# The Market for Data Privacy

**Tarun Ramadorai, Antoine Uettwiller, and Ansgar Walther**[1]

This draft: June 2019

[1]Ramadorai: Imperial College London and CEPR. Email: t.ramadorai@imperial.ac.uk. Uettwiller: Imperial College London. Email: a.uettwiller17@imperial.ac.uk. Walther: Imperial College London. Email: a.walther@imperial.ac.uk. We are grateful to Michelle Lee, Alex Hum, and Rehan Zahid for research assistance. We thank seminar participants at Imperial College Business School, and Stephen Hansen, Roxana Mihet, Emiliano Pagnotta, David Thesmar, Tommasso Valletti, and Andre Veiga for comments.

**Abstract**

We scrape a comprehensive set of US firms' privacy policies to facilitate research on the supply of data privacy. We analyze these data with the help of expert legal evaluations, and complement this analysis with data on firms' web tracking activities. We find considerable and systematic variation in privacy policies along multiple dimensions including ease of access, length, readability, and quality, both within and between industries. We also find that large firms with intermediate knowledge capital intensity have longer, legally watertight policies, but are more likely to share user data with third parties. We set up a simple theory of big data acquisition and usage which predicts the observed relationships between firm size, knowledge capital intensity, and privacy supply.

# 1 Introduction

Recent events, such as the Cambridge Analytica leak in 2018, have generated enormous public interest in possible privacy violations. Large data brokers and platforms such as Google, Facebook, and Amazon have a staggering ability to track consumers' behavior and personal details across a wide range of their online and offline activities (Varian, 2010; Jolls, 2012).[2] Faced with this scenario, Congress is considering whether the US should adopt stricter, European-style regulation,[3] but the economic principles underlying these debates are subtle. The classical view (Stigler, 1980; Posner, 1981) suggests that data collection allows firms to allocate resources, such as online advertising space, more efficiently (e.g., Goldfarb and Tucker, 2011). However, these efficiency benefits must be traded off against negative externalities that data sharing imposes on consumers (Varian, 2009), which are especially worrisome if consumers are unaware of data collection practices (Taylor, 2004).[4]

Given the complexity of the issue, it is crucial to have a more detailed understanding of how the market for privacy operates. There is evidence that consumer behavior deviates from the classical benchmark—while consumers routinely express a preference for privacy (Westin and Ruebhausen, 1967; Goldfarb and Tucker, 2012), they tend to exhibit "consent fatigue," failing to read firms' privacy policies.[5] Moreover, consumers often seem to interpret the mere presence of a policy as a signal of protection (Acquisti et al., 2015). In the presence of such consumer inaction, it becomes especially important to document firms' behavior.

We provide new evidence on the supply of privacy, by acquiring, processing, and

[2] Federal Trade Commission (2014) reviews data brokers' activities, and a growing literature in computer science documents the prevalence of web tracking (e.g., Krishnamurthy and Wills, 2009).

[3] See "Ad world flocks to Congress urging federal data privacy legislation", *The Drum*, 26 February 2019, and "Should Congress override state privacy rules? Not so fast.", *The Washington Post*, February 26, 2019.

[4] A large literature considers additional "second-best" arguments which speak either for or against privacy (e.g., Hirshleifer, 1971; Daughety and Reinganum, 2010; Calzolari and Pavan, 2006). Acquisti et al. (2016) provide a comprehensive review.

[5] See, for example, "Why your inbox is crammed full of privacy policies," WIRED May 24, 2018 , and "Getting a Flood of G.D.P.R.-Related Privacy Policy Updates? Read Them," *The New York Times,* May 23, 2018.

analyzing the privacy policies of a comprehensive set of US firms. To acquire the data, we first search for the privacy policies of all 5377 US firms in Compustat, using a combination of automated Google searches, web crawling techniques, and manual searches using each firm's main web domain. In total, we are able to obtain policies for 4078 firms (75.4% of the entire Compustat sample). We also leverage recent web measurement tools (Englehardt and Narayanan, 2016) to analyze the code of each firm's website, and to detect the presence of third-party cookies and other tracking devices.[6] We relate the attributes of firms' privacy policies, as well as firms' actual web tracking behavior, to a number of firm characteristics.

Our analysis of these data leads to three main contributions. First, we document considerable and systematic variation in firms' stated privacy policies, both within and across industries. Second, we contrast the content and quality of firm's privacy policies with their actual privacy practices. We show that firms whose stated privacy policies clearly outline contingencies and remedies to consumers are *more* likely to share consumer data with third parties; moreover, these firms' policies are generally longer, and use legal language that is harder to read. These findings suggest that greater legal clarity in privacy policies, rather than protecting consumers, might well facilitate or legitimize data sharing, and we explore this possibility more fully in the paper. Third, we provide a simple profit-maximizing theory in which firms choose whether or not to share and process data, and also choose the legal soundness of the privacy policies that they write, with a view towards limiting liability arising from their sharing practices. We use the theory to interpret the new facts that we uncover, and to confirm additional predictions of this theory in our data.

A common prior is that firms utilize simple "boilerplate" privacy policies that potentially vary only across, but not within industries. We find evidence contrary to this prior. First, we find considerable variance across policies in their length, as well as their basic paragraph structure. Second, the text of the privacy policies varies considerably. The median cosine similarity between individual policies and the sample

[6]We have made our data and code publicly available for other researchers at: https://github.com/ansgarw/privacy.

centroid is 0.57, which translates to a 55-degree median angle between policy word-frequency vectors and that of the grand average policy. Third, we find that most of this variation occurs within rather than between industries.[7]

Policies also vary considerably in the quality of their text. We first evaluate this using a common linguistic index, the Gunning (1952) "Fog" index of "readability," which is based on the sentence-level frequency of complex and polysyllabic words in the privacy policies, and heuristically measures the number of years of formal education required to understand a document at first reading. By this metric, one needs at least a college degree to follow the median privacy policy in our sample, highlighting the (lack of) readability in most privacy contracts.

We then move on to a more detailed analysis of the text in these documents, employing a human legal expert to read through a 10% sample of the privacy policies in the dataset and score them on their ostensible ability to protect consumer privacy along a set of dimensions. These are: Data Collection, User Consent, Responsible Use, Third-Party Sharing, and User Rights. The policies are also given an "Overall" score that seeks to amalgamate the scores on the individual categories,[8] but also takes a judgment call about whether the policy clearly spells out possible contingencies and remedies to an end-user who is able and willing to exercise rights over their personal data. We use the human scores in the 10% training sample to build a simple natural language processing (NLP) classifier which we apply to all sample firms' policies, which gives us a policy-specific measure of "Legal Clarity." We find substantial variation in policies across firms along this dimension, and find that Legal Clarity is also associated with several policy attributes that are easier to construct, and less reliant on a specific NLP model.[9] In particular, longer policies, as well as those that are more difficult-to-read (using the Gunning (1952) "Fog" index) also

[7]The median cosine similarity with the 3-digit SIC-level centroid is around 0.62, corresponding to a 51-degree median angle between firms' policy word frequency vectors and that of the industry-average vector.

[8]Note that we find evidence that the different attributes of policies are correlated with one another.

[9]We check how the specific NLP classifier used to construct the Legal Clarity index affects our work; we find that our results are qualitatively robust to the use of different classifiers.

exhibit higher Legal Clarity. In an environment strongly characterized by consent fatigue, these are interesting new facts about the text of privacy policies, as they suggest that policies are more than just a device to protect consumers.

We find that these linguistic differences between policies are not simply attributable to idiosyncratic noise in the quality and quantity of verbal expression across firms. We first find systematic variation in privacy policies with firm size. The largest firms more often *have* privacy policies, the word "Privacy" is more likely visible on their homepages, and their policies have higher Legal Clarity scores. However, large firms' policies are significantly lengthier and more complex. Moreover, large firms have a significantly higher incidence of third-party tracking cookies on their websites.[10]

We also find an interesting non-monotonic relationship between privacy policy attributes and firms' technical sophistication (which we measure using the Peters and Taylor (2017) measure of knowledge capital, expressed as a share of firms' total capital). Firms with intermediate technical sophistication have longer, more legally watertight policies, but are more likely to share data on their users' browsing history with third parties. However, firms with the very highest knowledge capital intensity have shorter, less complex, and less legally watertight policies, and simultaneously engage in less third-party sharing of user data from their websites.

To interpret these patterns in the data and to further discipline our empirical work, we build a model of firms' use of data, the clarity of their privacy policies, and their interactions with data intermediaries. We model a firm which can monetize consumer data by turning it into prediction-based products.[11] The firm can either process its data in-house, or share it with data intermediaries, who can monetize these data more

[10] These data on web-tracking are derived purely from analyzing the firms' websites, i.e., independently from the privacy policies. This helps to provide external validation for any inferences arising solely from the text of firms' privacy policies.

[11] A growing literature, complementary to our analysis, studies in detail how information should be monetized, e.g. by running auctions for goods and services in conjunction with information release (Eső and Szentes, 2007), garbling their information to elicit buyers' preferences (Bergemann et al., 2018), or selling information gradually over time (Hörner and Skrzypacz, 2016). A related theme is the analysis and sales of financial data (Admati and Pfleiderer, 1986), which has inspired recent research on big data and trading (e.g., Farboodi and Veldkamp, 2017; Begenau et al., 2018). See (Agarwal et al., 2018) for a book-length treatment of these themes.

efficiently. The firm also chooses the soundness of its privacy policy, balancing the legal costs associated with drafting a watertight policy against the benefits it obtains from mitigating any legal risks arising from its actions to share data.

The model shows that a firm's propensity to share data depends only on three sufficient statistics. These are: (i) the value of the firm's data when it is processed most efficiently, (ii) the total opportunity cost of the firm processing its data in-house rather than selling the data to the intermediary, and (iii) the cost-to-benefit ratio associated with having a high-quality privacy policy. This cost-to-benefit ratio in turn depends on the firm's bargaining power vis-à-vis the data intermediary, the risk mitigation arising from having a higher-quality policy, and the legal cost to the firm of putting the privacy policy in place.

To take the model to the data, we map the sufficient statistics in the model to the two variables mentioned earlier, namely, firms' size and technical sophistication. The resulting estimation reveals, consistent with our theoretical predictions, that large firms with intermediate knowledge capital intensity have longer, more legally watertight policies, but are more likely to share data on their users' browsing history with third parties. However, firms with the very highest knowledge capital intensity have shorter, less complex, and less legally watertight policies, and simultaneously engage in less third-party sharing of user data from their websites. This is consistent with firms deciding to process data "in-house" rather than share it when they achieve a sufficient degree of technical proficiency, and sharing data with third-parties otherwise.

The remainder of the paper is organized as follows. Section 2 describes our data on firm characteristics, privacy policies, and web tracking behavior, and provides basic descriptive statistics. In Section 3, we explore the variation in firms' privacy policies and behavior, and show how it systematically relates to firms' economic characteristics. We set up our theoretical model in Section 4, and test its additional predictions in Section 5.

# 2 Data

## 2.1 Firm Data

For all firms in the US Compustat database, we obtain data on market capitalization, book values of assets and equity, sales, intangible assets, R&D, SG&A, and marketing expenditures. For a more precise measure of intangibles, we obtain intangible capital as used by Peters and Taylor (2017). This is the sum of Compustat-recorded on-balance-sheet intangible capital, and the replacement values of knowledge and organizational capital. Peters and Taylor (2017) estimate the replacement value of knowledge capital by accumulating past R&D expenditures for firms assuming an industry-specific depreciation rate. They also estimate the replacement value of organizational capital almost identically to Eisfeldt and Papanikolau (2014), by accumulating a fraction of past SG&A expenditures.

For stock variables (market values, book values, assets, and intangibles) we take the latest available quarterly observation (2018Q1). For flow variables (sales, R&D, and marketing expenditures) we take the average over the last three years. As in Crouzet and Eberly (2018), we drop firms with missing sales data, or values of sales that are less than or equal to zero, firms with missing data on asset book-value, or those with asset book-value less than or equal to US$ 1MM, firms with missing market value,[12] and firms that do not list their web domain on Compustat. Following these filters, the Compustat sample comprises 5377 firms.

For each firm, we calculate the market-to-book ratio of assets, the firm's market share of sales in its (2-digit SIC code) industry, the share of its capital accounted for by intangible and knowledge capital (i.e., the fraction of knowledge capital and intangible capital in the firm to total capital, where total capital is the sum of (the replacement values of) knowledge capital and organizational capital, and total assets[13]), and the

[12] Where possible, we replace missing market values with the product of number of shares and price per share.

[13] Total assets in Compustat are the sum of Current Assets - Total (ACT), Property, Plant and

Table 1: **Descriptive Statistics of Firm Characteristics**

| | count | mean | median | std |
|---|---|---|---|---|
| Market Value | 5377.0 | 5693.413 | 664.697 | 15847.717 |
| Market to Book | 5345.0 | 2.207 | 1.408 | 2.216 |
| Market Share | 5377.0 | 0.011 | 0.001 | 0.032 |
| Intangible Share | 5306.0 | 0.159 | 0.043 | 0.216 |
| Knowledge Share | 5140.0 | 0.079 | 0.000 | 0.160 |
| R&D to Assets | 2621.0 | 0.029 | 0.010 | 0.051 |
| Marketing to Assets | 2029.0 | 0.022 | 0.005 | 0.043 |

Note: Market value is measured in millions of USD. Market Share is the firm's sales divided by industry sales at the 2-digit SIC code level. Intangible Share is intangible assets divided by total assets, and Knowledge Share is the replacement value of knowledge capital divided by the sum of intangible assets and the replacement values of knowledge and organizational capital (see Peters and Taylor (2017)). We winsorize all variables at their first and 99th percentile.

ratio of marketing and R&D expenditures to assets.

Table 1 shows descriptive statistics of these firm characteristics, following winsorization at the 1 and 99 percentile points. On average, knowledge capital accounts for roughly 8% of total capital, with total intangible capital roughly double this amount. The median knowledge share is zero, meaning that this is a skewed distribution, and some firms in the data exhibit very high fractions of knowledge capital. There is also a fairly high standard deviation across firms in knowledge share.

## 2.2 Privacy Policies

We search for firms' privacy policies using automated Google searches and web crawling techniques. We restrict this search to each firm's main web domain, as listed in Compustat. We supplement this method by a web crawl of the firm's domain, and finally, by manual checking, in cases where a policy was not found automatically. We scrape the text of each policy, and discard it if it does not contain the word "privacy". We also discard all paragraphs that have fewer than 100 characters (usually consisting

Equipment (Net) - Total (PPENT), Investment & Advances - Equity (IVAEQ), Investment & Advances - Other (IVAO), Intangible Assets - Total (INTAN), and Assets - Other - Total (AO).

Table 2: **Descriptive Statistics of Privacy Policy Attributes**

| | count | mean | median | std |
|---|---|---|---|---|
| Policy Found | 5377.0 | 0.758 | 1.000 | 0.428 |
| Policy Found on Google | 4078.0 | 0.918 | 1.000 | 0.274 |
| "Privacy" Visible on Homepage | 5377.0 | 0.647 | 1.000 | 0.478 |
| Number of Paragraphs | 4078.0 | 30.972 | 23.000 | 27.134 |
| Number of Words | 4078.0 | 1858.741 | 1433.000 | 1645.912 |
| Gunning Fog Index | 4078.0 | 17.792 | 17.695 | 2.579 |
| SMOG Index | 4078.0 | 15.616 | 15.569 | 1.770 |

Note: Policy Found is equal to one if we found a privacy policy by automated scraping or manual collection. Policy Found on Google is equal to one if, conditional on policy found, a link to the policy appeared during an (automated or manual) Google search. "Privacy" Visible on Homepage is equal to one if the firm's root web domain, as recorded in Compustat, contains a link with the word "privacy". We winsorize the length variables at their first and 99th percentile. For definitions of the Fog and SMOG indices, see Gunning (1952) and McLaughlin (1969).

of headings or snippets of HTML code). In total, we are able to obtain policies for 4078 firms (75.4% of the sample).[14]

Table 2 shows descriptive statistics for how easy it is to find privacy policies, and the length of policies in terms of paragraphs and words. The word "privacy" is visible on the homepages of only 65% of the sample of 5377 firms in Compustat. Roughly 92% of the 4078 policies acquired are found using google searches, while the remaining 8% were recovered manually. When found, the average privacy policy contains around 31 paragraphs, comprising roughly 1900 words, with considerable variance across policies in both length and paragraph structure. At first glance, the considerable variation in length and structure suggests that we should question the simple prior that firms all use a common boilerplate contract.

To prepare the policies for textual analysis, we remove all non-English words, and words associated with named entities such as organizations, persons, and locations. We also remove very common English words from a standard list of "stop words" that convey little semantic meaning (e.g., "is","in", "and", "or").[15]

[14]These policies are scraped from 4062 unique web domains. One web domain is shared by three firms in our sample, and 14 domains are shared by two firms each.

[15]We detect non-English with the pyenchant spellchecker (see https://github.com/rfk/pyenchant),

Figure 1 visualizes the textual content of our sample in terms of the most important bigrams (pairs of consecutive words). We measure the importance of bigrams in each policy with a standard TF.IDF (term frequency-inverse document frequency) metric, which attaches high importance to bigrams that are frequent within a policy relative to its overall length, and penalizes generic bigrams that occur in a large fraction of documents.[16] As might be expected, the policies prominently feature the word "privacy." They also prominently feature the terms "personal information," "personal data," and "personally identifiable information." Finally, an important and frequently used term that is clearly evident is "third party," which we will return to discussing later in the paper.

Table 2 also shows the descriptive statistics of two common linguistic indices of "readability", which are based on the sentence-level frequency of complex and polysyllabic words. For instance, the Gunning (1952) "Fog" index heuristically measures the number of years of formal education required to understand a document at first reading. By this metric, one needs at least a college degree to follow the median privacy policy in our sample, highlighting the (lack of) readability in most privacy contracts.[17]

and named entities with the Stanford NER tool (see Finkel et al. (2005) and https://nlp.stanford.edu/software/CRF-NER.shtml). We use the NLTK list of English stopwords (see https://www.nltk.org/nltk_data/).

[16] We divide each bigram's number of occurrences in each policy by the total number of bigrams in the policy, for the bigram's "term frequency" (TF). We then multiply the TF by the log of the inverse fraction of documents containing the bigram, known as the "inverse document frequency" (IDF). Let $P_{ij}$ be the number of times that bigram $j$ appears in document (in our case, policy) $i$. The TF.IDF metric is:

$$\hat{P}_{ij} = \underbrace{\left( P_{ij} / \sum_k P_{ik} \right)}_{TF} \cdot \underbrace{\log\left( \frac{N}{\sum_i 1\{P_{ij} > 0\}} \right)}_{IDF}$$

where $N$ is the total number of documents. See Rajaraman and Ullman (2011), and Gentzkow et al. (2018) for more detailed treatments of TF.IDF.

[17] These figures are similar to recent findings in the Computer Science literature (e.g., Fabian et al., 2017). The McLaughlin (1969) "SMOG" index is measured on the same scale, and since its correlation with Fog is 0.99, we choose to use the Gunning Fog index as our measure of readability in the remainder of the paper. The Gunning algorithm determines the average sentence length in the policies, and counts all words comprising three or more syllables. To get the final Gunning measure, the algorithm just adds the average sentence length and percentage of three-or-greater syllable words, and scales the result.

Figure 1: **Word Cloud of the Privacy Policy Sample**


Note: Bigrams are scaled by their average TF.IDF score across all 4078 privacy policies in our sample.

## 2.3 Web Tracking Data

We obtain tracking data for each firm's web domain using the methodology developed in Englehardt and Narayanan (2016).[18] This is to provide us with an independent measure of each firm's approach to data privacy, based on the detail with which they track behavior of individuals browsing their websites. This measure has no direct relationship with the privacy policy data, and helps to provide external validation for findings using the text of the privacy policies.

Descriptive statistics of these tracking data are in Table 3. The table shows two ways of measuring third-party tracking: The number of unique third parties who place cookies that are classified as "tracking cookies," and the *total* number of third-party requests, including, but not limited to cookies. The former measure is more conservative because tracking is not always done via cookies, but also quite

[18]Their open-source privacy measurement software is available at https://github.com/mozilla/OpenWPM. We obtain our data by scraping the results for each firm on https://privacyscore.org, which uses OpenWPM. There are null returns for which the crawler fails, and these reduce the sample size to 5184 of 5377 total firms in the sample.

Table 3: **Descriptive Statistics of Web Tracking Data**

| | count | mean | median | std |
|---|---|---|---|---|
| Third Party Tracking Cookies | 5184.0 | 2.449 | 1.0 | 4.805 |
| Third Party Requests | 5184.0 | 33.828 | 19.0 | 40.442 |

Note: Third Party Trackers is the number of unique third-party domains that place tracking cookies on each firm's homepage. Third Party Requests is the total number of HTTP requests from third-party websites on each firm's homepage.

accurate, because not all cookies are tracking cookies. The correlation between the two measures is 0.75. We use the (conservative ) "tracking cookies" measure in what follows.

# 3 Understanding Variation in Privacy Policies

To understand the information contained in the privacy policy text, we proceed in a series of steps. We first describe our use of a set of simple natural language processing (NLP) techniques to convert the text in the privacy contracts into quantitative information that is susceptible to empirical analysis. Our approach includes a legal assessment of the privacy protection afforded to consumers, for which we employ human input in addition to simple machine learning approaches.

Our next step is to assess a simple prior, namely, that all firms operate using a standard/industry-specific "boilerplate" privacy contract, and do not vary in any materially important manner. The descriptive statistics already reveal that there is cross-firm variation in how easy policies are to access, and how long they are. We go a bit further in this section, evaluating the content of the documents. We do so by first evaluating the cosine similarity between different policies to provide a sense of how varied the text in the policies is. We then take a first look at the relationship between privacy policy content and firm characteristics.

## 3.1 Evaluating Privacy Policy Content

There are several possible "unsupervised" approaches, such as topic models, that are commonly used to evaluate the content of text documents (see, for e.g., Gentzkow et al. (2018)). Given the documented complexity (the high Fog index seen above), and specialized nature of the language in privacy policies, we choose instead to adopt a simple supervised learning approach to interpret the content of these policies. To create an initial labelled "training" sample, we therefore use human expert input to carefully read and classify a subset of the policies.

We therefore sent 407 policies (10% of unique scraped policies) to a legal expert for evaluation. He determined that 54 of these policies did not contain meaningful legal text related to privacy. For the remaining policies, he assigned high (1), neutral (0) or low (-1) overall scores to each policy's protection of consumer privacy. He further assigned category-specific scores regarding the strength of the policy, taking the consumer's perspective, on six dimensions. These are:

1. Data collection: High scores given for policies that make data collection needs clear. If comprehensive data is collected, policies are scored highly if the types of data are in-line with industry standards. Low scores given for policies which collect data so comprehensively as to appear excessive, or vague in the sense that users would not understand the data they are providing.

2. User consent: High scores indicate the seeking of specific consent for different processes, and proactive notification of the user of changes to the policy. Low scores for consent clauses that presume the user's consent from their continued use (sometimes aggressively disclaiming their liability), and/or requiring the user to frequently check and review the policy with each use.

3. Responsible use: High scores mean proactive offers to the user of clear benefits and robust assurances. Low scores for policies specifying extensive use of user data, subjection to heavy advertising and additional services, and/or extensive

additional tracking and monitoring tools.

4. Third-party sharing: High scores mean that third-party sharing is clearly explained, appears legitimate, and the organization retains some liability and responsibility over the shared data. Low scores if the approach to sharing personal data with third parties is unclear or poorly explained, extensive, and/or not obviously justified by the potential interests of, or benefits to, the user.

5. User rights: High scores if significant and comprehensive rights are granted to the users over their data, and points given for the simplicity of the exercise of these rights by the user. Low scores if no rights over personal data are conferred at all onto the user, or if they are, they are minimal, poorly explained, difficult, and inaccessible for users to actually put into effect.

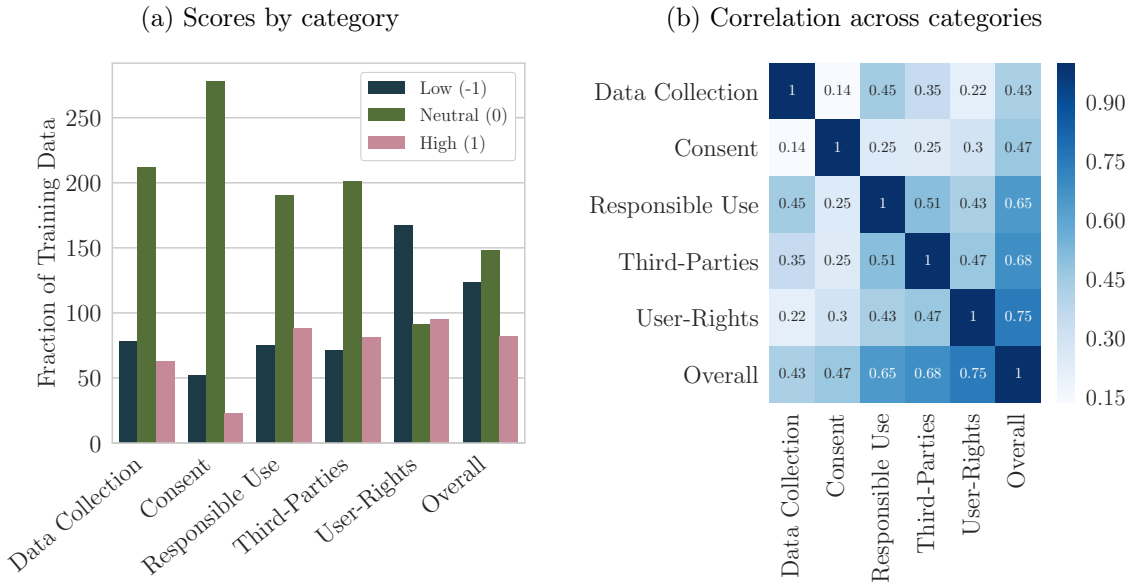
6. Overall: An amalgamation of the scores on the individual categories, alongside a judgment call about whether the policy clearly spells out possible contingencies and remedies to an end-user who is able and willing to exercise rights over their personal data.

In the appendix, we provide more detailed descriptions of the expert's definitions, as well as specific examples of firms whose policies satisfied the key evaluation criteria that were used to determine scores on each category.

Figure 2a shows the distribution of evaluations for each dimension into which policies are categorized, as well as the distribution of the overall score. A regression of the overall score on the individual category scores gives an $R^2$ of 81%, i.e., they are closely related, consistent with the overall category being close to an aggregate of individual category scores. 42% of policies are classified as "Neutral", with a larger fraction (35%) classified as "Low," than "High" (23%) as regards their overall protection of consumer privacy.

Figure 2b shows the correlation matrix of the numerical scores across categories. They seem to capture distinct concepts, but it appears that a high overall score has

Figure 2: **Distribution of policy-level expert evaluations**




a strong association with high scores on both Third-party sharing and User rights.

Figure 3 visualizes important bigrams associated with policies that receive a high or low overall ranking. To highlight bigrams that are specific to high and low rated policies, the importance of bigrams in this figure is normalized by the average importance in the population (i.e., those depicted in Figure 1).

We next construct a simple measure of overall policy soundness, based on the expert review, that can be extrapolated to our entire sample. For each privacy policy, we compute the total frequency of the most important 100 bigrams in policies rated "High" by our expert (i.e., the unweighted sum across all the bigrams in the word cloud in Figure 2a) and subtract the equivalent metric for "Low" policies (i.e., the sum across Figure 2b). We term the result the "Legal Clarity index," with different scores assigned to each of the 4078 policies in our dataset using this simple scoring technique.[19],[20]

[19]Let $G$ and $B$ be the set of bigrams appearing in Figure 2a and 2b respectively, and let $\hat{P}_{ij}$ be the TF.IDF frequency of bigram $j$ in policy $i$. Then, our measure of quality is $Q_i = \sum_{j \in G} \hat{P}_{ij} - \sum_{j \in B} \hat{P}_{ij}$. We normalize this measure to have mean zero and variance one.

[20]This can be thought of as a very simple supervised learning model using the classifications in the training data. Below, we test how different the results are when we use more sophisticated

Figure 3: **Expert Evaluation**

(a) High overall score


(b) Low overall score


Note: Bigrams are scaled by the difference between the average TF.IDF score among policies evaluated as high (low) by a legal expert, and the grand average TF.IDF score in our sample.

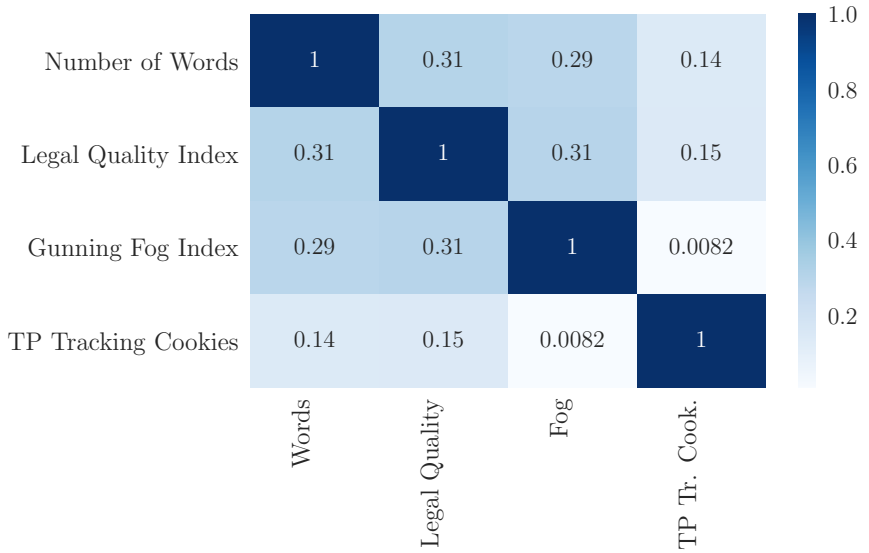
Figure 4 shows the correlations among the set of privacy policy attributes for all policies, using the simple length of each policy, the Legal Clarity index, the Fog index, and the incidence of tracking cookies placed by unique third parties. Most of these attributes are positively correlated with one another.

It is particularly interesting that the incidence of tracking cookies has a positive (15%) correlation with the Legal Clarity index. High Legal Clarity, at first glance, appears reassuring from the perspective of the user. However such clarity in the presence of consent fatigue may not necessarily be associated with firms refraining from sharing data with third-parties. Indeed, Legal Clarity might even facilitate such sharing if it empowers and legitimizes firms to take actions that they can claim have been clearly spelled out to customers, regardless of whether the customer has bothered to read the text of the privacy policies.

The Legal Clarity index is also highly correlated with the Fog index (31%), as well as the length of the policy (also 31%) which supports this interpretation—high Legal

approaches common in the NLP/machine learning literature.

Figure 4: **Correlations: Policy Attributes and Website Tracking**




Clarity policies are also longer, and more confusing to read for the average user. In an environment strongly characterized by consent fatigue, these are interesting new facts about the text of privacy policies.[21]

As a robustness check on these results, we also consider an alternative measure of legal clarity. We obtain this measure by fitting a supervised machine learning model (a penalized logistic regression) on the subsample of policies that we sent for expert review. We then obtain the predicted scores from this model for the remaining policies in our sample. Another variant of this model is that we augment it with a step that predicts and filters out policies that our expert classified in the training dataset as not containing any privacy-related information. In the appendix, we describe this alternative approach in further detail, but note here that the results from this robustness exercise are broadly similar to the results described here.

Our next step is to assess whether the policies contain similar text despite the differences in length and interpretability that we outline above. To dig deeper on this issue, we move to estimating cosine similarities between the privacy policies in the

[21]See, for example, "Why your inbox is crammed full of privacy policies," WIRED May 24, 2018 , and "Getting a Flood of G.D.P.R.-Related Privacy Policy Updates? Read Them," *The New York Times,* May 23, 2018.

data.

## 3.2 Are Firm Policies Standard Boilerplate?

Each policy can be described as a vector $P_i = (P_{i1}, ..., P_{iM})$ of term frequencies, where $P_{ik}$ is the frequency of term $k$ in policy $i$. The cosine similarity $C_{ij}$ between two policies is the cosine of the angle between their vector representations $P_i$ and $P_j$:

$$C(P_i, P_j) = \frac{P_i \cdot P_j}{\|P_i\| \, \|P_j\|}$$

For an intuitive interpretation, suppose the only two possible terms are "apple" and "orange". If policy $i$ only mentions apples, and policy $j$ mentions only oranges, then the angle between the two vectors is 90 degrees (i.e., they are orthogonal), and $C = 0$. If both policies mention only apples, then the angle is zero, and $C = 1$.[22]

To measure aggregate variation, we compute the cosine similarity between each policy vector $P_i$ and the centroid vector of all privacy policies in the sample, i.e., the "average" policy $\bar{P} = \left(\sum_j P_j\right)/N$. To isolate variation within industries, we compute the similarity between each policy and the associated industry-level centroid $\bar{P}_I = \left(\sum_{j \in I} P_j\right)/(\#I)$, where $I$ is the set of firms in an industry.

A situation in which all firms (or all firms within an industry) adopt roughly the same boilerplate policy would mean that these cosine similarities are close to one. As in the visualizations above, we use TF.IDF frequencies throughout, but as is customary in the literature, we focus on the frequencies of words (rather than bigrams) when computing cosine similarities.

Figure 5a shows the cumulative distributions of cosine similarities with the sample centroid (the centroid associated with each firm's SIC sector), as well as with the

[22] Note, however, that this measure is nonlinear due to the cosine transformation: If a third policy $k$ mentions apples and pears with equal frequency, then the angle is 45 degrees and $C_{ik} = \cos(45\text{deg}) = 0.71$. Since the cosine wave becomes steeper between zero and 90 degrees, the similarity measure is therefore more forgiving of small discrepancies between policies.

centroid associated with each firm's 2- and 3-digit SIC code bucket. The median cosine similarity between individual policies and the sample centroid is 0.57, which translates to a 55-degree median angle between policy vectors and the grand average policy.[23]

The figure also reveals that *within-industry* variation is marginally smaller than *total* variation: As we move to finer industry classifications, the distribution of similarities shifts in a first-order sense. For example, the distribution of similarities with SIC2 centroids lies strictly below similarities with the sample centroid.

However, there is still substantial variation within industries. For instance, the median cosine similarity with the 3-digit SIC-level centroid is about 0.62, corresponding to a 51-degree median angle between firms' policies and the industry-average vector. Figure 5b shows the associated mean cosine similarities, with 95-percent (bootstrapped) confidence intervals.

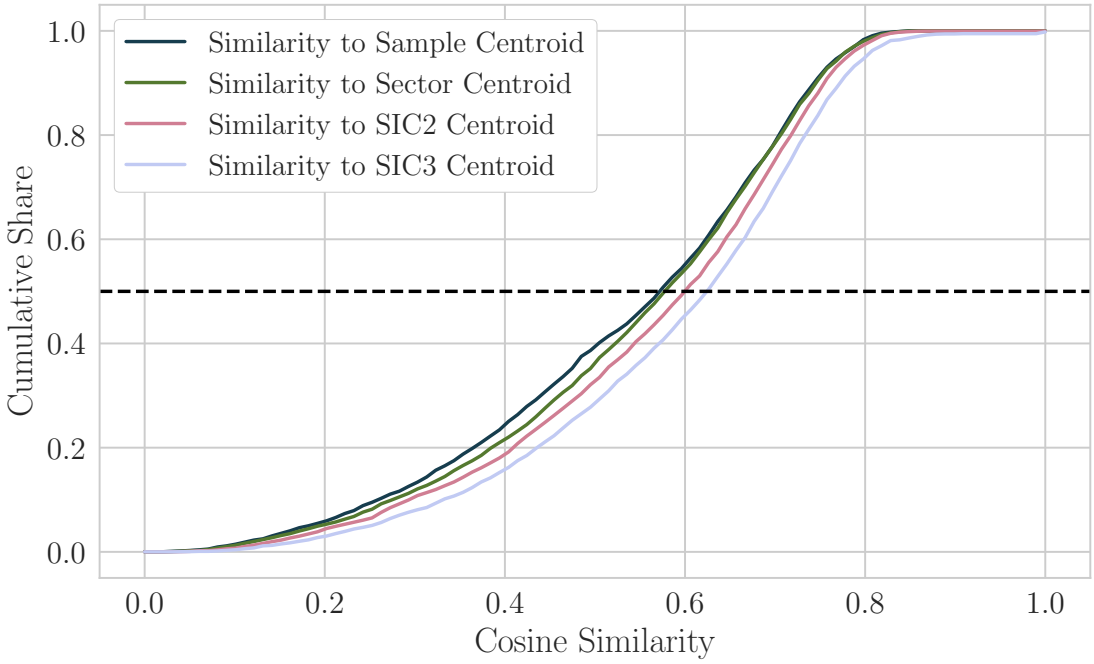
To verify the robustness of these inferences, in the online appendix, we repeat this exercise with a more specific measure of the semantic content of each policy. We use Latent Semantic Analysis (LSA), which amounts to a singular value decomposition on the term-document matrix, to reduce the dimension of the textual data from about 18000 words to 250 latent "topics". We expect this transformation to eliminate differences between policies that arise purely from idiosyncratic difference in wording, and to focus on semantic content. After the transformation, the cosine similarity with the 3-digit SIC-level centroid rises to about 0.73, which suggests that some of the differences are attributable to differences in vocabulary employed. However, this value is still far from the boilerplate prior of $C = 1$.

We also compute the cosine similarities *between* sector-level average policies in the appendix. With the exception of the agricultural and mining industries, we find

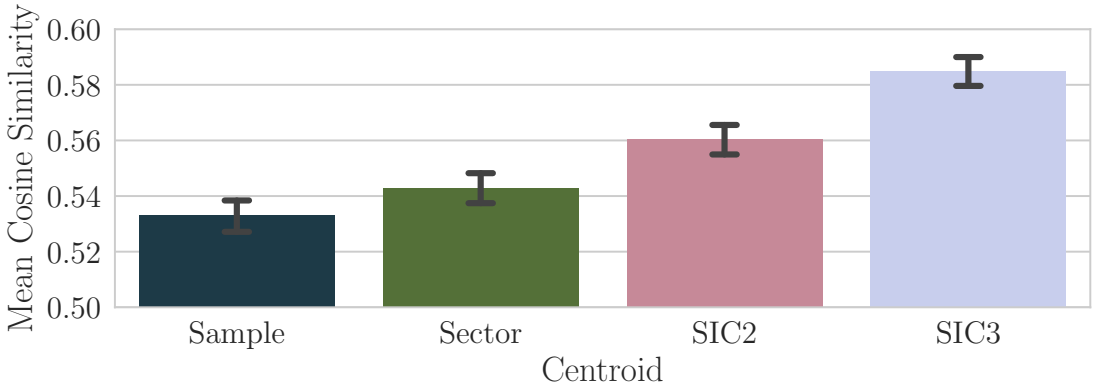
[23] An alternative measure of aggregate variation would be to compute the cosine similarity between all $N(N-1)/2$ pairs of policies in our sample. These similarities would also be close to one in a "boilerplate" world. The average similarity in this sense is 0.28, with a median of 0.27. Notice that it is natural for this measure to be quantitatively smaller than the distance from the centroid: In the example above, if policy $i$ only mentions apples, and policy $j$ mentions only oranges, then the pairwise-average similarity is zero, while the similarity to the centroid is cos(45deg).

Figure 5: **Variation in Privacy Policy Text**

(a) Cumulative Distributions of Cosine Similarities




(b) Mean Cosine Similarities




Note: The Sample centroid is the mean TF.IDF frequency vector across all 4078 privacy policies. Sector centroids are the mean TF.IDF frequency vectors in the 12 SIC divisions, which are Agriculture, Forestry and Fishing (SIC 0100-0999), Mining (SIC 1000-1499), Construction (SIC 1500-1799), Manufacturing (SIC 2000-3999), Transport, Communications, and Utilities (SIC 4000-4999), Wholesale Trade (SIC 5000-5199), Retail Trade (SIC 5200-5999), Finance, Insurance, and Real Estate (SIC 6000-6799), Services (7000-8999), Public Admin (SIC 9100-9729), and Nonclassificable (9900,9999). SIC2 and SIC3 centroids are mean frequencies at the 2-digit and 3-digit SIC code level, respectively.

that the centroids are very similar across industries, with cosine similarities in excess of 0.9 (i.e., an angle $> 26$ degrees). This is interesting, as it suggests that there are few industry-specific differences in privacy policies. Such differences might arise, for example, if the types of data collected, and the inferences derived from data about consumers, differ considerably across industries. The fact that there aren't such visible differences is consistent with similar uses of consumer data across industries, though not across firms within industries.

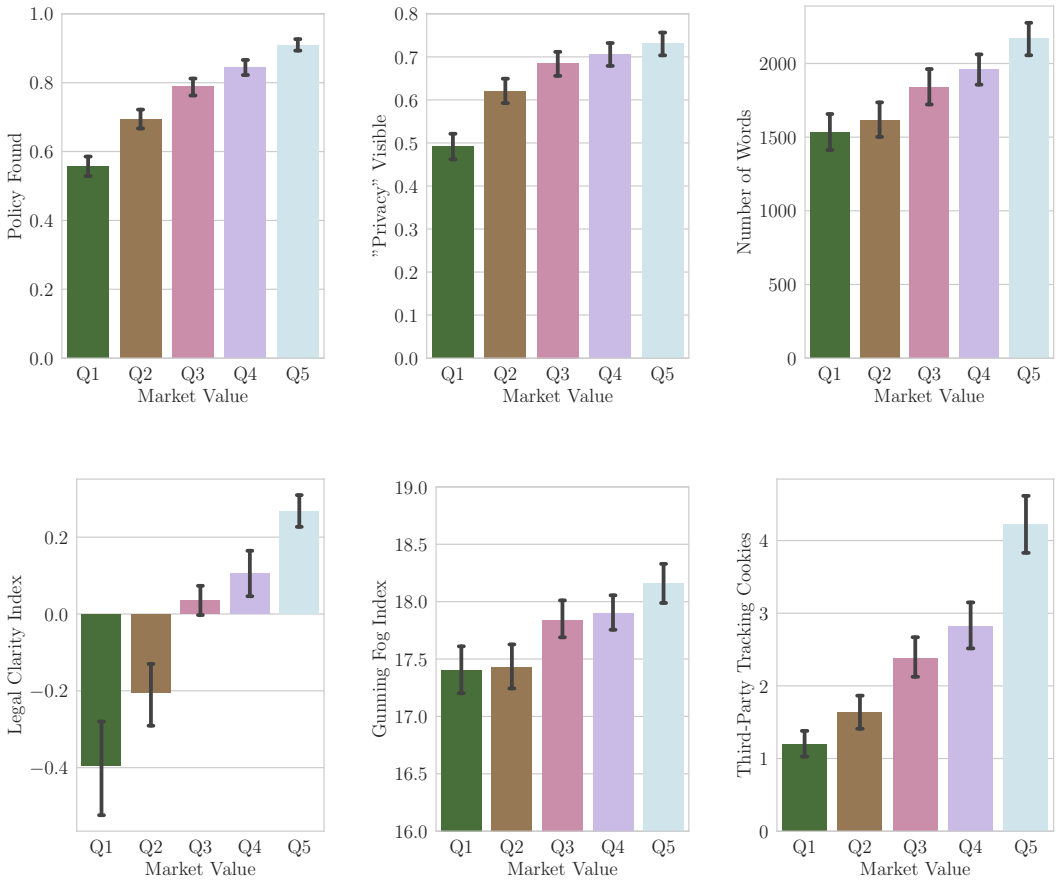
The upshot of this analysis is that, in contrast to the hypothesis that privacy policies are "boilerplates" that respond to industry-level regulation, we find that most of the variation in firms' policies occurs within industries. Moreover, this cross-firm variation is substantial. We next move to evaluating whether the content of the text contained in the privacy policies varies systematically with firm characteristics.

## 3.3 Firm Characteristics and Privacy Policy Text: A First Look

While we have established that firms' privacy policies differ considerably, and that there is seemingly no convergence to a "boilerplate" policy, it is entirely possible that the documented variation across policies is simply idiosyncratic noise in the quality and quantity of verbal expression across firms. This serves as a convenient null hypothesis to guide empirical investigation—that there is no systematic variation across policies that correlates with other characteristics of firms.

Keeping this null in mind, we start by investigating the relationship between the characteristics of privacy policies and two firm characteristics that seem intuitively important as a simple first step. We first evaluate the alternative hypothesis that privacy policy content and attributes are systematically related to firm size. We then evaluate the relationship between privacy policy content and attributes, and a measure of firms' technical sophistication, i.e., Peters and Taylor (2017)'s measure of "knowledge capital," which is essentially past accumulated R&D expenditures for firms

Figure 6: Privacy Policies and Firm Size




assuming an industry-specific depreciation rate. To make this measure comparable across firms, we simply scale it by firms' total (i.e., tangible plus intangible) capital as in Table 1, and term the result the "knowledge share."

We first plot the relationship between the important characteristics of privacy policies and firm size in Figure 6.

Rejecting the null hypothesis, Figure 6 clearly shows that numerous characteristics of privacy policies vary systematically with firm size. As we move from small firms to the largest firms, the likelihood of finding a privacy policy increases dramatically, from below 60% to above 90%. The likelihood that the word "Privacy" is visible on the firm's homepage also rises by a statistically significant 20%. Policies for larger firms are also significantly more verbose, with an average of above 2000 words in the

largest size quintile, compared to an average slightly above 1500 words in the bottom size quintile.

These differences are not merely cosmetic. The policies are also significantly different in their Fog scores, with policies from larger firms requiring an estimated additional year of education to interpret than those from the smallest two quintiles of firms, meaning that they are more complex. The Legal Clarity of the policies is also substantially different, with larger firms significantly more likely to have policies that clearly outline expected data usage, sharing policies, and consumer remedies in tightly-written legal language than the policies of small firms.

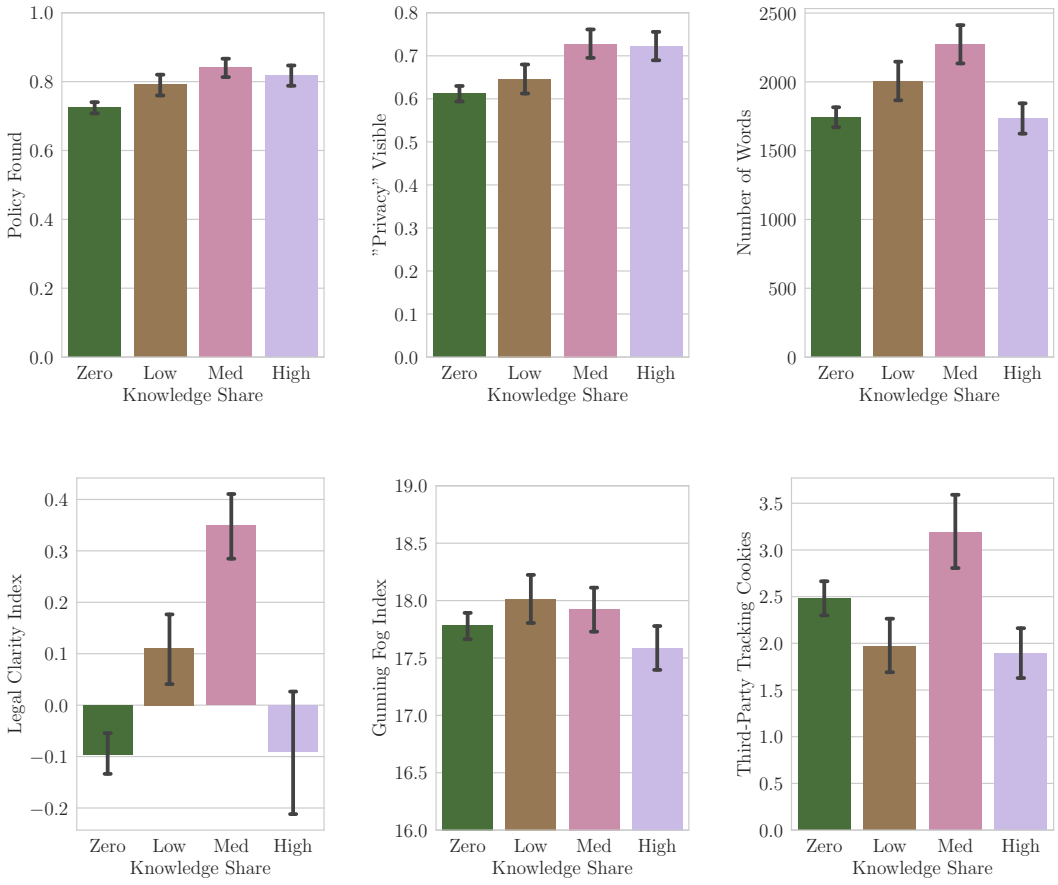
Finally, to provide an independent measure from those derived from the availability and text of the privacy policies, we document the incidence of third-party tracking cookies on firms' websites. This is also significantly higher for larger firms than it is for smaller firms.

We next turn to understanding the relationship between knowledge capital, our preferred proxy for firms' technical sophistication, and attributes of firms' privacy policies. Figure 7 sorts the firms by their share of knowledge capital, and provides a first look.

The figure presents an interesting pattern—there is a monotonic increase in the likelihood of finding a policy, the visibility of the word "Privacy" and the length of the policy, but only for the first three buckets, i.e., for firms with a zero, low, or medium knowledge share. However, there is no significant increase in any of these variables when firms move into the high knowledge share category. Indeed, the length of the policies significantly decreases when moving from medium knowledge share firms to high knowledge share firms.

We also see verification of this non-monotonic relationship when analyzing the text of these policies. The very high knowledge share firms have policies which are less confusing than those with low knowledge shares, and the Legal Clarity of the policies is also dramatically lower for the highest knowledge share firms than for the

Figure 7: Privacy Policies and Knowledge Capital



low knowledge share firms. This is also mirrored in the fact that the highest knowledge share firms have fewer third-party tracking cookies on their websites than the firms with medium knowledge shares.

These findings of a linear relationship with firm size and visible non-linearity with firm technical sophistication motivates further investigation of the underlying economic drivers of the differences in firms' privacy policies. In the next section, we propose a simple theory of data sharing which delivers a set of predictions about relationships between firm characteristics (i.e., firm size and technical sophistication) and firms' approaches to data sharing and the quality of the privacy policies that they choose to write. The theory provides some structure to aid interpretation of the simple relationships that we have detected in this section between firm characteristics and the attributes of their privacy policies, and generates additional predictions which we test subsequently.

We note here that any theory of privacy that needs to fit the attributes of the data faces an interesting challenge, as there are numerous attributes of these contracts that are not perfectly correlated—meaning that the theory will need to simultaneously (and ideally, parsimoniously) explain patterns in the different attributes of these policies.

# 4 A Theory of Data Sharing

In this section, we propose a theory of firms' use of data, their interactions with data intermediaries, and the quality of the privacy policies that they write. Our focus is primarily on firm's data sharing policies, i.e., whether or not firms choose to share customer data with intermediaries, who efficiently monetize these raw data by turning it into prediction-based products.

There are several reasons for this focus. First, most concerns about data privacy in recent policy debates and headlines arise from the power of data-rich third parties,

who can track consumers' browsing behavior across the internet (e.g., Federal Trade Commission, 2014). Second, looking at the data, firm's approaches to third-party sharing are frequently mentioned in their privacy policies, i.e., we see a high incidence of the "third party" bigram in the text of these policies. Finally, in our empirical work, we find that data sharing policies, along with the related category of user rights, are the most important determinants of the overall Legal Clarity of privacy policies in terms of the scores assigned by the human legal expert.

The theoretical model that we set up delivers a simple condition, in terms of three sufficient statistics, which determines whether firms optimally share data with data intermediaries. This condition allows us to interpret empirical relationships between the quality of firms' data policies and the extent to which they share data with third-parties, and firm characteristics such as firm size and technical sophistication.

## 4.1 Setup

**Agents and Data.** We model a firm and a data intermediary. Both can produce signals about an arbitrary state $\theta$ of the economy. Primitive to our model is an indirect utility function arising from such signals $x$ about the state of the economy. Let $V(x)$ be the maximized value of owning a signal $x$.

We make two remarks about $V(x)$ at this stage. First, thinking of $V(x)$ as a maximized value allows us to think of it as incorporating intermediate decisions that the owner of the signal could take, namely: whether to monetize the data in-house or to sell it;[24] whether to garble the signal to extract more surplus (see, e.g., Bergemann et al. (2018); Hörner and Skrzypacz (2016)); whether to purchase additional data externally to enhance the signal, and so on. Second, the ultimate value of information, i.e., the reason why $V(x) > 0$, has many possible origins. Some examples of "micro" uses of $x$, i.e., when it contains information about consumer demand are: (i)

[24]Relatedly, information can be sold indirectly, for example via consumer segmentation services. The structure of $V(.)$ functions is itself fascinating and complex, for example, when signals contain information about common shocks (see Bergemann and Bonatti, 2018).

third-degree price discrimination (i.e., making offers to specific segments of consumers based on the information in $x$, see Montes et al. (2018)); (ii) screening out high cost consumers in selection markets such as insurance and credit; (iii) targeted advertising in search markets.[25]

**Data Sharing.** The firm has three options. A trivial first option is not to use, process, or share its customers' data to produce any signals—in what follows, we describe this as the firm "discarding the data". A second option is that the firm can internally process and use its customers' data to produce a signal $x$, at a cost $\phi$. Third, it can share the data with an intermediary. The intermediary is more efficient at processing data, so does not pay $\phi$, and also produces a better signal using these data, i.e., $y$, with

$$V(y) \geq V(x) \tag{1}$$

In the online appendix, we discuss and explain this condition in more detail. To provide a simple example here, if a signal is used for Bayesian decision making, either by the firm itself or somebody else, then a "better" signal $y$ satisfies (1) if it is more informative than $x$ in the sense of Blackwell (1953) (i.e., $x$ is a "garbled" version of $y$).

For clarity, we focus on this simple market structure, where the intermediary becomes an active player only if the firm decides to share its data. Our model can easily be extended to allow for a richer industrial organisation, for example, where the firm and the intermediary compete in a market for data if they do not work together. We set up such a model in the online appendix, and show that the key sufficient statistics remain the same in such a model.

[25] One can also think of firms' choice of a product line or production technology that needs to be matched to consumers' tastes. See, for example, Veldkamp et al. (2019), who also model "data feedback effects" where firms want to produce more in order to generate data. An alternative modeling approach is to directly include data as a direct input to production functions Jones et al. (2018).

**Costs of Data Sharing.** We assume that if data is sent to an intermediary, the firm faces the risk of future litigation. Such litigation could ensue in the event that the data is misused, or if consumers believe they suffer harm from such sharing.

To help protect itself against this risk, the firm can put in place a privacy policy. A choice variable when writing a privacy policy is its legal "clarity" $q \geq 0$.[26] The legal clarity $q$ of a policy essentially measures its watertightness or soundness. The cost of writing a privacy policy is $\kappa(q)$, where $\kappa(0) \geq 0$ is the fixed cost of hiring a legal team, and $\kappa(q)$ is the variable cost of clarity, where $\kappa'(q) > 0$. The expected loss from future litigation is $L(q)$, where $L'(q) < 0$. As a result of these assumptions, it is costly to share data with the intermediary.

Effectively, the cost of sharing reflects the aversion of consumers to the risk of having their data shared, which is then manifested in the costs of "insuring" the firm legally against such aversion using a privacy policy of high quality.

**Data Ownership and Consumer Behavior.** It is important to note that our model represents an environment where the firm has automatic ownership of its data. The only cost of sharing these data is the potential of future litigation. This is in contrast to a possible alternative where the data is owned by the firm's customers. In this case, the firm needs to satisfy an additional "participation constraint," i.e., consumers' consent needs to be acquired for data collection. This introduces additional upfront costs of data collection, for example, if the firm needs to establish systems for obtaining, evaluating, and monitoring consent, or to offer discounts on products that are associated with heavy data collection.

Our model does not explicitly have such a participation constraint; we view our model as existing in an environment strongly characterized by "consent fatigue". Even if consumers own their data *de jure*, in such an environment, the firm owns the data *de facto* if consumers blindly agree to share their data in order to access services online.

[26] We use the term clarity rather than "quality" here to maintain consistent nomenclature with our empirical work.

We discuss below how the predictions of our model would change in an alternative environment in which customers more actively exercise their rights to their own data.

**Timing.** The timing of events is as follows:

1. The firm decides whether to share its data with the intermediary, process it in-house, or discard it. If the firm discards the data, the game ends.
2. If the firm decides to process its data in-house, then it sells the basic signal $x$ and receives its value $V(x)$.
3. If the firm decides to share its data, then:
   (a) The firm selects the clarity $q$ of its privacy policy and incurs the sunk cost $\kappa(q)$
   (b) The firm bargains with the data intermediary over the surplus generated by sharing its data, as we describe in more detail below. The firm's outside option when bargaining is to process the data in-house, or alternatively to discard the data.
   (c) The firm and the intermediary jointly sell the improved signal $y$ and receive the value $V(y) - L(q)$, where $L(q)$ is the expected loss due to litigation.

## 4.2 Surplus and Bargaining

We can define the total *efficiency advantage* of the intermediary (equivalently, the opportunity cost of processing data in-house) as:

$$C = V(y) - V(x) + \phi$$

This is the sum of the improvement $V(y) - V(x)$ in signal value, and the reduction $\phi$ in the cost of processing data, when it is sent to the intermediary.

We now define the total surplus generated by working with the intermediary. We consider the point in time where the firm bargains with the intermediary (i.e., step 3b in the above timeline). At this point, the clarity $q$ of the firm's privacy policy has been determined, and the costs $\kappa(q)$ of writing it have been sunk. Hence, these costs do not affect the surplus generated.

The total surplus generated by working with the intermediary is therefore:

$$\begin{aligned} S(q) &= \begin{cases} C - L(q), & \text{if } V(x) \geq \phi; \\ V(y) - L(q), & \text{otherwise.} \end{cases} \\ &= \min\{C, V(y)\} - L(q) \end{aligned} \tag{2}$$

There are two cases in the expression for $S(q)$. In the first case, the firm finds it worthwhile to process data in-house, since $V(x) \geq \phi$. Here, the surplus created by working with the intermediary is precisely the efficiency advantage of the intermediary $C$, net of the expected costs of litigation. In the second case, we have $V(x) < \phi$, so that the optimal strategy in the absence of an intermediary would be to discard the data. In this scenario, the surplus created by working with the intermediary is the value $V(y)$ of the new signal. In both cases the final surplus depends on the clarity $q$ of the firm's privacy policy because of its impact on litigation risk $L(q)$.

We assume a Nash bargaining process where the firm receives its outside option plus a share $\mu$ of the surplus generated, where $\mu \in (0, 1)$ is a parameter that measures the firm's bargaining power in negotiations with the intermediary. The firm's gain from negotiating with the intermediary therefore equals $\mu \cdot S(q)$.

## 4.3 Optimal Data Sharing

If the firm decides to share its data, then it will choose the clarity $q$ of its privacy policy to maximize its expected gain, net of the costs of writing the policy, solving

the problem:

$$\max_{q} \{\mu \cdot S(q) - \kappa(q)\} \tag{3}$$

The firm's bargaining power $\mu$ matters in the choice of $q$, because in the bargaining process, the intermediary compensates it for some of the expected loss from litigation. For example, if $\mu = 0$, then the firm always chooses $q = 0$, because all the surplus generated by a better privacy policy would go to the intermediary.

Overall, it is optimal to share data with the intermediary, rather than processing it in-house or discarding it, when the maximized value of problem (3) is positive. Combining this equation with equation (2), we obtain a simple characterization of optimal sharing decisions:

**Proposition 1.** *In equilibrium, the firm shares its data with the intermediary if and only if:*

$$\min\{C, V(y)\} \geq M, \tag{4}$$

*where $C$ is the (opportunity) cost of processing data within the firm, $V$ is the total value of the firm's data, $q^\star = \arg\max \{\mu S(q) - \kappa(q)\}$ is the firm's optimal choice of the quality of its privacy policy, and*

$$M \equiv \frac{\mu L(q^\star) + \kappa(q^\star)}{\mu}$$

*is the cost-benefit tradeoff associated with data-sharing.*

Proposition 1 suggests that firms generally fall into two categories. The first category comprises firms with low-value data $V(y) < C$, who would discard the data in the absence of a more efficient intermediary. These firms decide whether to share based on the total value $V$ of their data. This value is likely to increase both in the size and usefulness of the firm's data (which could, for example, be correlated with the size of the firm's customer base), and with the technical sophistication of the intermediary at processing these data.

The second category is firms with high-value data $V(y) > C$, who decide whether

to share based on the efficiency advantage $C$ of the intermediary. This efficiency advantage (equivalently the opportunity cost of processing data in-house) is also likely to be increasing in the size and usefulness of the firm's data and the technical sophistication of the intermediary, but decreasing in the firm's own technical sophistication.

At this point, we can discuss how the mechanics of the model will change if consumers take active ownership of their data. Consider an alternative setup where, before entering into negotiations with the intermediary, the firm has to ensure that consumers accept its privacy policy and opt-in to providing the firm access to their data. If accepting a policy that covers the firm for a wide range of sharing activities is costly for consumers, then this would effectively entail a larger upfront cost $\kappa(q)$. A full model of this cost function needs to account for any monetary incentives that are required to induce consumers to accept, and the fact that consumers' willingness to accept depends on their expectations about the firms' future data sharing approaches. That having been said, we conjecture that the key sufficient statistics that determine data sharing would be qualitatively similar to the setup which we present here.

## 4.4 Empirical Predictions

Condition (4) clarifies the empirical predictions of the model. A given firm's propensity to share data, and the quality of the privacy policy that it writes depends only on three sufficient statistics: The total value $V(y)$ of the firm's data, the total efficiency advantage $C$ of the data intermediary, and the cost-benefit tradeoff $M$ associated with the choice of clarity of the privacy policy. The predicted relationship between firms' observable characteristics and their decision to share will depend on how firms' observable characteristics map to these sufficient statistics from the model.

We now explore such predictions under a number of assumptions. Motivated by the stylized facts in the previous section, we focus on firm size (as measured by market value) and technical sophistication (as measured by the knowledge share) as two key observable characteristics.

**Assumption 1.** (The value of data) *The total value $V(y)$ of the firm's data is* increasing *in firm size, and* increasing *in the firm's technical sophistication.*

This assumption is natural, for example, in a world in which the value of a dataset is increasing both in the number of observations (rows), and the number of variables (fields) it includes. Larger firms gather more observations on average, as they have more consumers, and firms with greater technical sophistication are able to record more information about each consumer, i.e., they are able to capture a larger number of fields about each consumer, as well as to store customer data in a manner that enables easy processing.

**Assumption 2.** (The efficiency advantage of intermediaries) *The efficiency advantage $C$ of the data intermediary is* increasing *in firm size, but* decreasing *in the firm's technical sophistication.*

Asumption 2 captures the idea that, holding a firm's technical sophistication constant, a generic improvement in processing technology, i.e., increasing the ability of the data intermediary, is more valuable when applied to a large dataset collected by a large firm. On the other hand, holding firm size constant, increases in the firm's technical sophistication allows it to reduce its efficiency disadvantage vis-à-vis the intermediary. This could occur either by making the value $V(x)$ of the firm's in-house signal closer to the maximal value $V(y)$, or by reducing the firm's cost $\phi$ of data processing.

**Assumption 3.** (Cost-benefit trade-offs) *The cost-benefit tradeoff $M$ of data sharing is* weakly decreasing *in firm size, and* independent *of the firm's technical sophistication.*

The first part of Assumption 3 holds, for example, when bargaining power $\mu$ increases more rapidly with firm size and the minimized cost $\mu L(q^\star) + \kappa(q^\star)$ of data sharing moves relatively slowly with size. The second part is true when legal costs and bargaining power are determined by firm characteristics other than its technical systems.

Figure 8: Data Sharing and Technical Sophistication

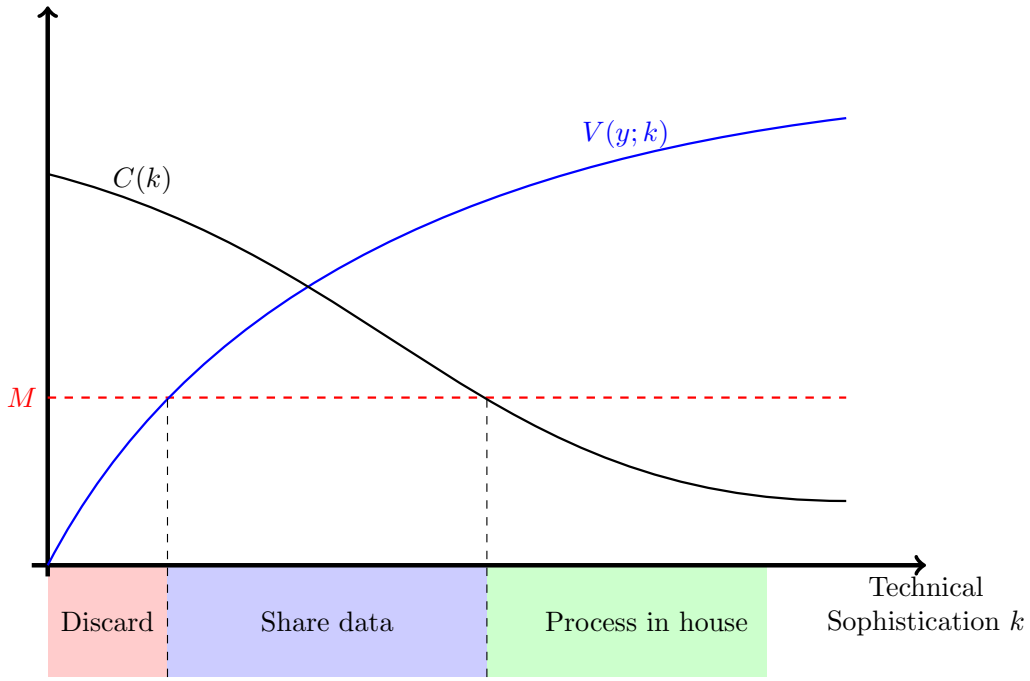



Figure 8 illustrates some subtle effects in the model under these assumptions. On the horizontal axis is a parameter $k$ that indexes the firm's technical sophistication, holding firm size constant. The upward-sloping (recall Assumption 1 above) blue line is the associated value $V(y;k)$ of data. The downward-sloping black line is the efficiency advantage $C(k)$ of the intermediary. Recall that the surplus created by sharing data (and hence the optimal data sharing condition (4)) is determined by the *minimum* of these two lines, which is non-monotonic in $k$. Intuitively, the surplus rises in $k$ for firms that would otherwise discard their data, because the total value of data $V(y;k)$ increases in sophistication. However, the surplus decreases in $k$ for firms that would otherwise process their data in-house, because the efficiency advantage $C(k)$ of the intermediary decreases in the firm's own technical sophistication. Hence, as illustrated by the labels along the horizontal axis, firms with *intermediate* technical sophistication are most likely to share data, whereas firms with the very highest degree of technical sophistication will prefer to process data in-house. The figure also suggests the robustness of this prediction to some differences in underlying assumptions. For example, the second part of Assumption 3 is not needed to obtain this qualitative

pattern, as long as $M$ moves relatively slowly with technical sophistication.

It is also worth noting that in our model, the data intermediary is a separate entity from the firm that accesses customer data, and the firm decides whether or not to share these data with the intermediary. In practice, of course, data intermediaries are also firms, and firms with sufficiently high technical sophistication may also decide to *become* data intermediaries. While we do not model these factors in the interest of keeping the model parsimonious, we note that these forces may well be at play in reality, and that non-linearities could certainly be associated with such choices by firms in an extended version of the model.[27]

We now state the formal predictions of our model under these assumptions:

**Corollary 1.** (Data Sharing, Privacy Policies and Firm Size) *Under Assumptions 1-3, firms' propensity to share data and the quality $q^\star$ of observed privacy policies are increasing in firm size.*

**Corollary 2.** (Data Sharing, Privacy Policies and the Firm's Technical Sophistication) *Under Assumptions 1-3, holding firm size constant, there are two possible scenarios:*

1. *If $V(y) > C$ for all firms, then firms' propensity to share data, and the quality $q^\star$ of observed privacy policies, are monotone decreasing in firms' technical sophistication, holding firm size constant.*

2. *If $V(y) < C$ for some (low sophistication) firms, and $V(y) > C$ for other (high sophistication) firms, then firms' propensity to share data, and the quality $q^\star$ of their observed privacy policies, will increase in technical sophistication for low-sophistication firms, but decrease with technical sophistication for high-sophistication firms.*

[27] Of course, the same issues are likely at work in our empirical results; we note that manual inspection of the firms with the very highest technical sophistication measured in the data reveals that they are a mix of specialized data processing firms (e.g., Destiny Media Technologies, Intrusion Inc.) and firms that are more likely to be processing their own customer data in-house (e.g., Zynga Inc., Kidoz Inc.).

We stress that these predictions in Corollaries 1–2 do not represent all possible predictions of the model. A rejection of the two scenarios in Corollary 2, for example, would not constitute a rejection of our model of data sharing. Instead, it would be a joint rejection of the assumptions about the relationship between the value of data, opportunity costs, sunk costs of sharing, and firm characteristics.

Put differently, estimated relationships between data sharing and firm characteristics in the data can be mapped back to the model through Condition (4), which allows us to interpret these relationships in terms of their implications for the key sufficient statistics in the model. We now return to more formally testing the model's predictions in the data.

# 5 Empirical Tests of Model Predictions

Earlier, we showed in Figures 6 that numerous attributes of privacy policies vary systematically with firm size and technical sophistication. To review, the largest firms have privacy policies that are more likely to be found; "Privacy" is more likely visible on large firms' homepages; large firms' policies are significantly lengthier and more complex; and have higher Legal Clarity scores—which, as discussed in the model, essentially facilitates/insures against expected losses from litigation arising from data sharing. Consistent with our interpretation on data sharing, we also show that large firms have a significantly higher incidence of third-party tracking cookies on their websites. We also find a non-monotonic relationship between firm technical sophistication and both the incidence of third-party sharing and the attributes of privacy policies.

Thus far, the evidence seems to line up well with the predictions of Corollary 1 and 2, which predict a monotone relationship of privacy policy quality and third-party sharing with firm size in both cases, i.e., $V > C$ for all firms, and the case in which there is firm heterogeneity around zero in $V - C$. The evidence also lines up

with the second case in Corollaries 1 and 3. These predict a non-monotonicity in the relationship between privacy policy attributes and firms' technical sophistication if there is significant heterogeneity in firms' technical sophistication. This in turn leads to differences around zero in $V - C$.

Thus far, we have estimated simple univariate relationships, but to test these predictions more clearly, we move to estimating multiple regressions to better pick up the independent patterns in the data. We also use these regressions to evaluate whether the patterns we detect in the univariate relationships apply only across industries, or whether there is also within-industry variation with firm characteristics in privacy policy attributes. We also consider robustness to a variety of different measurement approaches.

## 5.1 Multivariate Relationships: Firm Characteristics and Privacy

Our multivariate regression results are in Table 4. The top panel (a) of the Table contains regressions without industry fixed effects, while the bottom panel (b) of the table contains sector fixed effects at the level of SIC divisions.[28] The columns of the table correspond to the various attributes of the privacy policies that are on the left-hand-side of the regressions. The rows show the variables that are on the right-hand-side.

The first right-hand-side variable is the log Market Value (i.e., size) of each firm. The second is the log Market Share measured as the firm's sales divided by industry sales at the 2-digit SIC code level. In the model, we abstract from any effects of firms' market power relative to their own consumers. We note here that high market power could generate additional incentives not to share data, if the marginal buyer of the firms' core product has a strong concern about privacy—as in Spence (1975); or

[28] See Figure 5 for the definition of SIC divisions. In the online appendix, we show that the estimated coefficients are of similar magnitude, albeit less precisely estimated, with SIC2 level fixed effects.

Table 4: **Policy Attributes: Regressions**

(a) Without Sector Fixed Effects

| | (1) Policy Found | (2) Policy Visible | (3) Log Words | (4) Legal Clarity | (5) Fog Index | (6) $3^{rd}$-Party Trackers |
|---|---|---|---|---|---|---|
| Log Market Value | 0.0421*** | 0.0484*** | -0.00597 | 0.0426*** | 0.0296 | 0.330*** |
| | (12.22) | (12.13) | (-0.61) | (4.44) | (1.14) | (8.20) |
| Knowledge Share | 0.847*** | 0.695*** | 2.405*** | 2.605*** | 0.501 | 4.447*** |
| | (8.33) | (5.89) | (8.80) | (9.78) | (0.69) | (3.76) |
| Knowledge Share$^2$ | -0.813*** | -0.793*** | -2.821*** | -3.811*** | -0.264 | -7.114*** |
| | (-4.90) | (-4.12) | (-6.30) | (-8.74) | (-0.22) | (-3.69) |
| Log Market Share | 0.0157*** | -0.0105*** | 0.0874*** | 0.0615*** | 0.100*** | 0.119*** |
| | (5.41) | (-3.11) | (10.49) | (7.57) | (4.54) | (3.52) |
| Observations | 5140 | 5140 | 3918 | 3918 | 3918 | 4951 |

(b) With Sector Fixed Effects

| | (1) Policy Found | (2) Policy Visible | (3) Log Words | (4) Legal Clarity | (5) Fog Index | (6) $3^{rd}$-Party Trackers |
|---|---|---|---|---|---|---|
| Log Market Value | 0.0430*** | 0.0423*** | 0.0163 | 0.0417* | 0.0541 | 0.330* |
| | (8.31) | (5.53) | (0.75) | (1.88) | (0.76) | (2.07) |
| Knowledge Share | 0.659*** | 0.463** | 1.502** | 2.260*** | 0.759 | 4.968*** |
| | (12.66) | (3.08) | (2.60) | (6.34) | (0.42) | (3.31) |
| Knowledge Share$^2$ | -0.580*** | -0.482*** | -1.795** | -3.283*** | -0.299 | -6.852** |
| | (-12.91) | (-5.31) | (-2.75) | (-6.20) | (-0.18) | (-3.18) |
| Log Market Share | 0.0138** | -0.00547 | 0.0619* | 0.0618*** | 0.0878 | 0.110 |
| | (2.48) | (-0.41) | (2.14) | (5.73) | (0.78) | (1.02) |
| Observations | 5140 | 5140 | 3918 | 3918 | 3918 | 4951 |

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b).

the reverse, if having high market share increases the firm's bargaining power $\mu$ vis-à-vis the intermediary. To hold these effects constant while assessing the predictions in Corollary 1, we therefore control for the firm's market share throughout. We also include the knowledge share, i.e., the share of the firm's knowledge capital as a fraction of its total capital, and its square, to capture the nonlinearity we detected earlier in Figure 7.

The table shows that for virtually all of the privacy policy attributes, there is a positive and statistically significant relationship with firm size, and that this relationship appears to hold within industries as well as in the specification without fixed effects. There is attenuation in the statistical significance of the coefficients in some cases with the introduction of the fixed effects, but no attenuation in the economic magnitude of the coefficients, suggesting that this is primarily a power issue rather than an issue of between-industry variation being the proximate source of variation.

The knowledge share also continues to have a positive and statistically significant relationship with the policy attributes both with and without the inclusion of industry fixed effects, and the nonlinearity also shows up clearly in this case—the coefficient on the squared knowledge share is always negative and almost always statistically significant in the attributes for all regressions. Given the low correlations between the attributes seen in Figure 4,the consistent signs on both size and knowledge share across specifications which seek to explain different privacy contract attributes are noteworthy.

**Robustness** We conduct several robustness checks on these results in the online appendix:

- We confirm that the results are qualitatively unchanged when we employ an alternative specification in which policy length is included as a control variable rather than an outcome, further reinforcing that the results are not simply a manifestation of a single common dimension of the privacy policies—and that

there is independent explanatory power of firm characteristics for the residual variation in each policy's Legal Clarity, even after controlling for policy length.

- We reconfirm that these results hold when we control for a broader set of firm characteristics, including firms' marketing expenditures as a fraction of total assets, and firms' market-to-book ratios, as additional control variables.

- When we exclude manufacturing firms from the dataset, and focus only on non-manufacturing firms, we show that all of these patterns become substantially stronger, as might be expected given that manufacturing firms are less likely to be participants in the data sharing economy than services firms.

- We further confirm that the relationship between firm characteristics and third-party sharing persists at the extensive margin, i.e., large firms, and firms with intermediate knowledge share are also most likely to *allow* third party cookies.

- We show that the relationship between policy attributes and firm characteristics is not sensitive to the construction of our Legal Clarity index. Indeed, the results are qualitatively the same if we consider an alternative measure of Legal Clarity, which is based on a supervised learning model trained on the legal expert classifications.

Overall, our findings appear to line up with the second case described in Corollaries 1 and 2 of the model, in which there are pronounced differences between firms in their level of technical sophistication. In the data it appears that some firms have $V < C$, and the model predicts that such firms will have higher propensities to share data, and to write higher quality privacy policies. In contrast, other firms are more highly technically sophisticated, have $V > C$, and consequently lower propensities to share data. This in turn means that they will write lower quality privacy policies, preferring to incur costs $\phi$ and exploit their data in-house rather than sharing it with the intermediary.

# 6 Conclusion

In this paper, we take a first look at a large set of US firms' privacy policies, and bring new facts and analysis to the study of the market for data privacy.

We find that there is significant variation in the ease of acquiring and finding firms' privacy policies, and that when found, these policies do not follow a standard boilerplate. Instead, their text varies substantially both within and across industries. We find that this variation in policy text is systematic, with large firms and those with high levels of knowledge capital exhibiting longer and more complex policies with ostensibly more clearly specified and legally sound protections outlined in their text. However, we find that the firms with these characteristics also have websites with a *higher* incidence of tracking cookies from third-parties. This variation is both between and within industries.

To explain these new facts, we set up a simple theory of data acquisition and usage, in which firms optimally decide whether to process their own data in-house, or to sell these data to a third-party data intermediary for processing. In the model, firms determine the legal clarity and watertightness of the privacy policies that they write in order to insure themselves against future legal liability arising from such data sharing. Put differently, in the model, legal clarity in policies is a device to legitimize and facilitate third-party data sharing.

The model delivers predictions about the relationship between firm size, knowledge capital intensity, and the incidence of third-party sharing. While the theory predicts that firm size will be positively correlated with the incidence of third-party sharing and the legal clarity of firms privacy policies, it also predicts that firms with the very highest technical sophistication will choose to process data in-house rather than share it with third-parties. Consequently, such highly technically sophisticated firms will not need to write sound privacy policies. Consistent with our theoretical predictions, we find that large firms with intermediate knowledge capital intensity have longer, more legally watertight policies, but are more likely to share data on their users'

browsing history with third parties. However, firms with the very highest knowledge capital intensity have shorter, less complex, and less legally watertight policies, and simultaneously engage in less third-party sharing of user data from their websites.

We view our findings in this paper as a first step towards a broader and deeper empirical and theoretical analysis of data privacy policies, and hope to continue to refine our insights about this important area going forward.

# References


Acquisti, A., L. Brandimarte, and G. Loewenstein (2015): "Privacy and human behavior in the age of information," *Science*, 347, 509–514.

Acquisti, A., C. Taylor, and L. Wagman (2016): "The economics of privacy," *Journal of Economic Literature*, 54, 442–92.

Admati, A. R. and P. Pfleiderer (1986): "A monopolistic market for information," *Journal of Economic Theory*, 39, 400–438.

Agarwal, A., J. Gans, and A. Goldfarb (2018): *Prediction Machines: The simple economics of artificial intelligence*, Harvard Business Press.

Begenau, J., M. Farboodi, and L. Veldkamp (2018): "Big data in finance and the growth of large firms," *Journal of Monetary Economics*, 97, 71–87.

Bergemann, D. and A. Bonatti (2018): "Markets for information: An introduction," .

Bergemann, D., A. Bonatti, and A. Smolin (2018): "The design and price of information," *American Economic Review*, 108, 1–48.

Blackwell, D. (1953): "Equivalent Comparison of Experiments," *Annals of Mathematical Statistics*, 24, 265–272.

Calzolari, G. and A. Pavan (2006): "On the optimality of privacy in sequential contracting," *Journal of Economic theory*, 130, 168–204.

Crouzet, N. and J. Eberly (2018): "Intangibles, Investment, and Efficiency," *AEA Papers and Proceedings, 108 : 426-31.*

Daughety, A. F. and J. F. Reinganum (2010): "Public goods, social pressure, and the choice between privacy and publicity," *American Economic Journal: Microeconomics*, 2, 191–221.

Eisfeldt, A. and D. Papanikolau (2014): "The value and ownership of intangible capital," *American Economic Review: Papers and Proceedings 104, 1-8*.

Englehardt, S. and A. Narayanan (2016): "Online tracking: A 1-million-site measurement and analysis," in *Proceedings of ACM CCS 2016*.

Eső, P. and B. Szentes (2007): "Optimal information disclosure in auctions and the handicap auction," *The Review of Economic Studies*, 74, 705–731.

Fabian, B., T. Ermakova, and T. Lentz (2017): "Large-scale readability analysis of privacy policies," in *Proceedings of the International Conference on Web Intelligence*, ACM, 18–25.

Farboodi, M. and L. Veldkamp (2017): "Long run growth of financial technology," Tech. rep., National Bureau of Economic Research.

Federal Trade Commission (2014): "Data Brokers: A Call for Transparency and Accountability," Tech. rep.

Finkel, J. R., T. Grenager, and C. Manning (2005): "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*, Association for Computational Linguistics, 363–370.

Gentzkow, M., B. Kelly, and M. Taddy (2018): "Text as Data," .

Goldfarb, A. and C. Tucker (2011): "Online display advertising: Targeting and obtrusiveness," *Marketing Science*, 30, 389–404.

——— (2012): "Shifts in privacy concerns," *American Economic Review*, 102, 349–53.

Gunning, R. (1952): *The technique of clear writing*, McGraw-Hill, New York.

Hirshleifer, J. (1971): "The private and social value of information and the reward to inventive activity," *American Economic Review*, 61, 561–574.

Hörner, J. and A. Skrzypacz (2016): "Selling information," *Journal of Political Economy*, 124, 1515–1562.

Jolls, C. (2012): "Privacy and consent over time: the role of agreement in Fourth Amendment analysis," *Wm. & Mary L. Rev.*, 54, 1693.

Jones, C., C. Tonetti, et al. (2018): "Nonrivalry and the Economics of Data," in *Society for Economic Dynamics 2018 Meeting Papers*, vol. 477.

Krishnamurthy, B. and C. Wills (2009): "Privacy diffusion on the web: a longitudinal perspective," in *Proceedings of the 18th international conference on World wide web*, ACM, 541–550.

McLaughlin, G. H. (1969): "SMOG grading - a new readability formula," *Journal of Reading*, 12, 639–646.

Montes, R., W. Sand-Zantman, and T. M. Valletti (2018): "The value of personal information in markets with endogenous privacy," *Management Science*, 65.

Peters, R. H. and L. A. Taylor (2017): "Intangible capital and the investment-q relation," *Journal of Financial Economics*, 123, 251–272.

Posner, R. A. (1981): "The economics of privacy," *The American economic review*, 71, 405–409.

Rajaraman, A. and J. D. Ullman (2011): *Mining of massive datasets*, Cambridge University Press.

Spence, A. M. (1975): "Monopoly, Quality, and Regulation," *Bell Journal of Economics*, 6, 417–429.

Stigler, G. J. (1980): "An introduction to privacy in economics and politics," *The Journal of Legal Studies*, 9, 623–644.

Taylor, C. R. (2004): "Consumer privacy and the market for customer information," *RAND Journal of Economics*, 631–650.

Varian, H. R. (2009): "Economic aspects of personal privacy," in *Internet policy and economics*, Springer, 101–109.

——— (2010): "Computer mediated transactions," *American Economic Review*, 100, 1–10.

Veldkamp, L., M. Farboodi, R. Mihet, and T. Philippon (2019): "Big Data and Firm Dynamics," .

Westin, A. F. and O. M. Ruebhausen (1967): *Privacy and freedom*, vol. 1, Atheneum New York.

# Online Appendix: The Market for Data Privacy

**Tarun Ramadorai, Antoine Uettwiller, and Ansgar Walther**[1]

This draft: June 2019

[1]Ramadorai: Imperial College London and CEPR. Email: t.ramadorai@imperial.ac.uk. Uettwiller: Imperial College London. Email: a.uettwiller17@imperial.ac.uk. Walther: Imperial College London. Email: a.walther@imperial.ac.uk.

# Contents

# 1 Dimensions of Expert Evaluation

## DIMENSION 1: DATA COLLECTION

'Data Collection' clauses describe data gathering techniques, specifying what type of data is collected, and when or how it is collected and stored. Scoring was based on 1) clarity of the clauses, and 2) comprehensiveness of the data collected.

A high score meant that the data collection clauses were clear, and/or that the data collected was purposefully minimal. Sometimes these might conflict wherein a policy would state they collect comprehensive data for 'the sake of clarity and transparency – when this was the case, a policy was given a high score when the types of data collected were reasonable and in-line with industry standards and necessity.

A low score meant that the data collection clauses were either unclear, collected data so comprehensively to the point that it seemed unreasonable or excessive, or did not specify what type of data would be collected in sufficient detail that the user would not understand what data they are providing. A neutral score meant that the data collection clauses were sufficiently though not especially clear, and resembled a standard policy.

Examples:

- High score policy: Trinity Biotech – limited data collection, specific reference to their website mechanisms, implied exclusion of other types of collection
- Low score policy: Intuit – extensive data collection including location, camera, and contact data

## DIMENSION 2: CONSENT

'Consent' clauses specified where the policy was presuming the consent of the user, and was sometimes also used to identify where the organization expressly mentioned the legal basis they relied on for data processing.

Scoring was based on how onerous the presumed consent was on the user.

A high score meant that the consent clause stated it would ask specifically for consent for different processes, and would proactively notify the user of any changes to the policy.

A low score meant that the consent clause presumed the user's consent from their continued use (sometimes aggressively disclaiming their liability), and/or required the user to frequently check and review the policy with each use.

A neutral score meant that the consent clauses were resembled a standard policy. This often meant that consent was presumed but was not aggressively framed.

Examples:

- High score policy: WWE – no presumption of consent, will proactively inform users of changes via email or clear notice prior to the change taking effect
- Low score policy: Zynerba – presumed consent and onerous on user by requiring them to check the policy with each use

## DIMENSION 3: RESPONSIBLE USE

'Responsible Use' clauses describe how the organization will use or interact with the data, specifying any services, security measures, marketing, or other internal use that the data will be subject to. By nature, this dimension casts a wider net than the others.

Responsible Use also covers the use of particular tracking or monitoring techniques such as cookies or other third-party software. These exist on a boundary between Data Collection and Responsible Use, but was grouped with Responsible Use because the clauses are often found separate from other data collection clauses and will detail the function of those tools.

Scoring was based on 1) whether the use was either limited and favourable for the user or extensive and favourable for the organization, and 2) the extent of the use of additional tracking and monitoring tools.

A high score meant that the responsible use clauses proactively offered the user clear benefits and robust security assurance, and/or limited and specific use of user data. Further, a high score would indicate reasonable or restricted additional tracking and monitoring tools.

A low score meant that the responsible use clauses specified extensive use of user data, and/or subjection to heavy advertising and additional services. Further, a low score would indicate extensive additional tracking and monitoring tools.

A neutral score meant that the responsible use clauses were reasonable and resembled a standard policy.

Examples:

- High score policy: Palo Alto Networks – clear and specific explanation of use with limited use of additional tracking and monitoring tools
- Low score policy: Insight – extensive uses of user data and extensive use of additional tracking and monitoring tools including third-party tools

## DIMENSION 4: THIRD-PARTIES

'Third-Parties' clauses describe how the organization will share user data with third-parties, and what liability they accept or reject for that sharing. There are some categories of third-party sharing that are unavoidable, for example for the purposes of law enforcement, and then other reasons such as contract fulfilment, marketing purposes, and business interests that were assessed.

Scoring was based on 1) how clearly the sharing protocols were explained, 2) whether the sharing was restricted or not, and 3) whether the organization retained any liability or responsibility over the shared data.

A high score meant that third-party sharing was clearly explained, minimal in practice, and purely out of necessity. Further, it might indicate that the organization retained liability and responsibility over the shared data. While it is standard to disclaim liability for the actions of third-parties, some policies that scored well made an attempt to exercise some responsibility over the shared data to protect the user.

A low score meant that third-party sharing was unclear or poorly explained, leaving the user to wonder about the safety and use of their data. Further, it might indicate that sharing was extensive and not necessarily for the interest or benefit of the user. While it is standard to disclaim liability for the actions of third-parties, some policies that scored poorly did not make any attempt to protect the user or their data.

A neutral score meant that the third-party clauses were reasonable and resembled a standard policy, often disclaiming liability in unassuming terms.

Examples:

- High score policy: Vuzix – third-party sharing was clearly explained, limited to legitimate reasons for sharing, and at least attempts to impose some standards on sharing partners to protect user data in good faith
- Low score policy: Aps – extensive sharing with third-parties for marketing purposes and no attempt to impose standards on their sharing partners

## DIMENSION 5: USER-RIGHTS

'User-Rights' clauses describe what protection and remedies users have in response to the organization's use of their personal data. This includes clauses about data retention and deletion, information redress, requests for access, and complaint procedure. Although the GDPR does not directly apply to these American websites, the GDPR still inspired this dimension insofar as it provides users with the tools and language to understand what rights they may exercise over their data. Clauses that explained opt-out clauses were also included.

Scoring was based on 1) whether or not users were granted any rights over their data, 2) how clearly these rights were explained, and 3) how simple it was for users to exercise these rights.

A high score meant that significant and comprehensive rights were conferred onto the user over their data. It might also indicate that the rights were clearly explained and that the organization was forthcoming in providing users with a straightforward avenue to address any issues.

A low score meant that no rights were conferred at all onto the user, or if they were, they were minimal, poorly explained, or difficult and inaccessible for users to actually put in effect.

A neutral score meant that the user-rights clauses offered some reasonable form of redress that is neither particularly forthcoming nor overly minimalistic.

Examples:

- High score policy: Huron Consulting Group – rights are clearly identified and conferred onto the user. They are laid out in order and all in one place at the end of the policy.
- Low score policy: Marcus Millichap – some user rights loosely explained throughout policy but the totality of the user's exercisable rights is unclear and not found in one place as with the vast majority of other policies.

## DIMENSION 6: OVERALL

This dimension is the culmination of the other dimensions, and introduces perhaps the most interesting part of the data but also the most subjective to human error. The entire policy was considered as a whole, more than simply the sum of the other dimensions, because it accounted for each policy's tone, clarity, and style, all assessed from the perspective of whether a lay-user would be able to understand the policy and be able to exercise their data rights from it.

Scoring was based on 1) the overarching tone and legibility of the policy, and 2) how many of the other dimensions were positively or negatively scored.

Most policies that had a high overall score were clear throughout and had minimal shortcomings, whereas most policies that poor overall were lacking throughout.

There was room for discrepancy in some policies that received mixed evaluations across the dimensions. For example, a policy with too many negatively scoring dimensions would have been difficult to read or use as a whole, but where only one aspect of an otherwise good policy fell-short, it could still be a relatively effective policy for users.

By contrast, a policy that was relatively clear or at least measured up to standard in most dimensions, but was very poor in a crucial dimension such as user-rights, the policy may be scored negatively overall because it would be difficult for a user to apply towards protecting their data rights.

Examples:

- High score policy: Image Sensing – clearly laid out with different sections for each crucial aspect of the policy
- Low score policy: Tabularasa Healthcare – policy offers almost nothing meaningful for the user and is only there to vaguely disclaim any liability

# 2 Word Clouds of Good and Bad Policies

Figure 1: **Word cloud of high and low score policies: data collection**

(a) Good data collection score


(b) Bad data collection score


Note: Bigrams are scaled by the difference between the average TF.IDF score among policies evaluated as high (low) by a legal expert and the grand average TF.IDF score in our sample.

Figure 2: **Word cloud of high and low score policies: consent**

(a) Good consent score


(b) Bad consent score


Note: Bigrams are scaled by the difference between the average TF.IDF score among policies evaluated as high (low ) by a legal expert and the grand average TF.IDF score in our sample.

Figure 3: **Word cloud of high and low score policies: responsible use**

(a) Good responsible use score


(b) Bad responsible use score


Note: Bigrams are scaled by the difference between the average TF.IDF score among policies evaluated as high (low ) by a legal expert and the grand average TF.IDF score in our sample.

Figure 4: **Word cloud of high and low score policies: third-party sharing**

(a) Good third-party sharing score


(b) Bad third-party sharing score


Note: Bigrams are scaled by the difference between the average TF.IDF score among policies evaluated as high (low ) by a legal expert and the grand average TF.IDF score in our sample.

Figure 5: **Word cloud of high and low score policies: user rights**

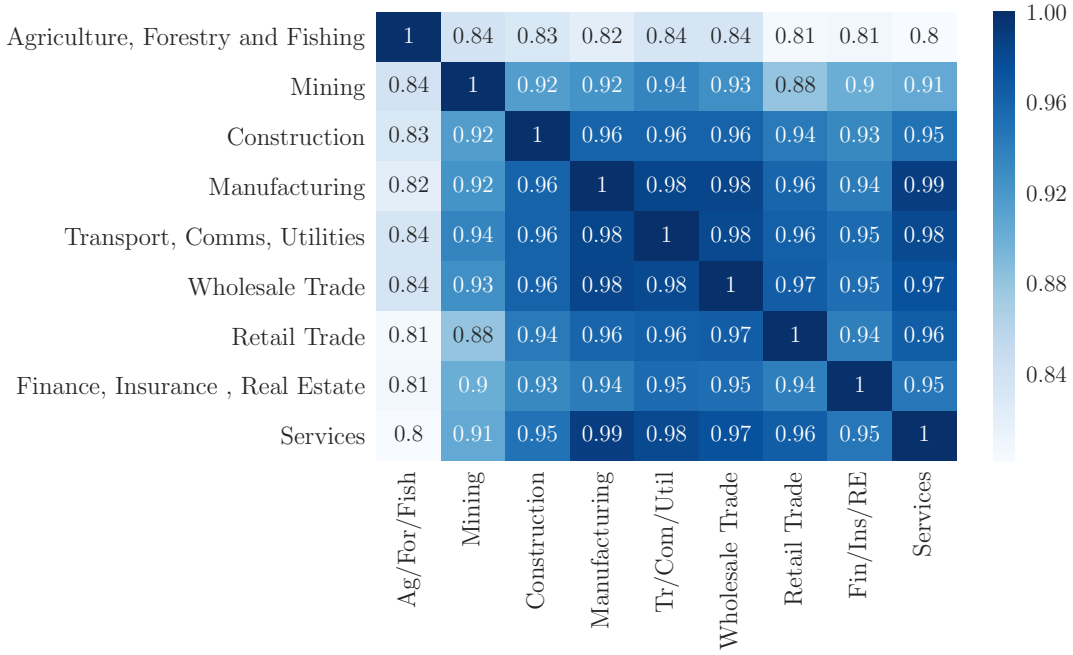
(a) Good user rights score


(b) Bad user rights score


Note: Bigrams are scaled by the difference between the average TF.IDF score among policies evaluated as high (low ) by a legal expert and the grand average TF.IDF score in our sample.

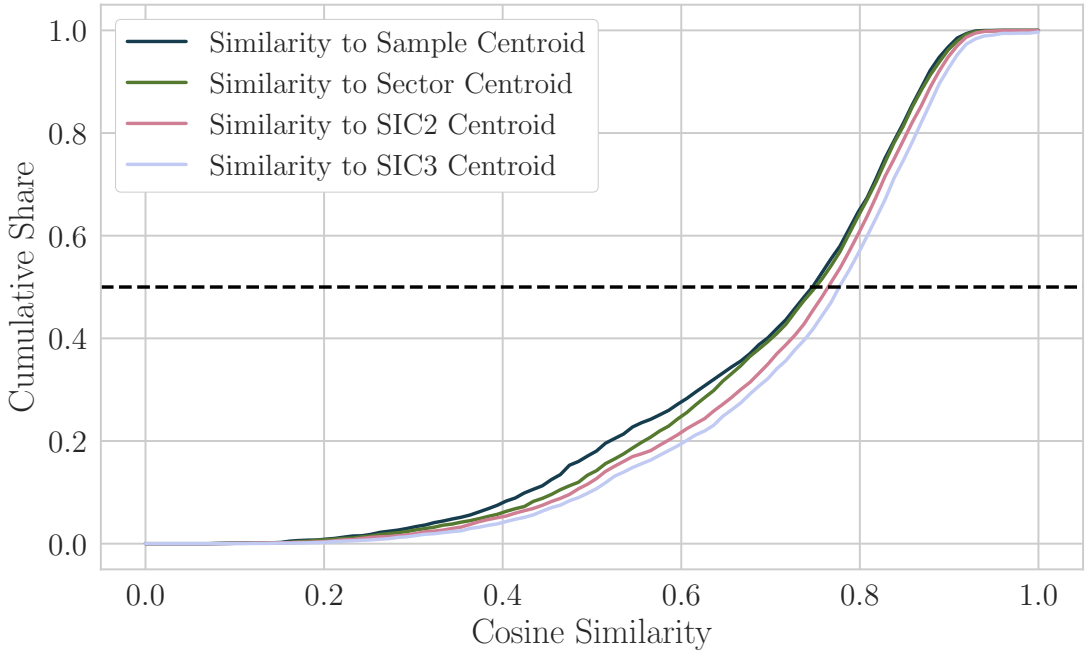
# 3 Variation in Policy Text Between Industries

Figure 6: Variation in Policy Text Between Industries



# 4 Variation in Policy Text with LSA

Figure 7: Variation in Policy Text in 100 Latent Dimensions



# 5 Additional Results on Policy Attributes

Figure 8: Privacy Policies and Market Share

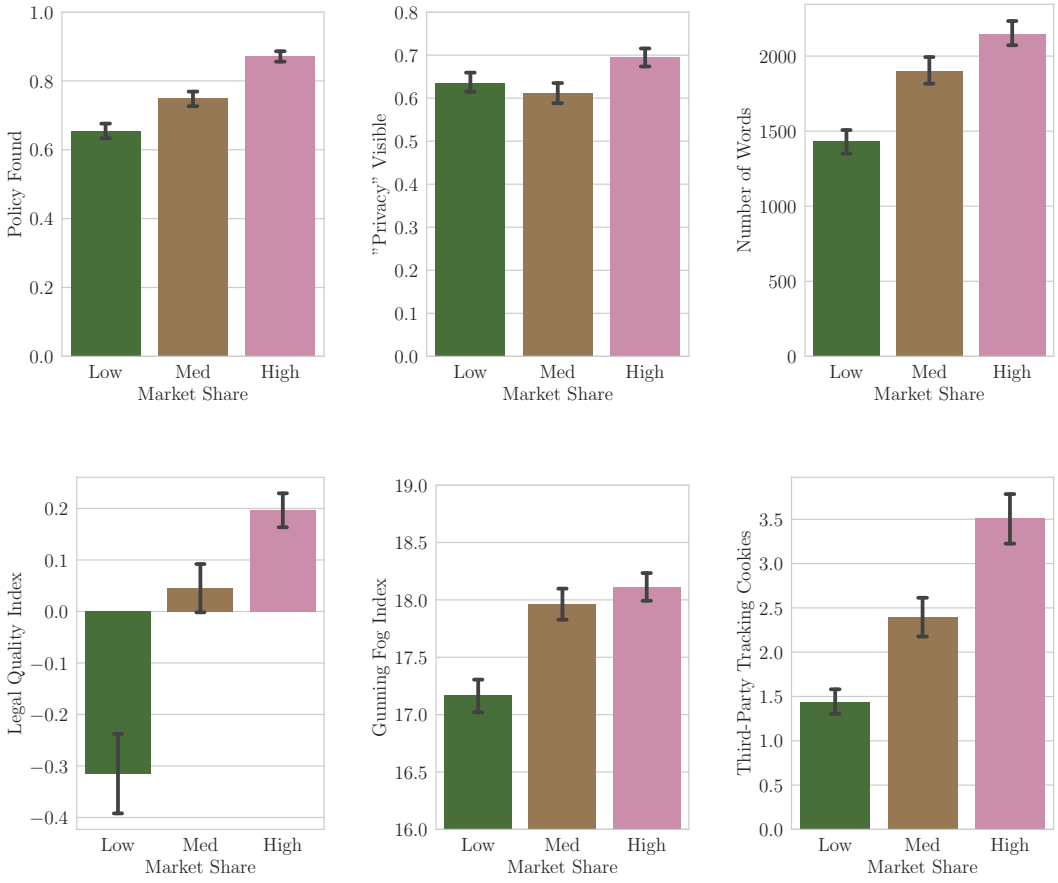


Figure 9: Privacy Policies and Intangible Share

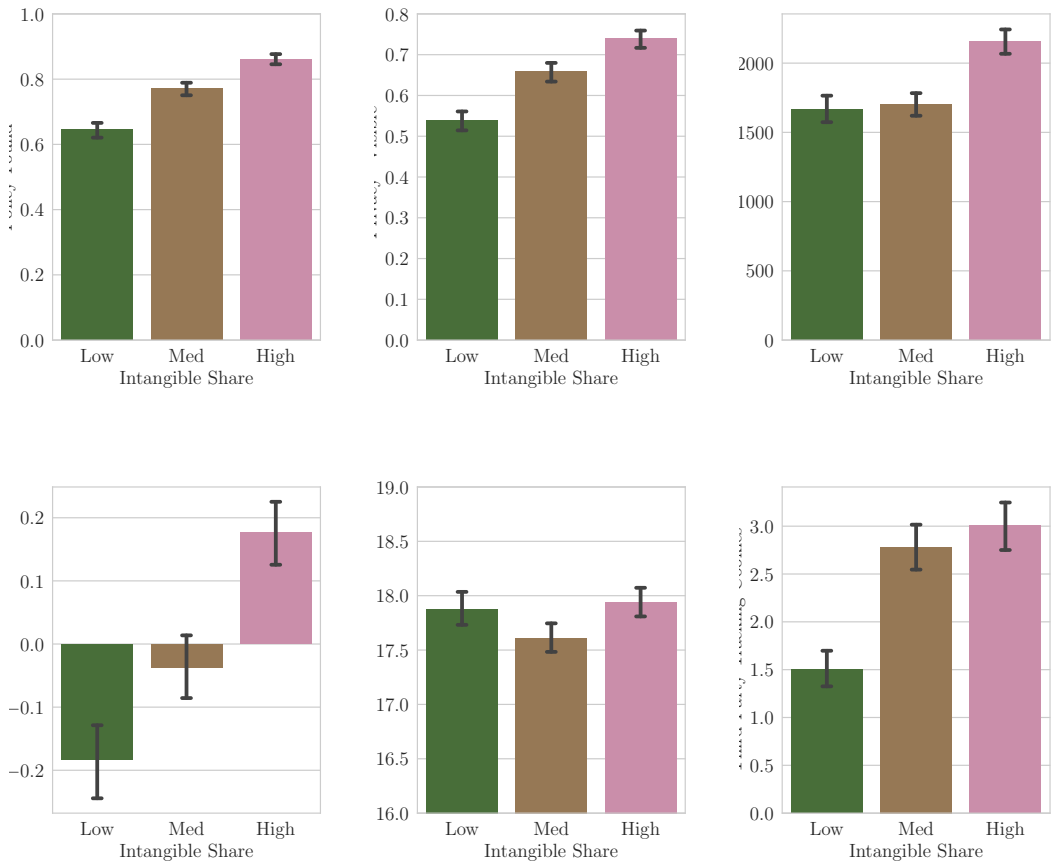


Table 1: **Policy Attributes: Controlling for Policy Length**

(a) Without Sector Fixed Effects

| | (1) Legal Clarity | (2) Fog Index | (3) $3^{rd}$-Party Trackers |
|---|---|---|---|
| Log Market Value | 0.0449*** | 0.0349 | 0.298*** |
| | (5.09) | (1.42) | (5.71) |
| Knowledge Share | 1.682*** | -1.622** | 2.056 |
| | (6.80) | (-2.36) | (1.41) |
| Knowledge Share$^2$ | -2.728*** | 2.226** | -4.840** |
| | (-6.77) | (1.98) | (-2.04) |
| Log Market Share | 0.0279*** | 0.0229 | 0.0824* |
| | (3.69) | (1.09) | (1.84) |
| Log Words | 0.384*** | 0.883*** | 0.568*** |
| | (26.82) | (22.14) | (6.71) |
| Observations | 3918 | 3918 | 3798 |

(b) With Sector Fixed Effects

| | (1) Legal Clarity | (2) Fog Index | (3) $3^{rd}$-Party Trackers |
|---|---|---|---|
| Log Market Value | 0.0355* | 0.0398 | 0.293 |
| | (2.11) | (0.74) | (1.42) |
| Knowledge Share | 1.689*** | -0.561 | 4.199** |
| | (6.37) | (-0.40) | (2.82) |
| Knowledge Share$^2$ | -2.601*** | 1.279 | -6.243** |
| | (-6.60) | (1.06) | (-2.91) |
| Log Market Share | 0.0382*** | 0.0333 | 0.102 |
| | (4.85) | (0.37) | (0.73) |
| Log Words | 0.380*** | 0.879*** | 0.494** |
| | (4.76) | (7.71) | (3.13) |
| Observations | 3918 | 3918 | 3798 |

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b).

Table 2: **Policy Attributes: Further Firm Controls**

(a) Without Sector Fixed Effects

| | (1) Policy Found | (2) Policy Visible | (3) Log Words | (4) Legal Clarity | (5) Fog Index | (6) $3^{rd}$-Party Trackers |
|---|---|---|---|---|---|---|
| Log Market Value | 0.0439*** | 0.0508*** | -0.00928 | 0.0443*** | 0.0214 | 0.321*** |
| | (12.25) | (12.30) | (-0.90) | (4.39) | (0.79) | (7.79) |
| Knowledge Share | 0.715*** | 0.593*** | 1.083*** | 1.676*** | -0.417 | 7.094*** |
| | (5.21) | (3.74) | (2.93) | (4.62) | (-0.43) | (4.49) |
| Knowledge Share$^2$ | -0.596*** | -0.531** | -1.343** | -2.689*** | -0.0426 | -8.688*** |
| | (-3.03) | (-2.34) | (-2.53) | (-5.17) | (-0.03) | (-3.85) |
| Zero Knowledge Capital | -0.0227 | -0.0180 | -0.170*** | -0.153*** | -0.185 | 0.793*** |
| | (-1.35) | (-0.93) | (-3.70) | (-3.40) | (-1.52) | (4.09) |
| Log Market Share | 0.0142*** | -0.0123*** | 0.0856*** | 0.0572*** | 0.0961*** | 0.146*** |
| | (4.63) | (-3.47) | (9.82) | (6.69) | (4.17) | (4.15) |
| Marketing / Assets | 0.328 | 0.125 | 2.936*** | 1.399** | 0.669 | 8.339*** |
| | (1.51) | (0.50) | (5.11) | (2.48) | (0.44) | (3.31) |
| Marketing Missing | -0.0455*** | -0.125*** | -0.0238 | -0.0493 | 0.798*** | -1.680*** |
| | (-3.61) | (-8.62) | (-0.69) | (-1.46) | (8.81) | (-11.58) |
| Market_to_book | -0.00716** | -0.0131*** | 0.0341*** | 0.00163 | 0.0482** | 0.0309 |
| | (-2.44) | (-3.86) | (3.94) | (0.19) | (2.11) | (0.92) |
| Observations | 5109 | 5109 | 3902 | 3902 | 3902 | 4920 |

(b) With Sector Fixed Effects

| | (1) Policy Found | (2) Policy Visible | (3) Log Words | (4) Legal Clarity | (5) Fog Index | (6) $3^{rd}$-Party Trackers |
|---|---|---|---|---|---|---|
| Log Market Value | 0.0462*** | 0.0470*** | 0.0109 | 0.0423* | 0.0365 | 0.348** |
| | (7.94) | (6.28) | (0.42) | (1.93) | (0.58) | (3.00) |
| Knowledge Share | 0.651*** | 0.483*** | 0.907** | 1.647*** | -0.0939 | 6.491*** |
| | (13.83) | (4.16) | (2.95) | (3.83) | (-0.07) | (4.53) |
| Knowledge Share$^2$ | -0.514*** | -0.401** | -1.087*** | -2.511*** | 0.0802 | -7.445*** |
| | (-8.05) | (-2.69) | (-3.83) | (-4.44) | (0.06) | (-3.35) |
| Zero Knowledge Capital | -0.00694 | 0.00925 | -0.119** | -0.175 | -0.326* | 0.721*** |
| | (-0.90) | (0.28) | (-2.27) | (-1.78) | (-1.94) | (5.39) |
| Log Market Share | 0.0111** | -0.00917 | 0.0658* | 0.0597*** | 0.0927 | 0.113* |
| | (2.59) | (-0.87) | (2.06) | (5.12) | (1.02) | (1.92) |
| Marketing / Assets | 0.195 | 0.203 | 2.347*** | 1.153*** | -0.000740 | 5.969 |
| | (1.04) | (0.51) | (4.70) | (5.79) | (-0.00) | (1.12) |
| Marketing Missing | -0.0294 | -0.0850* | -0.0293 | -0.00442 | 0.687** | -1.439*** |
| | (-1.50) | (-1.85) | (-0.32) | (-0.13) | (2.37) | (-5.79) |
| Market_to_book | -0.0102** | -0.0139** | 0.0194 | -0.00619 | 0.0341 | -0.0150 |
| | (-2.70) | (-2.62) | (1.20) | (-0.76) | (0.90) | (-0.51) |
| Observations | 5109 | 5109 | 3902 | 3902 | 3902 | 4920 |

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b).

Table 3: **Policy Attributes: Excluding Manufacturing Firms**

(a) Without Sector Fixed Effects

| | (1) Policy Found | (2) Policy Visible | (3) Log Words | (4) Overall Score | (5) Fog Index | (6) $3^{rd}$-Party Trackers |
|---|---|---|---|---|---|---|
| Log Market Value | 0.0417*** | 0.0521*** | -0.0317** | 0.0254** | -0.0379 | 0.462*** |
| | (9.29) | (10.20) | (-2.47) | (2.39) | (-1.11) | (8.52) |
| Log Market Share | 0.0122*** | -0.0204*** | 0.107*** | 0.0584*** | 0.176*** | 0.0496 |
| | (3.27) | (-4.81) | (10.05) | (6.62) | (6.21) | (1.10) |
| Knowledge Share | 1.205*** | 0.971*** | 4.722*** | 4.361*** | 5.190*** | 17.22*** |
| | (5.38) | (3.81) | (7.91) | (8.84) | (3.28) | (6.44) |
| Knowledge Share$^2$ | -1.351*** | -1.328** | -6.668*** | -6.100*** | -7.560** | -28.43*** |
| | (-2.63) | (-2.27) | (-4.71) | (-5.21) | (-2.01) | (-4.57) |
| Observations | 3380 | 3380 | 2515 | 2515 | 2515 | 3232 |

(b) With Sector Fixed Effects

| | (1) Policy Found | (2) Policy Visible | (3) Log Words | (4) Overall Score | (5) Fog Index | (6) $3^{rd}$-Party Trackers |
|---|---|---|---|---|---|---|
| Log Market Value | 0.0435*** | 0.0426** | -0.00258 | 0.0226 | -0.0153 | 0.505*** |
| | (5.06) | (3.20) | (-0.08) | (0.68) | (-0.15) | (3.39) |
| Log Market Share | 0.00866 | -0.0131 | 0.0748 | 0.0579** | 0.169 | -0.0236 |
| | (1.40) | (-0.69) | (1.62) | (2.60) | (1.07) | (-0.27) |
| Knowledge Share | 0.673*** | 0.422 | 2.754*** | 2.979*** | 4.006 | 8.995** |
| | (4.90) | (1.80) | (4.16) | (14.07) | (1.70) | (3.18) |
| Knowledge Share$^2$ | -0.635* | -0.622 | -3.817*** | -4.178*** | -5.466 | -17.27** |
| | (-2.25) | (-1.54) | (-4.18) | (-6.48) | (-1.63) | (-3.28) |
| Observations | 3380 | 3380 | 2515 | 2515 | 2515 | 3232 |

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b).

Table 4: **Policy Attributes: Service Sector Only**

| | (1) Policy Found | (2) Policy Visible | (3) Log Words | (4) Overall Score | (5) Fog Index | (6) $3^{rd}$-Party Trackers |
|---|---|---|---|---|---|---|
| Log Market Value | 0.0328*** | 0.0246*** | 0.0471** | 0.0996*** | 0.0886 | 0.512*** |
| | (4.32) | (2.59) | (2.12) | (4.72) | (1.53) | (3.98) |
| Log Market Share | 0.0178** | 0.0199** | 0.0133 | 0.0195 | 0.0220 | 0.125 |
| | (2.51) | (2.25) | (0.63) | (0.97) | (0.40) | (1.03) |
| Knowledge Share | 0.644** | 0.606* | 1.830** | 2.451*** | 1.962 | 11.90*** |
| | (2.54) | (1.92) | (2.55) | (3.58) | (1.04) | (2.76) |
| Knowledge Share$^2$ | -0.499 | -0.573 | -2.579* | -2.791** | -2.930 | -20.43** |
| | (-0.96) | (-0.88) | (-1.75) | (-1.99) | (-0.76) | (-2.28) |
| Observations | 730 | 730 | 617 | 617 | 617 | 707 |

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 5: **Policy Attributes: 2-digit SIC Code Fixed Effects**

| | (1) Policy Found | (2) Policy Visible | (3) Log Words | (4) Overall Score | (5) Fog Index | (6) $3^{rd}$-Party Trackers |
|---|---|---|---|---|---|---|
| Log Market Value | 0.0342*** | 0.0326*** | 0.0350** | 0.0409 | 0.0773* | 0.268*** |
| | (5.26) | (5.75) | (2.39) | (1.65) | (1.71) | (4.89) |
| Knowledge Share | 0.691*** | 0.588*** | 0.994*** | 1.823*** | -0.0590 | 4.691** |
| | (7.56) | (4.10) | (3.00) | (3.64) | (-0.06) | (2.26) |
| Knowledge Share$^2$ | -0.536*** | -0.607*** | -1.177*** | -2.521*** | 0.218 | -4.790* |
| | (-4.88) | (-3.24) | (-3.01) | (-4.49) | (0.18) | (-1.67) |
| Log Market Share | 0.0241*** | 0.00758 | 0.0390*** | 0.0658*** | 0.0358 | 0.207*** |
| | (4.20) | (1.03) | (2.67) | (3.19) | (0.80) | (4.39) |
| Observations | 5140 | 5140 | 3918 | 3918 | 3918 | 4951 |

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the SIC2 industry level.

Table 6: **Policy Attributes: Extensive Margin of Third-Party Trackers**

| | (1) $3^{rd}$-Party Trackers > 0 | (2) $3^{rd}$-Party Trackers > 0 |
|---|---|---|
| Log Market Value | 0.0339** | 0.0364** |
| | (2.27) | (2.64) |
| Knowledge Share | 0.654*** | 0.644 |
| | (3.40) | (1.73) |
| Knowledge Share$^2$ | -1.060*** | -1.133* |
| | (-4.20) | (-2.06) |
| Log Market Share | 0.00833 | 0.00673 |
| | (0.70) | (0.65) |
| Sector FE | No | Yes |
| N | 5140 | 5140 |

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level.

# 6 Alternative Measure of Legal Clarity

We fit a penalized logistic regression model which predicts the Overall score (High, Medium or Low) that our expert attached to each policy in the reviewed subsample, as a function of the TF.IDF weights on the most common 5000 words in our corpus of documents. We split the reviewed sample into a training set containing 80% ($N = 341$) of the observations, and a test set containing the remaining 20% ($N = 86$). On the training set, we first use SMOTE oversampling (Chawla et al., 2002) in order to obtain a balanced sample of High-, Medium- and Low-scoring policies. On this training set, we estimate a three-class logistic model, with an $L_2$ regularization penalty that is proportional to the squared norm of the parameter vector.[2] Based on a cross-validation exercise, we choose the weight on the regularization penalty to be equal to $C = 1$. Then, we check the performance of the classifier on the test set, and extrapolate its predictions to all privacy policies that our expert did not evaluate. In addition, we use the same methodology to predict a dummy variable provided by the expert to indicate policies that do not contain privacy-relevant text. Figure 10 shows the performance of both classifiers on the training and test data, in terms of a "confusion matrix" that compares true scores (in the rows of the matrix) with predicted scores (in the columns).

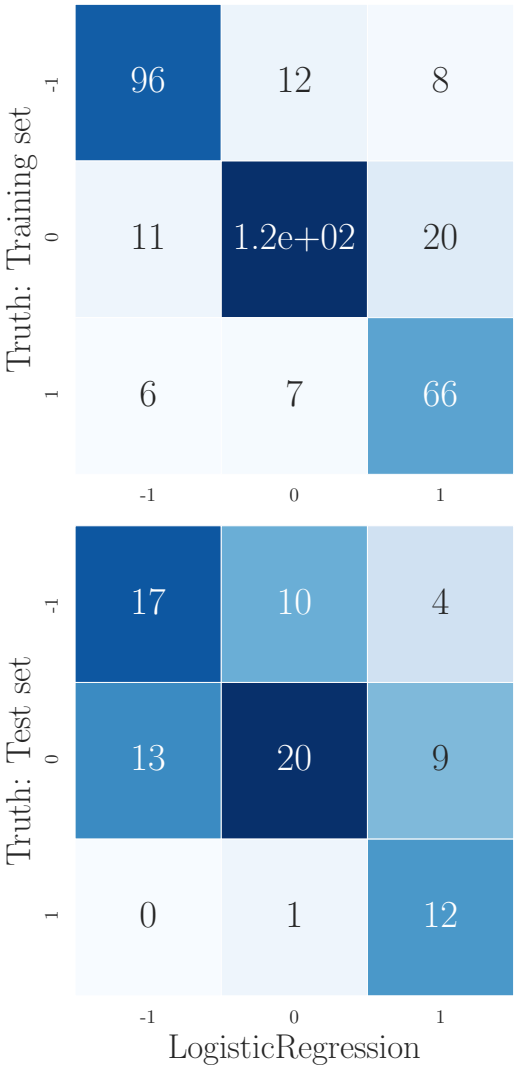
Out of sample, our classifier is quite good at distinguishing High scores form Medium and Low ones, but not very good at distinguishing Medium and Low scores from each other. Hence, we use the predicted probability that the score is High as our alternative measure of clarity.

Table 7 shows a regression of these alternative measures on our baseline set of firm characteristics. We consider two cases: In the first column, we include all policies in our sample. In the second column, we exclude policies with a probability greater than 50% of not containing relevant text.

[2] We use the *sklearn* package in Python to fit this model. For further details, see: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

Figure 10: **Performance of Classifiers**

(a) Predicting Overall Score




(b) Predicting "irrelevant" dummy

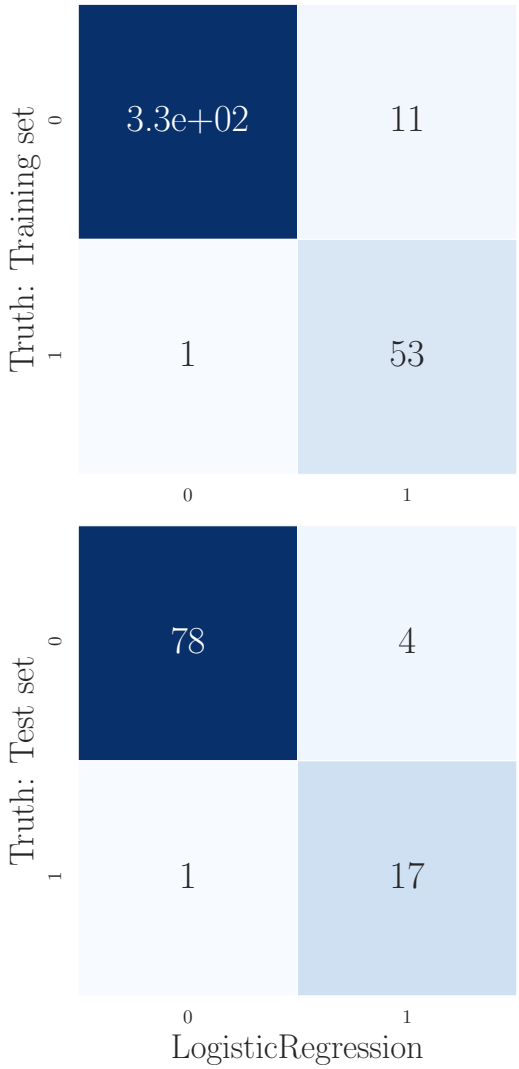


Table 7: **Policy Attributes: Alternative Legal Clarity Measure**

(a) Without Sector Fixed Effects

| | (1) Pr[High Score] | (2) Pr[High Score] |
|---|---|---|
| Log Market Value | 0.00488 | 0.00512 |
| | (0.92) | (0.96) |
| Knowledge Share | 0.829*** | 0.767*** |
| | (4.38) | (3.48) |
| Knowledge Share$^2$ | -1.060*** | -1.007*** |
| | (-4.23) | (-3.53) |
| Log Market Share | 0.0123** | 0.0120* |
| | (2.31) | (1.99) |
| Filtered | No | Yes |
| N | 3918 | 3099 |

(b) With Sector Fixed Effects

| | (1) Pr[High Score] | (2) Pr[High Score] |
|---|---|---|
| Log Market Value | 0.00762 | 0.00813 |
| | (1.53) | (1.75) |
| Knowledge Share | 0.521*** | 0.495** |
| | (3.94) | (3.11) |
| Knowledge Share$^2$ | -0.707*** | -0.687*** |
| | (-4.38) | (-3.51) |
| Log Market Share | 0.00858 | 0.00859 |
| | (1.77) | (1.75) |
| Filtered | No | Yes |
| N | 3918 | 3099 |

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b).

# 7 Additional theoretical analysis

## 7.1 The sources of data valuations

In the paper, we impose

$$V(y) \geq V(x) > 0 \tag{1}$$

This assumption means that *(i)* the firm's data is valuable, and *(ii)* it becomes even more valuable when processed by an intermediary. Recall that both $x$ and $y$ are signals about some underlying state of the world $\theta$ (which can also be a vector containing many pieces of fundamental information such as the characteristics of demand for each consumer in some database). In this appendix, we consider two possible underlying explanations for these assumptions. One is based on the contribution of data to the informativeness of decisions, and the second is based more on concepts related to statistical learning.

**Bayesian decisions.** Suppose that the firm has the opportunity to sell its signal to a population of information buyers $i \in \{1, ..., I\}$, whose payoff is $u_i(a_i, \theta)$, where $a_i$ is a vector of actions that buyer $i$ needs to choose. In addition, the firm's own profits are $u_0(a_i, \theta)$. Hence, both the firm's profits and the payoffs of information buyers are sensitive to information about $\theta$. This can capture, for example, a setting where the firm uses information about its consumers to target costly advertisements, and where the information buyers are firms in other industries that can also use this information for targeted advertisements, or indeed, for price discrimination.

Generically, the value that an agent with utility $u_i(a_i, \theta)$ derives from owning signal $x$ is

$$v_i(x) = E\left[\max_{a_i} E\left[u_i(a_i, \theta)|x\right]\right] - E\left[\max_{a} E\left[u_i(a_i, \theta)\right]\right]$$

This expression measures the difference in her expected maximized value if she makes decisions having observed $x$, versus its equivalent if she makes decisions based only on prior beliefs about $\theta$.

Assume, for simplicity, that the firm can make take-it-or-leave-it offers to individual information buyers (i.e., the firm engages in first-order price discrimination). Then it will charge each buyer $i$ a price $p_i = v_i(x)$ that equals their willingness to pay. The total value that the firm can derive from owning signal $x$ is therefore

$$V(x) = v_0(x) + \sum_{i \in I} v_i(x)$$

The first term is the value that the firm extracts from the data when using these data it for its own decisions. The second captures the value it can extract by monetizing the data and selling the resulting signal to information buyers. This equation can

easily be adjusted if the firm cannot perfectly price-discriminate against information buyers.[3]

We can now state conditions under which our assumption (1) is satisfied in this more concrete model. By standard arguments in Bayesian decision theory, we know that $V(x) > 0$ whenever there are realizations of the signal $x$ which occur with positive probability, and which induce a decision-maker (either the firm or one of the information buyers) to take a different action $a_i$ from the "default" action she would take based on prior beliefs. Moreover, we have $V(y) \geq V(x)$ if the signal $y$ produced by the data intermediary is more informative than $x$ about $\theta$, in the sense of Blackwell (1953).[4]

**Statistical learning.** Consider the case of Bayesian decisions, as above, but now specify that the utility of each agent is quadratic. For a concrete example, suppose that there is a single state variable $\theta$ (this easily extends to the vector case with additional notation), and suppose that each agent (i.e., the firm itself or an information buyer) needs to take a single action $a_i$, with associated utility

$$u_i(a_i, \theta) = -(a_i - \theta)^2$$

We can re-interpret a signal $x$ in this context as delivering a prediction $\hat{\theta}(x)$ of the state, which is obtained by fitting a statistical model (e.g., a linear regression or a machine learning model) to the data that the firm collects. It is well known that the value $v_i(x)$ of a signal $x$ in this context is proportional to the negative *mean-square error* of this prediction:

$$v_i(x) = -E\left[(\hat{\theta}(x) - \theta)^2\right] \equiv -MSE(x)$$

The goal of supervised machine learning methods is precisely to obtain predictions that minimize $MSE(x)$. This objective is known as the "test error". It measures the *population* average of a squared error loss, i.e. the loss that an agent would experience on average if predicting $\theta$ on "test" data that were not used in estimation. Typically, modern methods operate by minimizing the sum of the average loss on a "training" dataset, and a penalty term that is designed to prevent overfitting.

In this context, our assumption (1) can be motivated by the relative sophistication of statistical technology. Indeed, it is satisfied when the intermediary produces a signal $y$ with $MSE(y) \leq MSE(x)$. This can arise because the intermediary has access to a more sophisticated machine learning technology or, alternatively, to a pool of workers that can operate the available technology more effectively.

[3] If we call $\alpha_i$ the maximal share of buyer $i$'s valuation that the firm can extract, then we can replace the second term above by $\sum_{i \in I} \alpha_i v_i(x)$.

[4] The Blackwell ranking says that $s$ is more informative about $\theta$ than $s'$ if every Bayesian decision maker whose objective depends on $\theta$ would prefer observing $s$ to observing $s'$. Blackwell's theorem states that this ranking is equivalent to being able to express $s'$ as a "garbling" of $s$.

## 7.2 The industrial organization of the data market

In our model in the paper, only the firm collects a dataset that can be monetized. Hence, the intermediary is only an active player in the data market if the firm decides to share its data.

We now consider a richer setting, where the intermediary has a dataset of its own. If the firm shares its data with the intermediary, it can produce a signal $y$ as before. If the firm does not share its data, the firm can produce a signal $x$ at a cost $\phi$ as before, but the intermediary can also use its own data to produce a signal $z$.

We let $V(x|z)$ denote the monetary value of owning a signal $x$ when agents already have access to signal $z$. We assume that

$$V(y) \geq V(z) + V(x|z)$$

Therefore, the value of the signal that can be produced with data sharing is strictly greater than the total joint valuation of the signals that the firm and intermediary can produce on their own. This condition is satisfied, for example, in the setting with Bayesian decisions that we considered above, as long as $y$ is more informative than the combined signal $\{x, z\}$ in the sense of Blackwell (1953).

For concreteness, we assume that if the firm does not share its data, the intermediary acts as a Stackelberg leader: The intermediary extracts value $V(z)$ from its signal (e.g., by making take-it-or-leave-it offers to information buyers, as discussed above). Then, the firm extracts value $V(x|z)$ (e.g., by offering $x$ for sale to buyers who have already purchased $z$).

The remainder of the model is the same as in the body of the paper. In this setting, all our results go through, modulo a change in the definition of the intermediary's efficiency advantage. This advantage is now equal to

$$C = V(y) - [V(z) + V(x|z)] + \phi.$$

One can imagine alternative settings, such as differentiated Bertrand competition in a data market with many competing intermediaries and firms. However, the qualitative conclusions are likely to remain similar, so long as the total value of the signals that can be monetized increases when the firm shares its data with the intermediary.

# References

Blackwell, D. (1953): "Equivalent Comparison of Experiments," *Annals of Mathematical Statistics*, 24, 265–272.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002): "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, 16, 321–357.