. I am a safe and honest driver.")

Other revealing words in Topic 7 include *pay*, *paid*, *load*, and *shipper*. These terms initially appear to be outside the scope of the pending rules as defined by the agency, but they suggest a set of concerns that places the electronic monitoring debate in the context of broader problems plaguing the trucking industry, regarding long waits for trucks to be loaded and unloaded at shippers' and receivers' terminals (called "detention time"). These wait times, sometimes several hours, are often unpaid for drivers and can cause significant problems for HOS compliance, as a driver facing a long delay may run out of available hours to drive to his next destination. Notably, this was also the only topic that included *trucker* as a high-ranking word.

What does electronic monitoring have to do with these industry problems? In some respects, they are discrete issues; the proposed regulations do not concern detention time or driver pay. But our additional qualitative research on this issue suggests that to many truckers, the issues are inextricably linked in two ways. First, many drivers facing long delays at shippers' terminals have traditionally "fudged" their time logs in order to make up for lost time and money, and electronic monitors would remove this flexibility in executing their work. Second, many drivers feel that the electronic monitoring effort treats *them* as "the problem"—as untrustworthy liars—rather than what many see as the root cause, disorganized shippers who delay their passage. A common sentiment among drivers is that EOBRs are used to treat drivers like criminals, rather than professionals who are doing their best to negotiate difficult logistics in the face of countervailing demands. Taken together, then, Topic 4 and Topic 7 demonstrate starkly different frames around what monitoring devices mean: in one sense, the regulatory debate is about the EOBR as a technological device, and the salient terms of debate are over things like technical specifications. In another, the debate is about the EOBR as a tracking device, and the discourse is bound up with much broader concerns about the industry.

Topic 8 also addresses broader trucking industry issues, also containing words like *box*, *load*, *pay*, and *problem*, and adding terms like *unload*, *home*, *family*, and *park*. Another broad industry problem, from truckers' perspectives, is the dearth of safe parking near the highway for drivers who have run out of hours; drivers commonly fudge their logs in order to find somewhere safe to stop for the night. In addition, many are concerned that stricter enforcement of HOS rules would keep them away from their homes for longer periods, since trips are expected to take more time.

Topic 1 and Topic 3 both concern fatigue and sleep, a central motivator for regulating truckers' work time. Though we expected that fatigue would be a prominent theme in the comments, these topics revealed something we did not expect: two different frames for talking about sleep and fatigue. Although both topics contain some of the same high-ranking words (*fatigue*, *sleep*, *period*, *perform*), Topic 3 is much more trucking-specific and reflects the relationship between sleep and road safety (*crash*, *highway*, *risk*, *vehicle*, *carrier*), while Topic 1 reveals a frame that relates sleep to human biology (*circadian*, *human*, *health*, *alert*). The division between themes is not 100% distinct (e.g., the word *accident* appears in both topics), as is to be expected given the fact that all comments were submitted to address a trucking safety issue; yet, the multiple sleep-related topics suggest that alternative frames around the topic of sleep and fatigue exist in the EOBR debate.

Though space constraints prevent us from discussing the remaining topics in detail, they also suggested themes that struck us as plausible and informative. Topic 5 suggests themes having to do with quantification of the costs and benefits of the proposed rules (*cost*, *benefit*, *percent*, *estimate*, *economic*, *increase*, *number*). Topic 2 and Topic 6, with words like *utility/electric/power/outage* and *construction/concrete/mix*, respectively, make sense in light of utility and construction companies' vigorous arguments that their work is distinguishable from general trucking and should be exempted from the regulations; Topic 6 also contained a number of words relating to agency entities and addresses that are likely in part an artifact of the comment data (e.g., address blocks on some comments).

Revealing Differences Between Stakeholder Groups

We also were interested in understanding how different groups of commenters framed their remarks around electronic monitoring. By pairing the themes identified by the topic model with data about commenter identity, we are able to analyze relationships between these identity groups and the language used in comments submitted by those groups. In order to facilitate comparisons between subsets of the data, we coded each comment based on whether the commenter was an individual (e.g., a truck driver or concerned citizen) or an organization (like a trucking company or a technology vendor). This determination was largely based on comment metadata, that is, fields for commenter's first and last name and organization on the *regulations.gov* submission page. In some cases, when metadata were not available or were ambiguous, we looked for identifying information in the text of the comment, based on explicit identification or other characteristics of the comment (e.g., the use of "I").[2]

We selected the organization/individual axis for comparison for three reasons. First, previous research on regulatory processes has noted that organizations and individuals tend to bring different types of concerns to the table in commenting on proposed regulations (Cuéllar, 2005; Yackee & Yackee, 2006; Wagner, 2010). Second, our intuition based on our research in the trucking industry was that truck drivers (who would generally submit as individual commenters) would likely frame electronic monitoring issues differently than large organizational interests. The third reason was practical: because the determination could usually be made using document metadata, it was feasible for us to hand-code this attribute for over 3,500 comments.

Because LDA generates topic proportions for each comment as well as the topics themselves, we were able to discern the degree to which each comment was proportionally "about" each of the topics in the model. For interpretation, a comment would have proportion 0.15 for topic Y if the model assigned 15% of the words in the comment to topic Y based on probabilistic approximation. We determined proportions for each topic within each comment and then were able to make comparisons between subsets of the data.

Figure 1 displays the mean topic proportions for each topic for individual and organizational submitters; all differences are significant to the $p < .0001$ level, except for Topic 1. The graph makes two dynamics clear. First, comments submitted by individuals were dominated by the concerns described in Topics 7 and 8—the topics that included unique terms about [black] boxes and tracking (Topic 7), home and family (Topic 8), and problems around loading and unloading delays (both). On average, an individual comment was 47% comprising Topic 7 and 25% comprising Topic 8. No

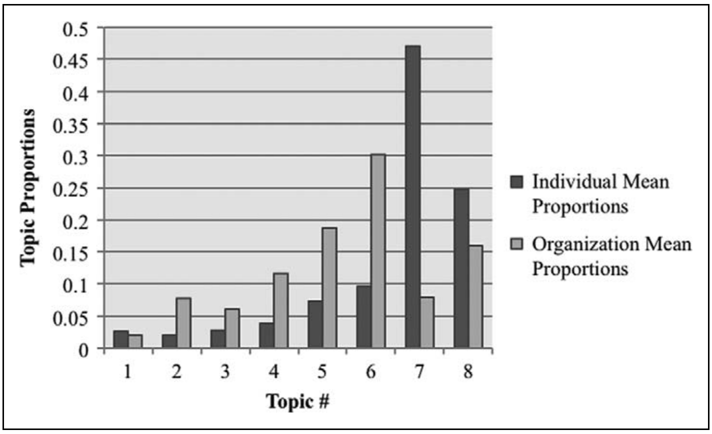

Figure 1. Mean topic proportions for individual and organizational comment submitters ($N = 3531$). Except for Topic 1, all differences in means are statistically significant at $p < .0001$. The mean difference for Topic 1 is statistically significant at $p < .1$.

other topic in the model accounted for more than 10% of the words in the average individual comment. This suggests that these themes were dominant frames through which individuals (a very large number of whom, based on our reading of many of the comments, were truck drivers) viewed the regulatory process around EOBRs. The themes underlying organizational comments, on the other hand, were more evenly distributed across topics.

Comparing the mean proportions for specific topics is also instructive. For instance, contrast Topic 4 (the EOBR "technological" frame) and Topic 7 (EOBRs as tracking boxes, coupled with concerns about pay and shippers). Organizations more commonly employed the technological frame in their comments as compared to individuals (11% mean topic proportion for organizational comments, 4% for individuals) and vice versa for the broader tracking boxes frame (8% mean topic proportion for organizations, 47% for individuals).

Looking at the distribution of proportions for particular topics can also be informative. For instance, Figure 2 displays the distribution of proportions for Topic 5 (quantification and cost/benefit terms) across comments by individuals and organizations. (Note that a large percentage of comments—79% of individual and 50% of organizational comments—have a topic proportion at or close to zero; this was the case for every topic and simply indicates that a large number of documents contain no words that are closely associated with the topic, which is unsurprising given the variety of themes that characterize the corpus.) Beyond these values, we see that the lines are roughly horizontal until about document proportion 0.2 and then slope quickly down toward the x-axis, indicating that very few comments are highly concentrated around Topic 5.

Contrast this with Figure 3, which displays the distribution for Topic 7 (again, the tracking boxes/shipper/pay concerns topic). Here we see a similar pattern for the organization line—but the individuals line slopes sharply upward at its end, indicating that 20% of individuals' comments were highly (more than 90%) concentrated on this topic. This suggests that the frames and concerns represented in Topic 7 were less likely to be "mixed" with others, perhaps indicating that individual commenters using this frame considered it to represent the entirety of their concerns.

Confirming Our Results

As we have mentioned, reliance on statistical measures of model fit may lead to the selection of topic models that are less substantively meaningful (Chang et al., 2009). As such, we do not rely on such

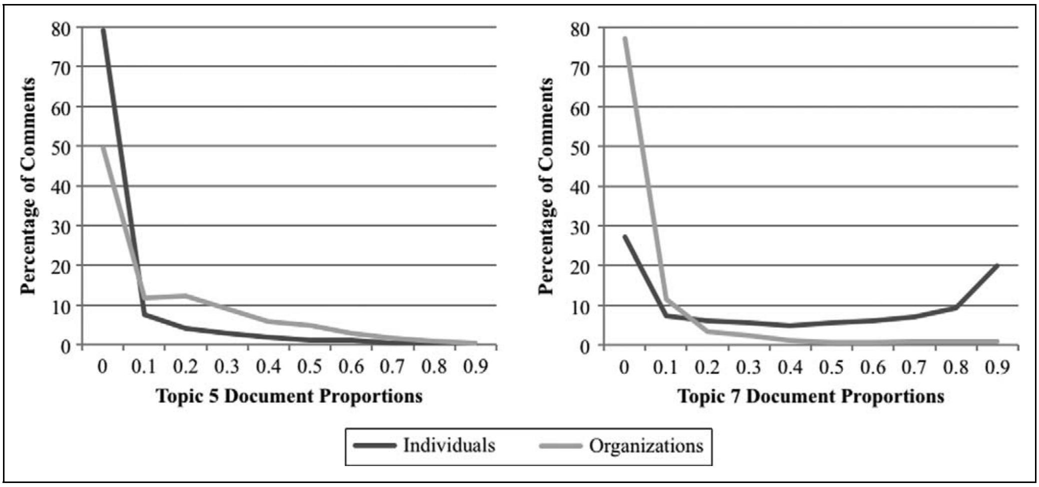

Figure 2. Distribution of Topic 5 proportions across comments ($N = 3{,}531$).
Figure 3. Distribution of Topic 7 proportions across comments ($N = 3{,}531$).

measures here. Instead, we use three “confidence checks” in order to assess the validity and robustness of our findings. First, the salient themes returned by our model are confirmed by qualitative assessment, based on the first author’s multiyear ethnographic, archival, and interview-based research into the electronic monitoring debate in the trucking industry (see Grimmer & Stewart, 2013, for discussion of expert evaluation as a validation tool for text analysis). Second, we ran multiple iterations of the model, starting from different randomly selected seeds and with varying parameters (number of topics, model sensitivity, etc.). Our findings were quite consistent across these iterations.

Third, in order to confirm the robustness of our findings with respect to differences in salient themes between identity groups, we ran secondary models on subsets of the data. For the primary analysis discussed above, we modeled our topics based on the dataset as a whole. In contrast, for this procedure, we first divided the dataset into identity groups (i.e., comments submitted by individuals and comments submitted by organizations), created separate topic models for each subset, and estimated the topic proportions for each document within the subset. We then applied a statistical inference procedure in order to infer the distribution of these topics on the *opposite* subset. In other words, we built a model to fit the set of comments submitted by individuals and applied that model to the comments submitted by organizations (and vice versa). Comparing the document proportions generated by this procedure confirmed our primary findings. For instance, the model built on individuals’ comments contained topics about problems like shippers’ delays and parking; by comparing the estimated and inferred document proportions, we saw that these topics were much less prevalent in comments submitted by organizations.

Discussion

Previous scholars (Coglianese, 2004b; Shulman et al., 2003) have suggested that e-rulemaking might open up exciting opportunities for social scientists to reflect on democratic processes and political participation. We see our project as a response to their call. Newly accessible data sources coupled with innovative methods invite possibilities for new types of political analysis.

Any new form of research faces complications in execution. Here, the processing steps required to analyze comment data are nontrivial, particularly if the researcher is interested in older comments (which are more likely to have been submitted in paper form and would need to be processed using

OCR, or if handwritten, manual transcription). And because *regulations.gov* does not yet permit bulk download of comment text (only metadata), we were dependent on the efforts of Sunlight Labs to bring transparency to government processes by scraping *regulations.gov* and making the results available to us. These complications are likely to be ameliorated in the future as *regulations.gov* continues to improve.

Regulatory comments are submitted to further tactical political aims (to influence the agency toward a particular course of action) and cannot always be considered a pure reflection of interests. Nonetheless, our data provide revealing clues about how stakeholders attempt to shape the discourse around electronic monitoring—what justifications and concerns they pose and how they frame them. Our use of comment data for this purpose is distinct from prior research in that we are not approaching such data principally to make findings about the fairness or efficacy of the regulatory process; rather, we use them to attempt to understand the discursive construction of a particular policy debate.

That said, it is worth noting that some of the topics that arose in our analysis—particularly topics that suggest values (like privacy) rather than quantifiable technical or economic evidence and which situate the electronic monitoring debate in the context of broader industry problems around detention time, driver pay, and parking—are probably less likely to be recognized by an agency as it formulates new policies. In fact, this was indeed the case in the trucking debate: in 2010, the owner-operator drivers' association successfully sued the FMCSA over its failure to address driver privacy and harassment in EOBR rulemakings. Previous research indicates that agencies are far less likely to address "unsophisticated" comments (Cuéllar, 2005) and that value-based information is more difficult for agencies to manage and less likely to have substantive impact on rules (Steelman, 1999). Our analysis here suggests that individuals were more likely to voice such broad and value-based concerns, while the frames used in organizations' submissions were more commonly tailored around quantifiable data and technical specifications.

Topic modeling is a useful and innovative tool for discovering thematic patterns and alternative frames in political discourses, as we found around monitoring devices in our data. Other textual analysis methodologies (like word counts) or researcher-driven categorizations (like coding prespecified categories) are quite likely to miss differences like these. We expect topic modeling to gain traction as a methodological strategy as new online sources make rich textual data increasingly accessible to social scientists.

Acknowledgments

The authors gratefully acknowledge Paul DiMaggio, Sean Gerrish, Michael Herz, and Kim Lane Scheppele for their support and guidance as well as three anonymous reviewers for their helpful comments. Andrew Pendleton at the Sunlight Foundation was instrumental in helping us to acquire our data.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Financial support for the research was provided by the National Science Foundation (SES #1228436), Intel Labs, and the Horowitz Foundation for Social Policy.

Notes

1. Daily and weekly limitations apply; the specific limits, which have changed a few times over the years, are not necessary to rehash for our purposes. The current rules are found at 49 C.F.R. § 395.3.

2. This technique almost certainly resulted in some categorization errors, though it struck us as infeasible to read the text of thousands of comments in order to categorize commenter identity with perfect accuracy. However, our review of several hundred comments suggested that the categorizations generated by this technique were at least 95% accurate.

References

Belzer, M. H. (2000). *Sweatshops on wheels: Winners and losers in trucking deregulation*. New York, NY: Oxford University Press.

Blei, D. M. (2011). Introduction to probabilistic topic models. *Communications of the ACM*. Retrieved from http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. Srivastava & M. Sahami, (Eds.), *Text Mining: Classification, Clustering, and Applications* (pp. 71–93). Boca Raton, FL: Taylor & Francis Group.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boltanski, L., & Thévenot, L. (1999). The sociology of critical capacity. *European Journal of Social Theory*, *2*, 359–377.

Cardie, C., Farina, C., Aijaz, A., Rawding, M., & Purpura, S. (2008). A study in rule-specific issue categorization for e-rulemaking. In *Proceedings of the 2008 International Conference on Digital Government Research* (pp. 244–253). Digital Government Society of North America, Montreal, Canada.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta, (Eds.), *Proceedings of Advances in Neural Information Processing Systems 2009* (pp. 288–296).

Coglianese, C. (2004a). E-Rulemaking: Information technology and the regulatory process. *Administrative Law Review*, *56*, 353–396.

Coglianese, C. (2004b). Information technology and regulatory policy: New directions for digital government research. *Social Science Computer Review*, *22*, 85–91.

Coglianese, C. (2006). Citizen participation in rulemaking: Past, present, and future. *Duke Law Journal*, *55*, 943–968.

Cuéllar, M. (2005). Rethinking regulatory democracy. *Administrative Law Review*, *57*, 411–500.

Golden, M. M. (1998). Interest groups in the rule-making process: Who participates? Whose voices get heard? *Journal of Public Administration Research and Theory*, *8*, 245–270.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267–297.

Krawiec, K. D. (2013). Don't "screw Joe the plummer": The sausage-making of financial reform. *Arizona Law Review*, *55*, 53–103.

Kwon, N., Shulman, S. W., & Hovy, E. (2006). Multidimensional text analysis for eRulemaking. In *Proceedings of the 2006 International Conference on Digital Government Research*. Digital Government Society of North America, Montreal, Canada.

Lau, G. X., Law, K. H., & Wiederhold, G. (2005). Analyzing government regulations using structural and domain information. *Computer*, *38*, 70–76.

Mendelson, N. A. (2011). Rulemaking, democracy, and torrents of e-mail. *George Washington Law Review*, *79*, 1343–1380.

Shulman, S. W., Schlosberg, D., Zavestoski, S., & Courard-Hauri, D. (2003). Electronic rulemaking: A public participation research agenda for the social sciences. *Social Science Computer Review*, *21*, 162–178.

Steelman, T. A. (1999). The public comment process: What do citizens contribute to national forest management? *Journal of Forestry*, *97*, 22–26.

Wagner, W. E. (2010). Administrative law, filter failure, and information capture. *Duke Law Journal*, *59*, 1321–1432.

Wagner, W., Barnes, K., & Peters, L. (2011). Rulemaking in the shade: An empirical study of EPA's air toxic emission standards. *Administrative Law Review*, *63*, 99–158.

Yackee, J. W., & Yackee, S. W. (2006). A bias towards business? Assessing interest group influence on the U.S. bureaucracy. *The Journal of Politics*, *68*, 128–139.

Author Biographies

Karen E.C. Levy is a PhD candidate in the Department of Sociology at Princeton University. Her research focuses on intersections among technological systems, legal rules, and social control, with emphasis on surveillance and monitoring. She holds a JD from Indiana University. She may be contacted at kelevy@princeton.edu.

Michael Franklin holds a bachelor's degree in computer science from Princeton University. His interests include information security and more broadly the influences of technology on society and vice versa. He may be contacted at m.franklin.128@gmail.com.