

Privacy Policies and Consumer Data Extraction: Evidence From U.S. Firms

Tarun Ramadorai, Antoine Uettwiller, Ansgar Walther*

February 20, 2025

Abstract

Using a comprehensive dataset of privacy policies, firm characteristics, consumer tracking, and cybersecurity incidents, we document several stylized facts about the heterogeneity of firms' data extraction practices and the influence of privacy regulations. Rather than adopting standardized boilerplate privacy policies, we find substantial within-industry differences correlated with firms' technical sophistication; firms engaging in data extraction have lengthier policies, seeking to hedge legal risks. Firms with intermediate technical sophistication appear to follow a "collect and share" model, collecting large amounts of consumer data and sharing it with third-parties for processing, thus creating cybersecurity risks. Conversely, high sophistication firms appear to implement a "receive and process" model, consistent with a two-tier data market in which data flows from intermediate to high sophistication firms.

*Ramadorai: Imperial College London and CEPR. Email: t.ramadorai@imperial.ac.uk. Uettwiller: Queen Mary University of London. Email: a.uettwiller@qmul.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, the NBER IT and Digitization Summer Institute, the Cambridge AI/ML in Finance Lecture, the University of East Anglia, and Alex Edmans, Xavier Giroud, Avi Goldfarb, Stephen Hansen, Roxana Mihet, Emiliano Pagnotta, David Thesmar, Tommaso Valletti, Andre Veiga, and anonymous referees for comments.

1. Introduction

There is growing public interest in consumers' data privacy and firms' ability to track consumer behavior and discover personal information. In this context, academic research on data privacy has discovered interesting facts about consumer demand and willingness to pay for privacy (e.g., [Tucker, 2012](#); [Loewenstein, 2015](#); [Tang, 2019](#); [Chen et al., 2021](#)). While the increasing reliance on artificial intelligence and data analytics has brought transformative changes from strategic insights to operational efficiencies, it has also heightened privacy and cybersecurity concerns for firms¹.

We analyze how firms navigate this dual objective to harness consumer data through collection, extraction and sharing, while taking into account cybersecurity and other risks. A long literature in financial economics, starting with the seminal work of [Pfleiderer \(1986\)](#), emphasizes that firms who possess private information have strong incentives to trade and monetize their data. In addition, recent work emphasizes the unique industrial organization of data markets: large platforms and data intermediaries can use their central position to acquire data cheaply from its owners, and to extract a large share of the associated surplus when selling it on to firms (e.g., [Bonatti, 2019](#); [Acemoglu et al., 2022](#)). The usual forces of competition, in this context, are muted ([Ichihashi, 2021](#)). In short, a strong theoretical prior suggests that markets for data gravitate towards a two-tier model: Sophisticated firms/intermediaries receive and process data, while other agents tend to collect data at source and then share it with the intermediaries.

In the absence of measurement of data flows between firms, an empirical analysis of these effects is challenging. In order to make progress, we collect a comprehensive dataset tracking more than 5,000 US firms' privacy policies, and their communications with consumers, from 2016 to 2023. We merge this data with detailed information on how firms

¹In a recent report, [of the Treasury \(2024\)](#), the U.S. Treasury highlighted artificial intelligence-related cybersecurity and fraud risks in the financial services industry, putting a particular emphasis on third-party risks.

track consumers using cookies on their websites. We then analyze how these tracking and communication measures relate to firm characteristics and mandatory disclosures including their 10-K risk factors, and estimates of cyberbreach costs.

Our analysis proceeds in two parts. First, we highlight five stylized facts, described in detail below, which validate the various sources of our data and highlight some basic structures in the market for consumer data. Second, we argue that the stylized facts, in conjunction with a further analysis of firms' disclosures about data use, are consistent with the two-tier model of the market for data. This part of our work reveals novel patterns consistent with a market in which firms with intermediate levels of technological sophistication collect and share data with more technically advanced firms. We also find that the firms that we identify as "collect and share" appear to bear greater cybersecurity risks than their more technically advanced "receive and process" counterparts. Since investors demand higher risk-premia from cyber-exposed firms (Florackis et al., 2023), this heterogeneity is a potentially important factor, given that these risks are only expected to increase.² We view these new findings as a useful first step in the analysis of the supply of privacy and, more broadly, the way in which firms interact with consumer data.

Our first stylized fact is that there is considerable variation in the text of privacy policies, and most of this variation is within, rather than between, industries. This is a simple validation exercise which suggests that there is information (as opposed to pure boilerplate) in firms' communication with their consumers. This finding motivates our deeper analysis of what drives firms' choices, but it is an initial demonstration that firms tailor their privacy policies to their individual business models. Concretely, the median cosine similarity between individual policies and the sample centroid is 0.61, which translates to a 52-degree median angle between policy word-frequency vectors and that of the average policy across all firms in our sample. This measure of similarity does not change

²The SEC released [new rules](#) on July 26, 2023, heightening cybersecurity disclosure requirements by firms within both the 8-Ks and 10-Ks. The director of the Division of Corporation Finance at the SEC, Erik Gerding, mentioned in a [recent statement](#) the "prevalence of cybersecurity incidents" as a reason for those.

substantially if, instead of comparing policies to the grand sample average, we compare them to averages within increasingly fine industry buckets.³

Our second stylized fact is that, by comparing privacy policies with tracking activity, we find evidence consistent with these communications being used by firms to hedge legal risk, rather than to protect consumers. Firms that actively extract consumer data using tracking cookies write more elaborate policies containing clauses to mitigate the legal risks associated with extracting consumer data. This finding, which is consistent with firms behaving strategically about their data extraction choices in a market populated by inattentive consumers, speaks directly to the privacy literature, which increasingly emphasizes potential behavioral biases that may affect consumers' privacy-related choices (e.g., [Loewenstein, 2013](#); [Xiong, 2020](#)).⁴

Third, we observe that extraction activities, which we measure in two ways—through the use of cookies, as well as the consumer data that firms describe that they collect in their privacy policies—are linked to increased cybersecurity risks. This holds true despite mitigation efforts by firms that engage in greater data collection, such as crafting careful privacy policies and investing in cybersecurity measures. These risks are evident both in terms of perceived or disclosed risks ex-ante—such firms mention privacy risks more frequently in their mandatory risk disclosures, as well as realized risks ex-post—such firms exhibit a higher likelihood of cyberbreaches.

Fourth, we relate firms' privacy policies and data collection practices to their economic

³The median cosine similarity with the 3-digit SIC-level centroid is around 0.66, corresponding to a 49-degree median angle between firms' policy word frequency vectors and that of the industry-average vector. We find similar results when using topic frequencies from a simple language model. This suggests that the variation within industries is not driven by spurious semantic differences, such as two firms choosing different words to describe the same practice.

⁴Indeed, many consumers can be reassured by the mere presence of legal text such as privacy policies or other generic information, even if such information is unrelated to firms' actual data extraction practices ([Adjerid et al., 2013](#); [Urban, 2014](#); [Tucker, 2017](#)). In the 2009 *Berkeley Privacy Survey*, 62% of respondents believed that "If a website has a privacy policy, it means that the site cannot share information about you with other companies, unless you give the website your permission." ([Urban, 2014](#), p.309). [Tucker \(2017\)](#) show that, in an experiment, the introduction of generic information about privacy protection, which was irrelevant to the actual extraction policy, made participants less likely to avoid surveillance. Relatedly, [Adjerid et al. \(2013\)](#) show experimentally that small, irrelevant distractions reduce consumers' sensitivity to the policy

characteristics. We find strong size effects: the largest firms conduct the most data extraction. We also find an interesting non-monotonic relationship between privacy policy attributes and firms' technical sophistication, which we measure using the [Taylor \(2017\)](#) measure of knowledge capital, expressed as a share of firms' total capital: firms with intermediate technical sophistication are most extractive, measured by their use of cookies. Once again, these firms who extract more consumer data (i.e., large firms and firms with intermediate knowledge capital shares) have more visible and longer privacy policies. We later show that such firms are also more concerned with privacy risks and have larger costs associated with cybersecurity incidents. This is interesting in the context of the finance literature that has established a clear link between cybersecurity and financial performance ([Florackis et al., 2023](#)).

Finally, we study the effects of these patterns of a recent spate of privacy regulations, in particular, the EU General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). We find that these regulatory actions have had significant effects on firm behavior—as regulation becomes binding over time, we show that privacy policies become more visible, more detailed, and less shrouded/more truthful.

We argue on the basis of these findings that firms can be classified into three principal groups, two of which are particularly important to study as data becomes an ever-increasing source of competitive advantage. The first group comprises firms that are not very active in data extraction. These are typically small firms and those with low technical sophistication as measured by Peters-Taylor knowledge share. These firms use fewer cookies, are least likely to have visible/elaborate privacy policies, and have low cyber risk associated with data and privacy. The second type of firm appears to engage in “collect and share” behavior, which is especially prevalent among firms with intermediate technical sophistication. Such firms extract data via cookies, monetise data by sharing it with third parties, hedge their legal risks with more elaborate privacy policies, and have higher realized cyber-risks, which are apparently not fully offset by these firms' greater

investments in protection. The third group of firms “receive and process” data, and are more often seen in the set of firms with high technical sophistication. While they do collect some data directly, they have fewer tracking cookies than “collect and share” firms, process data in-house to monetise, have less sophisticated privacy policies, and exhibit intermediate levels of cyber risk.

The existence of these groups hints at a two-tier market for data, in which data flows from medium to high technical sophistication firms but financial risks are borne by medium sophistication firms. While the flows of data are difficult to measure directly, the distinction between firms who share and firms who receive data in our textual analysis of regulatory risk disclosures is striking and consistent with this explanation.

The remainder of the paper is organized as follows. Section 2 describes our data on firm characteristics, privacy policies, web tracking behavior, and provides basic descriptive statistics. Section 3 establishes stylized facts showing that policies are not boilerplate (Section 3.1), are a hedge against risk coming from extracting consumer data (Sections 3.2 and 3.3), which is mainly practised by large and medium sophistication firms (Section 3.4). In Section 3.5, we focus on the times series and demonstrate that privacy regulations have influenced both privacy policies and tracking practices. In Section 4, we interpret our results differentiating between two business models of consumer data: “collect and share” and “receive and process”. Finally, Section 5 concludes.

2. Data

2.1 Privacy Policies and Cookies

We compile a broad cross-section and time series of privacy policies. To create this dataset, we followed a multi-step process. First, we used automated Google searches and web crawling techniques to find firms’ privacy policies, focusing on each firm’s main web domain as listed in Compustat. We further supplemented this with a web crawl of the

firm’s domain, and in cases where policies were not automatically found, we conducted manual checks. We scraped the text of each policy, discarding any that did not include the word “privacy.” Ultimately, we collected policies for 6,019 firms⁵, which forms our primary sample. We extended this cross-section into a time series using the Wayback Machine, and added additional metrics, as outlined below. Recognizing that privacy links may change over time, we minimized survivorship bias by collecting those links at a midpoint, in 2019.

2.1.1 Privacy Policies over time

Having gathered the privacy links and homepages, we used the Wayback Machine⁶ to get snapshots of privacy policies and homepages between 2016 and mid-2023. This allows us to track changes in privacy policies as well as the visibility of those privacy policies over time. Figure A.1a in the Appendix shows two histograms of the number of privacy policies we gathered per month and year. On average, we get snapshots for about 3,000 policies per year.

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 prevalent English words from a standard list of “stop words” that convey little semantic meaning (e.g., “is”, “in”, “and”, “or”).⁷

Figure 1a visualizes the textual content of our sample in terms of the most important bigrams (pairs of consecutive words), keeping one policy per firm per year. 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

⁵These policies were obtained from 5,260 unique web domains, with some domains being shared by multiple firms.

⁶Available at web.archive.org

⁷We detect non-English with the pyenchant spellchecker (see github.com/pyenchant/pyenchant), and named entities with the Stanford NER tool (see [Manning \(2005\)](https://nlp.stanford.edu/software/CRF-NER.shtml) and nlp.stanford.edu/software/CRF-NER.shtml). We use the NLTK list of English stopwords (see nltk.org).

a large fraction of documents.⁸ As might be expected, the policies prominently feature the word “privacy.” They also prominently feature the terms “personal information,” “personal data,” and “identifiable information.” Finally, an important and frequently used term that is evident is “third party,” which we will return to discussing later in the paper.

⁸We 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 the document (in our case, policy) i . The TF-IDF metric is:

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

where N is the total number of documents. See [Ullman \(2011\)](#), and [Taddy \(2019\)](#) for more detailed treatments of TF-IDF.

Figure 1b shows the histograms of the average privacy policy length for the domains in our sample⁹. When found, the average privacy policy contains 2,800 words. The distribution of length, which the figure shows on a log scale, is skewed to the right, and a large number of policies feature between 1,000 and 10,000 words. Figure 1c plots the Gunning (1952) Fog index of “readability”, which is based on the sentence-level frequency of complex and polysyllabic words.¹⁰ The Fog index heuristically measures the number of years of formal education required to understand a document at first reading. The distribution of the Fog index suggests that the average policy requires the reading level of a high school sophomore to be understood on first reading (gunning fog of 10). with a substantial number of policies requiring the level of a college freshman (gunning fog of 13).

We should expect that firms disclose the information they gather on their customers in their privacy policies. This is clearly stated in laws that came to pass during our sample in various jurisdictions. For example, GDPR which took effect on May 25th, 2018, states that “[i]t should be transparent to natural persons that personal data concerning them are collected, used, consulted or otherwise processed”. CCPA, which took effect on January 1st, 2020, states - perhaps more clearly - that “a notice at collection must list the categories of personal information businesses collect about consumers and the purposes for which they use the categories of information”. Extracting this information from policies is not straightforward and has been the subject of academic research (e.g. Bui et al., 2021). Those methods are conceivably outdated in light of recent advances in textual analysis via large language models. We use OpenAI and its GPT-4 language model to extract the data points firms collect. In doing so, we sent the privacy policies to the OpenAI API and asked the question “Can you give me the list of the personal information they collect?”. For

⁹For each domain, we take the average length and fog of the snapshots we gathered.

¹⁰These figures are similar to recent findings in the Computer Science literature (e.g., Lentz, 2017). 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.

example, for the July 29th, 2016 snapshot of the privacy policy of American Airlines¹¹, the returned answer was the following:

The personal information that American Airlines collects may include, but is not limited to: 1. Name 2. Date of birth 3. Address 4. Telephone number 5. Credit/debit card number(s) and associated billing address(es) and expiration date(s) 6. Travel companion(s) names 7. Emergency contacts 8. Photographs 9. Seating preferences 10. Special dietary or medical needs 11. Corporate contract, employer, and/or other corporate affiliation information (e.g., employer name, title, work address, and contact information) 12. Business Extra account information (e.g., tax ID, business type, number of employees, number of business travellers, and travel manager information). Please note that this is not an exhaustive list, and American Airlines may collect additional information as necessary to facilitate travel and manage their business

From those answers, we then extract the data points collected into different categories such as personally identifiable information, government IDs, education, interests, demographics, financial, physical, and medical information, risk tolerance, device information, and geo-localization, for a total of over 60 possible data points. Figure 1d plots the average number of data points collected per the privacy policies (one per year per firm). The number rose from about 4 in 2016 to over 6 in 2023. Figure 1e shows the most collected categories, with addresses, names and emails mentioned in over 60% of the policies. We also notice that cookies are mentioned in over 10% of policies.

2.1.2 Data Extraction Behaviour: Cookies

We obtain data on each firm's data extraction behaviour using the "OpenWPM" crawler developed in Narayanan (2016).¹² We emphasize that this measure is collected separately from, and has no direct relationship with the privacy policy data. There are null returns for which the crawler fails, which reduces the sample size. We ran the crawler twice, once in February 2019, and the second time in May 2023. Descriptive statistics of these tracking data are in Table 1.

¹¹Available through the [Wayback Machine](#).

¹²Their open-source privacy measurement software is available at github.com/mozilla/OpenWPM. We obtain our data by scraping the results for each firm on privacyscore.org, which uses OpenWPM.

Table 1: Cookies Summary Statistics

	2019				2023			
	Count	Mean	Median	Std.	Count	Mean	Median	Std.
First Party Cookies	8713.00	7.28	6.00	6.47	3072.00	6.86	4.00	8.33
Third Party Requests	8713.00	33.59	19.00	41.80	3072.00	40.50	27.00	44.48
Third Party Tracking Cookies	8713.00	3.03	1.00	5.42	3072.00	1.80	0.00	3.35
Total Cookies	8713.00	14.59	8.00	17.03	3072.00	13.93	7.00	17.08

Note: The table reports the number of first-party cookies placed by firms on their websites, the number of unique third parties placing tracking cookies, and the total number of third-party requests. These measures provide insight into firms' tracking practices, with total cookies serving as a comprehensive proxy for data extraction. Data extraction measures are obtained from privacyscore.org, which uses the OpenWPM web measurement method. For details on OpenWPM, see [Narayanan \(2016\)](#) and webtap.princeton.edu. We winsorize the extraction measures at their first and 99th percentile. The scan dates were February 28th, 2019, and May 10th, 2023.

The first row shows the number of “first-party” cookies each firm places on its website. The second row shows the number of unique third parties (i.e., agents other than the firm itself) who place cookies on the firm’s website classified as “tracking cookies.” For an alternative measure of third-party activity, the third row shows the *total* number of third-party requests encountered on the firm’s website, including, but not limited to cookies. The correlation between the two measures of third-party activity is 70%.¹³ For simplicity, we use total cookies, as measured by OpenWPM. This measure shows a high correlation with the number of first-party cookies (76%) and third-party requests (92%), as our overall measure of data extraction in what follows.

In the 2023 data, a large number of scans failed. Those scan failures happened for a variety of reasons. We perform a balance test to study those failures in Table 2. We find that firms for which the scans failed tend to be smaller and have slightly harder-to-read privacy policies.

¹³Using third-party tracking cookies instead of requests is more conservative because tracking is not always done via cookies. This approach is also quite accurate because not all cookies are tracking cookies.

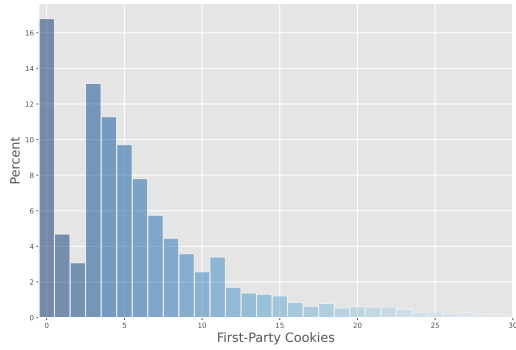
Table 2: Balance Tests

Variable	Scan Successful - Mean	Scan Failure - Mean	T-test
Log Market Value	19.70	19.53	-2.68
Knowledge Share	0.11	0.10	-0.75
10-K Security Risks	0.52	0.53	0.84
Policy Length	2668.25	2806.30	1.67
Policy Fog	9.59	9.76	2.98
Obs	3072	5899	

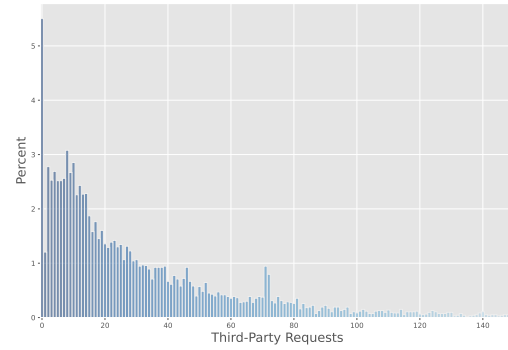
Note: This table compares firms with successful and failed scans across key attributes, such as firm size and privacy policy readability. The results indicate that firms with failed scans tend to be smaller and have slightly more complex privacy policies.

Figure 2 shows histograms of our data extraction measures. For all measures, the distribution of data extraction activity is skewed to the left, with many firms having zero or a small number of cookies/requests, and has a long right tail, with up to more than 100 third-party requests and up to around 60 in total.

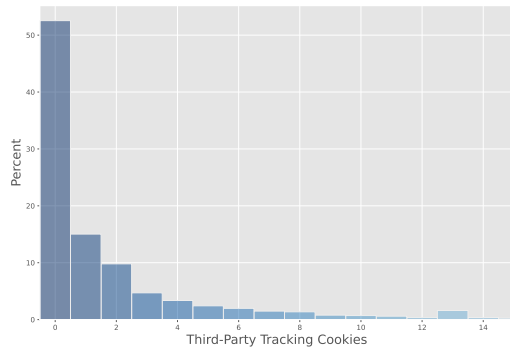
(a) First-Party Cookies



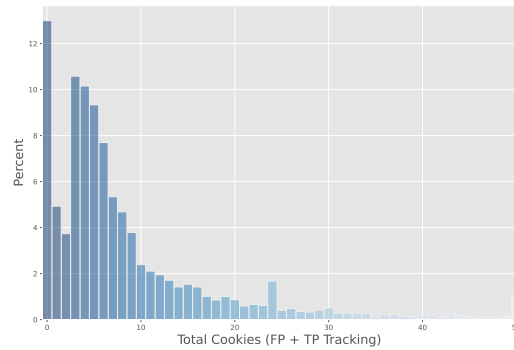
(b) Third-Party Requests



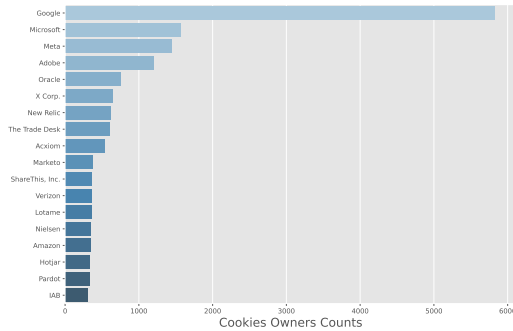
(c) Third-Party Tracking



(d) Total Cookies



(e) Cookies Ownership



(f) Cookies Type

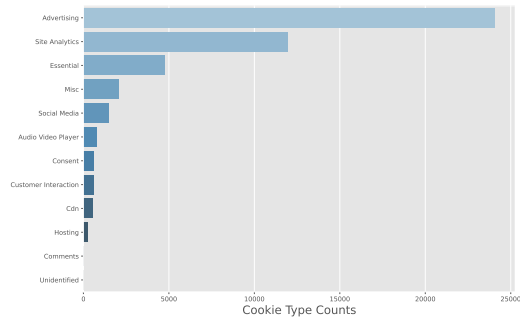


Figure 2: OpenWPM Data Collection Measures and Cookie Ownership

Note: Figures 2a-2d present histograms of data extraction measures, illustrating the distribution of cookies and third-party requests collected through OpenWPM. Figures 2e and 2f provide insights into cookie ownership and type. Data extraction measures are obtained from privacyscore.org, which uses the openWPM web measurement method. For details on OpenWPM, see [Narayanan \(2016\)](#) and webtap.princeton.edu. We winsorize the extraction measures at their first and 99th percentile. Data on cookie ownership and type come from [WhoTracks.me](https://whotracks.me).

Ultimately, cookies' owners also gather data on the consumers. As such, the number of cookies on homepages may not fully reflect the data any firm gathers on individuals. To alleviate those concerns, we use [WhoTracks.me](#) to show in Figure 2e that most of the cookies are owned by few firms, with Google being the most prominent by a large margin. Additionally, we want to make sure the cookies we have in our sample can be used to increase firms' revenues. [WhoTracks.me](#) allows us to do so and confirms, as shown in Figure 2f, that the vast majority of tracker requests fall in one of two categories: advertising or site analytics.

2.2 Firm Characteristics, Knowledge Capital, and Cybersecurity Risks

2.2.1 Compustat & Knowledge Capital

For all firms in our sample, 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 construct intangible capital following [Taylor \(2017\)](#)¹⁴. This is the sum of Compustat-recorded on-balance-sheet intangible capital and the replacement values of knowledge and organizational capital. [Taylor \(2017\)](#) estimate the replacement value of knowledge capital by accumulating past R&D expenditures for firms assuming an industry-specific depreciation rate, and the replacement value of organizational capital almost identically to [Papanikolaou \(2014\)](#), by accumulating a fraction of past SG&A expenditures.

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 assets¹⁵), and the ratio of market-

¹⁴Note that the [Taylor \(2017\)](#) dataset stops in 2022

¹⁵Total assets in Compustat are the sum of Current Assets - Total (ACT), Property, Plant and Equipment

ing and R&D expenditures to assets.

Table 3 shows descriptive statistics of the firm characteristics (average of firm-level characteristics between 2016 and 2023), following winsorization at the 1 and 99 percentile points to reduce the influence of outliers. On average, knowledge capital accounts for roughly 10% of total capital, while total intangible capital sits around 12%. 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.

Table 3: Descriptive Statistics of Firm Characteristics

	Count	Mean	Median	Std.
Market Value	5,870.00	5,912.72	476.30	18,235.30
Market To Book	6,016.00	2.83	1.22	5.74
Market Share	6,016.00	0.00	0.00	0.00
Intangible Share	4,262.00	0.12	0.04	0.17
Knowledge Share	4,288.00	0.10	0.00	0.18
R&D To Assets	2,261.00	0.16	0.04	0.30
Marketing To Assets	6,016.00	0.01	0.00	0.02

Note: This table presents descriptive statistics on firm characteristics, averaged at the firm level between 2016 and 2023. 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. 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 [Taylor \(2017\)](#)). We winsorize all variables at their first and 99th percentile.

2.2.2 Annual Reports (10-K): Data Intensive Firms and Privacy Risks

Ideally, we need to be able to distinguish the sub-sample of firms that have consumers (B2C) and generate consumer data. Regardless of utilization, having consumer data is a risk given the increase in leaks, cyberattacks, regulations, and penalties. As a result, firms with consumer data should have identified this as a risk factor and included it in their 10-K disclosures, as noted by [Florackis et al. \(2023\)](#). Furthermore, we should expect

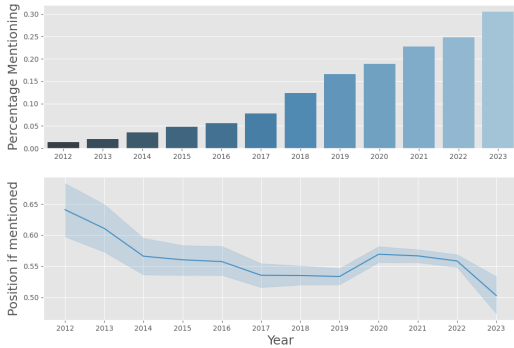
(Net) - Total (PPENT), Investment & Advances - Equity (IVAEQ), Investment & Advances - Other (IVAO), Intangible Assets - Total (INTAN), and Assets - Other - Total (AO).

the relative placement of the paragraph that mentions those risks - if it exists - to have decreased over the years¹⁶. Figure 3a shows that this is the case.

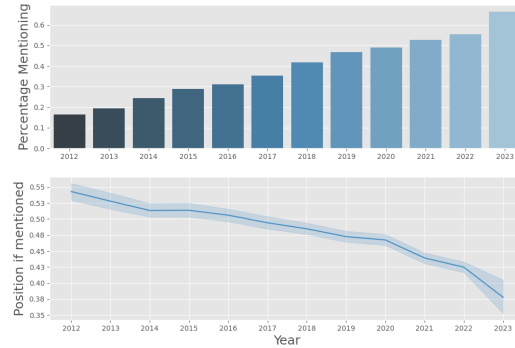
Similarly, we gather the positional risk of "privacy", as a measure of perceived privacy risk from the firms' perspective. Figure 3b shows that the proportion of firms mentioning privacy as a source of risk has increased over the years, with over 60% of firms mentioning it in 2023. Most notably, the relative importance of privacy as a risk factor has increased, being mentioned on average in the top 40% of the text.

¹⁶According to the SEC: "Item 1A - "Risk Factors" includes information about the most significant risks that apply to the company or its securities. Companies generally list the risk factors in order of their importance."

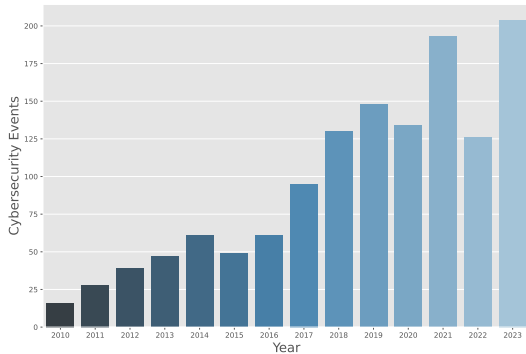
(a) Data Breaches 10-K Positional



(b) Privacy 10-K Positional



(c) Cybersecurity Events



(d) Cost of Breaches

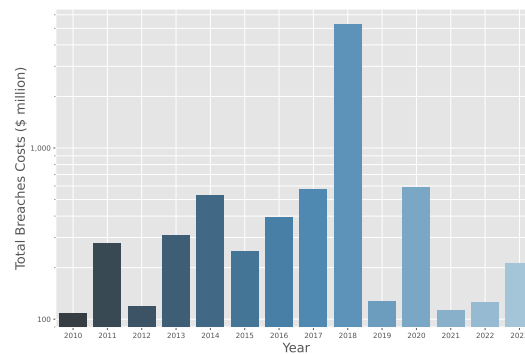


Figure 3: Cybersecurity Risks

Note: Figure 3a tracks the relative placement of data breach risk mentions in annual reports, showing a declining trend in their positioning, suggesting increasing importance. Figure 3b shows that the proportion of firms mentioning privacy risks in their disclosures has increased over time, with privacy risks appearing as of 2023, on average, in the top 40% of risk factors. Figures 3c and 3d demonstrate a significant increase in both the number of reported cybersecurity incidents and the total reported costs of data breaches between 2010 and 2023.

2.2.3 Cybersecurity Incidents, Fines, and Remediation Costs

Finally, some firms in our sample have had cybersecurity incidents, ranging from unauthorized access to malware and ransomware. Those incidents come with costs such as investigations, legal fees, and remediation. We use the Audit Analytics Cybersecurity database which contains a list of data breaches that have affected companies registered with the SEC. Figure 3c plots the number of cybersecurity incidents per year; the number

is quite clearly increasing with only 16 incidents reported in 2010 and over 200 in 2023. Figure 3d plots the total breach costs per year, which is only reported for a subsample of incidents. The median cost per year between 2010 and 2023 is \$1,500,000 with the largest fine imposed on Facebook (now Meta) by the FTC for the September 2018 data breach incident: over \$5B. We winsorize the total breach costs at the 99 percentile.

3. Stylized Facts

3.1 Variation in Privacy Policy Text

This section presents our first set of stylized facts, which is based on a decomposition of the variation in firms’ privacy policy text. We use the common *cosine similarity* measure, which quantifies the distance between the content of different text documents, to compare text documents across and within industries.

Concretely, the text of each firm’s privacy policy can be described as a vector $P_i = (P_{i1}, \dots, P_{iM})$ of word frequencies, where P_{im} is the frequency of word m in policy i , and M is the total number of words used in our sample of privacy policies. 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 $C = 1$.¹⁷ To measure aggregate variation, we compute the cosine similarity between each policy vector P_i and the centroid

¹⁷Note, 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 when moving from zero towards 90 degrees, the similarity measure is therefore more forgiving of small discrepancies between policies.

vector of all privacy policies in the sample, i.e., the “average” policy $\bar{P} = (\sum_j P_j) / N$. To isolate variation within industries, we compute the similarity between each policy and the associated industry-level centroid $\bar{P}_I = (\sum_{j \in I} P_j) / (\#I)$, where I is the set of firms in an industry. Figure 4a shows the cumulative distributions of cosine similarities with the sample-year centroid, the centroid associated with each firm’s SIC sector for the same year, with the centroids associated with each firm’s 2- and 3-digit SIC code bucket, once again for the same year, as well as a the Hoberg-Phillips centroid¹⁸. The median cosine similarity between individual policies and the sample centroid is 0.61, which translates to a 52-degree median angle between policy vectors and the grand average policy. This measure rises to a median cosine similarity of about 0.66, corresponding to a 49-degree median angle when we compare firms’ policies and the industry-average vector at the 3-digit SIC level. Figure 4b shows the associated mean cosine similarities, with 95-percent (bootstrapped) confidence intervals.

¹⁸As created by Phillips (2010). Data available at hobergphillips.tuck.dartmouth.edu.

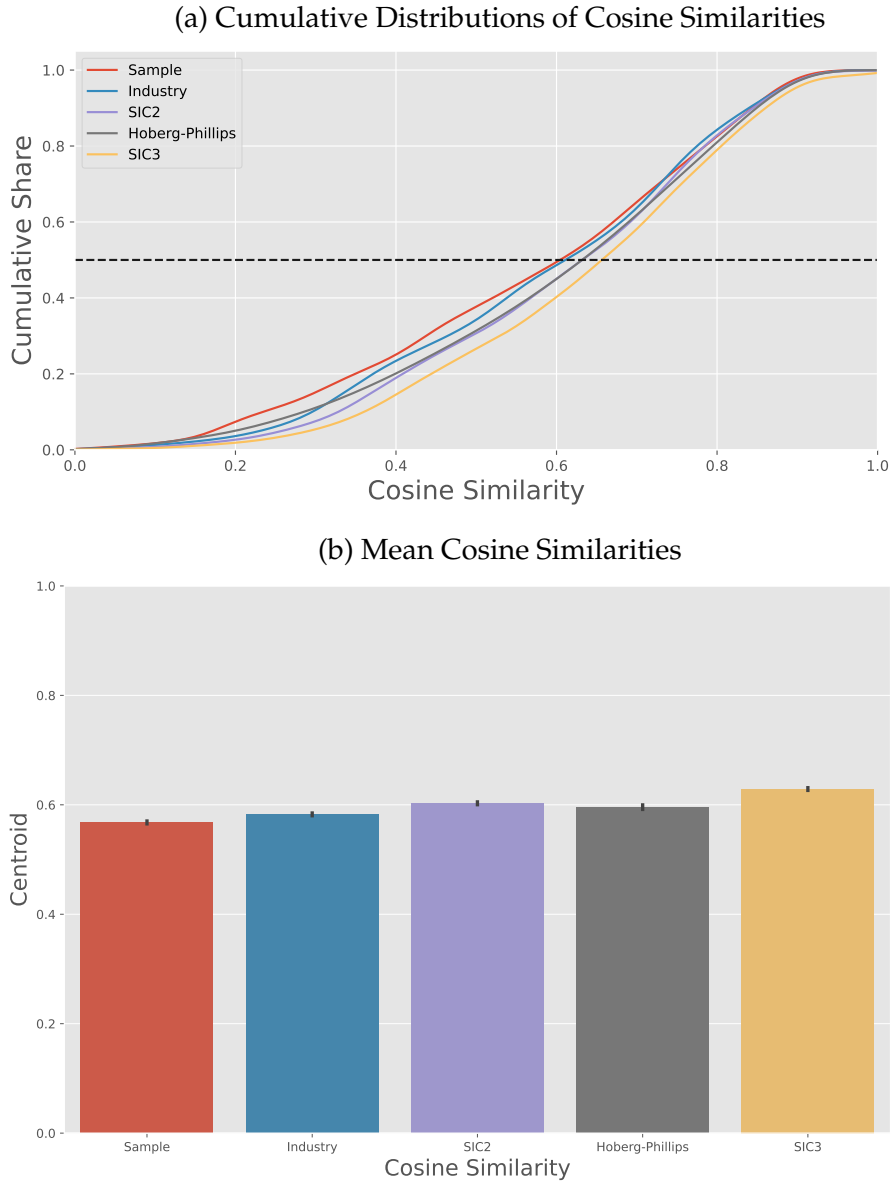


Figure 4: Cosine Similarities of Privacy Policies

Note: The Sample centroid is the mean TF.IDF frequency vector across our privacy policies, keeping one policy per year per firm. Industry 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 Nonclassifiable (9900,9999). SIC2 and SIC3 centroids are mean frequencies at the 2-digit and 3-digit SIC code levels, respectively. Hoberg-Phillips centroids are based on firms competitors as defined by Phillips (2010) (data available at hoberg-phillips.tuck.dartmouth.edu).

The figures show that *within-industry* variation is only marginally smaller than *total* variation in privacy policies. Indeed, only about 4 degrees of the median angle between policies and the average policy can be “explained” by accounting for the industry-specific frequency of words. The upshot of this analysis is that privacy policies are not “boilerplates” that respond to industry-level regulation (under a pure “industry boilerplate” hypothesis, the cosine similarity between each firm’s policy and the industry-level centroid would equal one). By contrast, we find that most of the variation in firms’ policies occurs within rather than between industries.

It is important to verify that these conclusions do not arise purely from idiosyncratic differences in wording. To check the robustness of our conclusions, we repeat the decomposition exercise above with a more specific measure of the semantic content of each policy in Figure A.2b in the Appendix. 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 10,000 words to 1,000 latent “topics”. We expect this transformation to eliminate low-level differences in wording, and to focus on semantic content. After the transformation, we find that the level of all cosine similarities naturally rises,¹⁹ but it remains the case that only a small fraction of the existing variation can be explained by industry effects.

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. We now investigate whether data extraction practices and firm characteristics are systematically associated with privacy policy content.

¹⁹The median cosine similarity relative to the sample centroid is about 0.63 in latent topic space, while the median similarity to the SIC-3 industry centroid is about 0.66.

3.2 Data Extraction Practices and Privacy Policies

In this section, we present our second set of stylized facts, which focuses on the association between firms' data extraction practices, as measured by total cookies in our data, and the content of their privacy policy text. Recall that our measure of cookies is obtained separately from privacy policy text, using the openWPM crawler of [Narayanan \(2016\)](#).

Figure 5 presents a simple, non-parametric visualization of the relationship between data extraction and privacy policies. In each panel, we sort firms into terciles of data extraction intensity along the horizontal axis and plot the associated means of privacy policy attributes on the vertical axis. The error bars denote (bootstrapped) 95-percent confidence intervals.

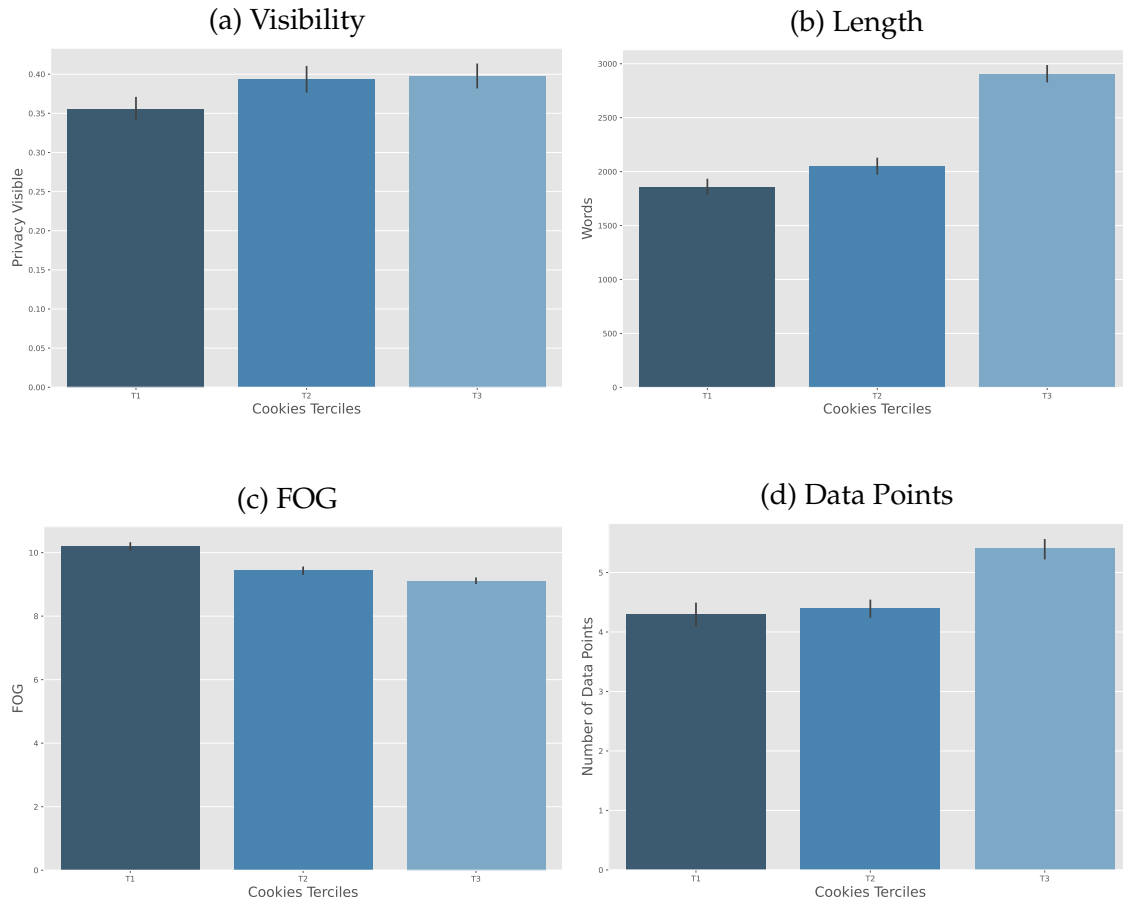


Figure 5: Average Attributes of Privacy Policies by Terciles of Total Cookies

Note: These figures present the average policy visibility, length, fog and number of data points collected for firms sorted by terciles of the total number of cookies they collect. For each firm in our sample, we keep the home page (Figure 5a) or privacy policy (Figures 5b, 5c and 5d) closest to the two scan dates performed through privacyscore.org: February 28th, 2019 and May 10th, 2023.

Figure 5a demonstrates that firms that use more cookies for tracking tend to make their privacy policies more visible by placing a link on their homepage. Figure 5b shows that conditional on having a privacy policy, firms in the top tercile of data extraction use significantly longer. Figure 5c shows that there is a significant association between data extraction and the Gunning Fog index of readability; firms that extract more data from their consumers - as also corroborated by Figure 5d, which shows that firms with more cookies also gather more data points from their consumers and have policies that are

easier to read, perhaps as mandated per regulation.

These facts, along with the substantial cognitive costs of reading privacy policies (e.g., [Cranor, 2008](#)), suggest that there is a higher cognitive cost to consumers of reading the policies of firms in the top tercile of data extraction. Figure [A.3](#) in the Appendix shows that the same patterns continue to arise if we sort firms by data extraction intensity after removing industry-year fixed effects, which is important, given the possibility of cross-industry variation in the importance of a web presence for sales and marketing purposes. Note however that the monotonic pattern on the Fog index is no longer. The Fog result in Figure [A.3](#) suggests that the “intelligible” aspect of privacy policies, as required by regulators, is more commonly implemented by firms that engage in extensive tracking, while companies that engage in minimal tracking have little reason to make their privacy policies opaque.

Another pattern in the data suggests that our results are not driven by the mechanical effect of firms whose business models are orthogonal to data collection (e.g., soup manufacturers): Figure [A.4](#) in the Appendix shows an association between data extraction practices and privacy policies holds even when we condition on firms that are relatively active data extractors, those identifying data as a source of risk in their 10-K filings. Note that the confidence intervals increase in size because this sample restriction reduces the sample size.

Our preferred interpretation of these patterns is twofold. First, it is in line with the recent evidence of consumer behaviour, which suggests that some consumers are easily reassured by visible privacy policies, even if those policies do not contain information that genuinely protects them (e.g., [Adjerid et al., 2013](#); [Urban, 2014](#); [Tucker, 2017](#)). Second, our empirical results on firm behaviour are consistent with a “hedging” effect in which, as we show next, firms are aware of risks stemming from the collection of consumer data, and create sophisticated policies to hedge those.

3.3 Data Extraction Practices and Cybersecurity Risks

This section presents our third set of stylized facts. We show that data extraction, whether in the form of cookies or personal data, increases cybersecurity risks. Figure 6a plots the number of cookies and data points collected by firms grouped in three categories. The first are those that do not, in a given year, mention “privacy” in their risk factors. The second, firms that mention it, but further down compared to the median for that year, thus implying lower perceived risks from the firm’s perspective. Finally, the third category groups firms that mention “privacy” earlier in their disclosed risk factors than the median firm. The graphs show that data collection linearly increases with the perceived risk level. Figure 6b shows that it is not only perceived risks that increase with data collection but also realized risks. We plot the average number of cookies and data points collected for firms between 2015 and 2023 against whether or not those firms have suffered a cyberintrusion during that same period. The figure demonstrates that firms that gather more data are more likely to suffer a cyberintrusion.

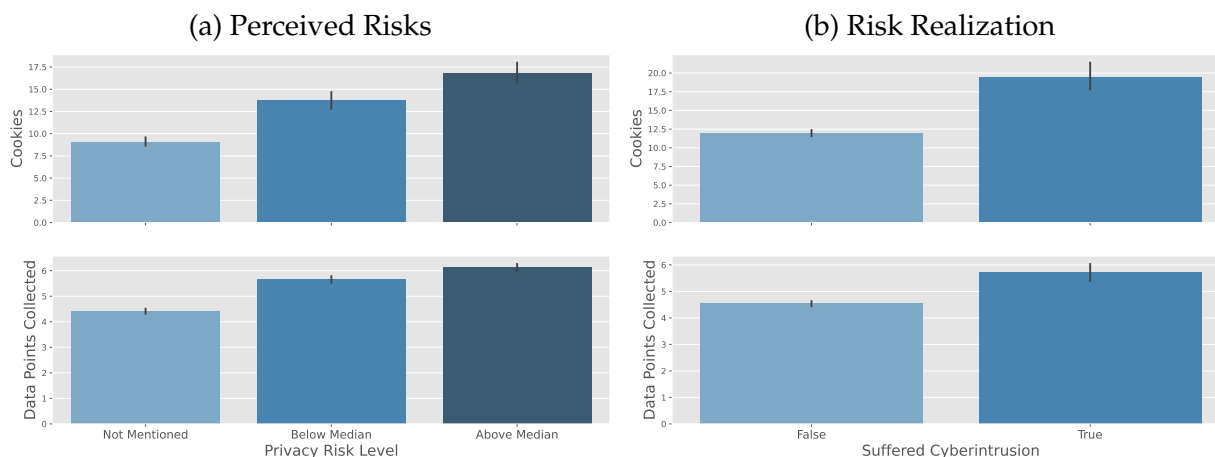


Figure 6: Data Extraction Practices and Privacy Risks

Note: Figure 6a shows the number of cookies and data points collected by firms, divided into three categories: firms that do not mention privacy in their 10-K, those that mention it below the yearly median, and those that mention it above the median (i.e. sooner in the 10-K). Figure 6b plots the average number of cookies and data points collected by firms between 2015 and 2023 that have (“True”) and have not (“False”) suffered a cyberintrusion between 2015 and 2013.

3.4 Firm Characteristics and Privacy

This section contains our fourth set of stylized facts. We investigate whether firms’ economic characteristics can be used systematically to predict data extraction behaviour and the nature of their privacy policies. We start by analyzing the relationship between the characteristics of privacy policies and two firm characteristics that seem intuitively important as a simple first step, namely, firm size and a measure of firms’ technical sophistication. Our measure of technical sophistication is Taylor (2017)’s measure of “knowledge capital”, which is essentially past accumulated R&D expenditures for firms 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 3, and term the result the “knowledge share.”

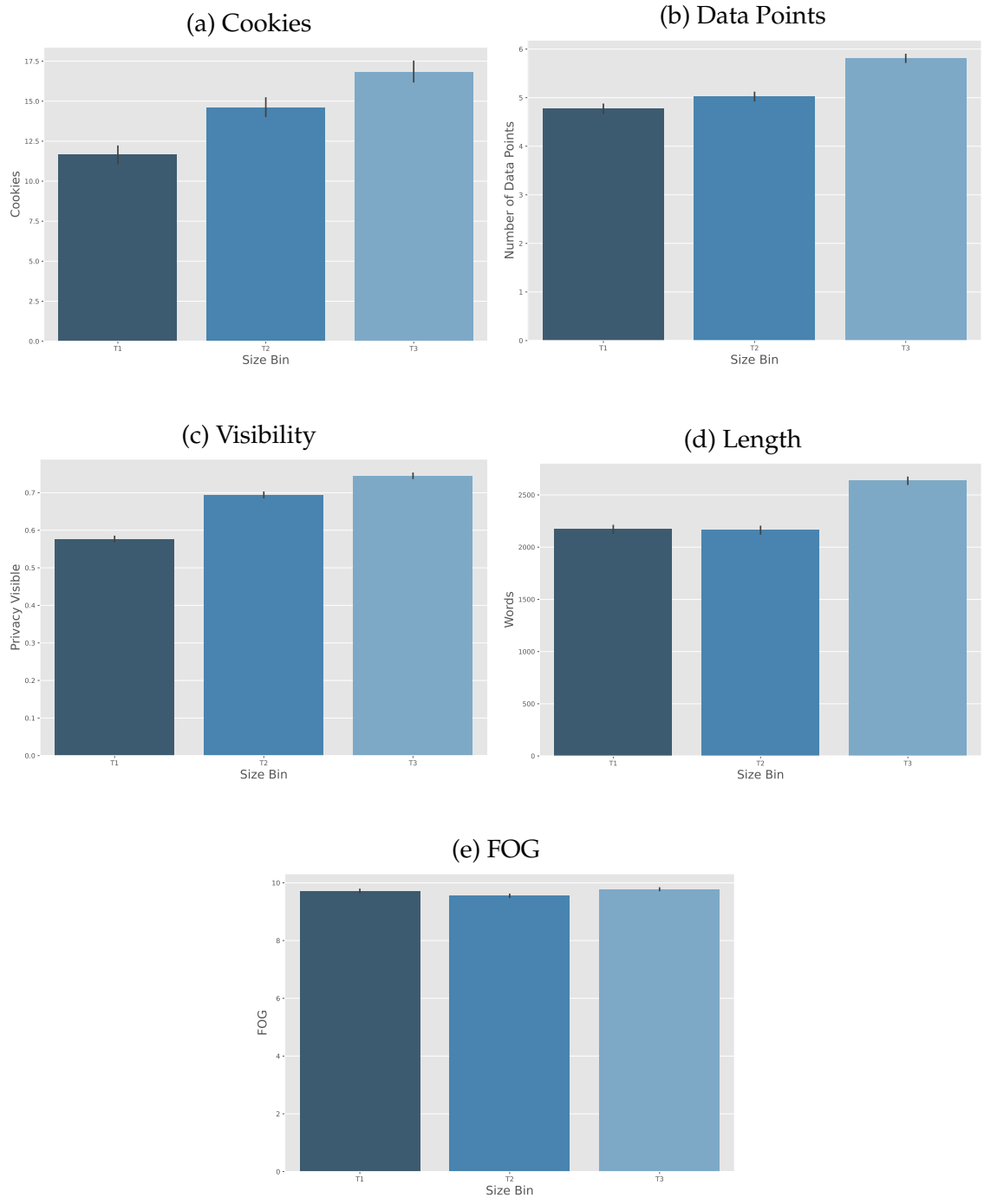


Figure 7: Market Value Bins

Note: These figures respectively plot the number of cookies, data points gathered, visibility, length and readability index of firm privacy policies by firm size terciles. Figure 7a is based on policies gathered during the year of the scan dates (2019 and 2023). For Figures 7b, 7d and 7e we keep at most one policy per firm per year. For Figure 7c we keep at most one home page per firm per year.

We first sort firms into terciles of size, as measured by market value, and plot the associated characteristics of privacy policies and data extraction behaviour. Figure 7a and 7b shows that larger firms are significantly more likely to extract consumer data. Moreover, Figures 7c to 7e show patterns that are consistent with our results in the previous section: Larger firms, who extract more data, are also more likely to have a privacy policy and to display it visibly. Conditional on having a privacy policy, large firms write longer but more intelligible policies.

Figure 8 sorts firms into firms with zero knowledge capital (about 55% of our sample), and then into three terciles conditional on having positive knowledge capital. Figure 8a shows an interesting pattern. The average number of cookies is higher for the second tercile and lowest for the third; a result corroborated by Figure 8b which shows an identical pattern. Those results are consistent with our results in Section 3.2. Indeed, firms in the second tercile of knowledge share are most likely to have a privacy policy and to make it visible, and they also write the longest privacy policies.

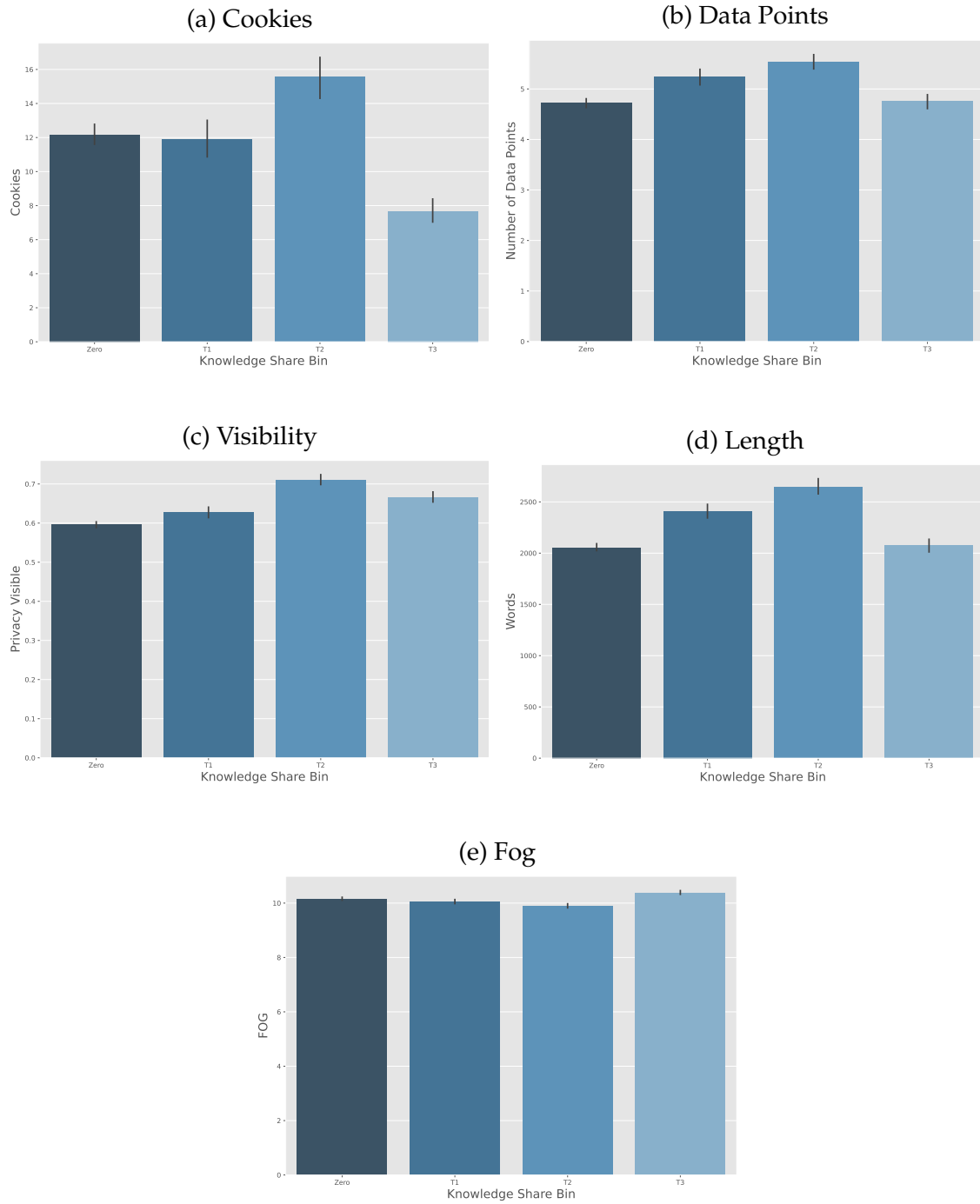


Figure 8: Knowledge Share Bins

Note: These figures plot the number of cookies, data points gathered, visibility, length, and readability index of firm privacy policies by firm knowledge share categories. Figure 8a is based on policies gathered during the year of the scan dates (2019 and 2023). For Figures 8b, 8d and 8e we keep at most one policy per firm per year. For Figure 8c we keep at most one home page per firm per year.

Table 4 confirms these findings in multivariate regressions of data extraction and privacy policy attributes on firm characteristics. The top panel (a) of the Table contains regressions without fixed effects, while the bottom panel (b) of the table contains year and sector fixed effects at the level of SIC divisions.²⁰ 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. We further 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 8. The fourth right-hand side variable is the log Market Share measured as the firm's sales divided by industry sales at the 2-digit SIC code level; we include this because the classical theory (e.g., [Spence, 1975](#)) suggests that market power is a possible determinant of non-price attributes of firm behaviour, such as privacy.

²⁰See Figure 4b for the definition of SIC divisions.

Table 4: Policy Attributes: Regressions

This table presents multivariate regression results examining the relationship between firm characteristics and both data extraction practices and privacy policy attributes. The table is divided into two panels: Panel (a) reports regressions without fixed effects, while Panel (b) includes year (when possible) and sector fixed effects at the SIC division level. The dependent variables represent different privacy policy attributes, while the independent variables include firm size, knowledge share (along with its square), and market share.

(a) Without Fixed Effects

	Policy Visible	Log Words	Fog Index	Cookies	Data Collected
Log Market Value	0.046*** (0.008)	0.040*** (0.012)	-0.123 (0.078)	1.217** (0.473)	0.262*** (0.038)
Log Market Share	-0.006 (0.014)	0.053*** (0.012)	0.051 (0.106)	0.463 (0.290)	0.125*** (0.031)
Knowledge Share	0.480 (0.339)	1.833*** (0.513)	-0.288 (1.977)	16.426 (13.173)	4.036* (2.098)
Knowledge Share ²	-0.449 (0.313)	-2.089*** (0.728)	1.897 (1.633)	-25.886 (20.017)	-3.984 (2.722)
Intercept	-0.346 (0.270)	7.051*** (0.337)	13.025*** (2.449)	-7.671 (12.131)	0.654 (1.156)
Obs	25894	12661	12661	4826	12661

Note: Standard Errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.
Standard Errors are clustered at the Sector level.

(b) With Fixed Effects

	Policy Visible	Log Words	Fog Index	Cookies	Data Collected
Log Market Value	0.038*** (0.006)	0.042*** (0.008)	-0.071 (0.061)	1.224** (0.534)	0.217*** (0.024)
Log Market Share	0.000 (0.012)	0.045*** (0.008)	0.022 (0.088)	0.418 (0.332)	0.146*** (0.030)
Knowledge Share	0.464** (0.185)	1.324*** (0.338)	-0.586 (0.942)	16.105*** (5.546)	3.239*** (0.799)
Knowledge Share ²	-0.375*** (0.127)	-1.490*** (0.454)	2.177*** (0.526)	-22.510*** (8.415)	-2.433*** (0.786)
Sector Fixed Effects	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	No	Yes
Obs	25894	12661	12661	4826	12661

Note: Standard Errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.
Standard Errors are clustered at the Sector level.

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 in-

roduction of the fixed effects, but only limited 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. Table B.1 in the Appendix shows the robustness of our results when we control for a wider set of firm characteristics, including firms' marketing expenditures as a fraction of total assets, and firms' market-to-book ratios.

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. Table B.2 in the Appendix confirms that these patterns continue to arise, and indeed become stronger, as per the point estimates, when we focus on data-intensive firms. This once again suggests that our results are not driven by the mechanical impact of firms that have no interest in consumer data.

Overall, we find robust size effects, which are consistent with large firms engaging in more data extraction. This result can perhaps be explained easily in a model where there are economies of scale, or fixed costs, associated with the collection and processing of consumer data. We also find a persistent non-linear effect of technical sophistication, as measured by firm capital generated through research and development. The most sophisticated firms behave differently, engage in less data extraction, and write more concise privacy policies than those with intermediate sophistication. This finding suggests that there may be multiple dimensions of technical sophistication. For instance, it is conceivable that the firms with the most research-related capital have alternative ways of acquiring consumer data (e.g. because they are themselves data intermediaries). Additionally, such firms may simply have the ability to process consumer data “in-house” rather than sharing it with third-party intermediaries, thus reducing privacy risks and obviating the need for an extensive privacy policy.

3.5 The Promising Effects of Privacy Regulations

This section contains our final set of stylized facts, focusing on the time series. Privacy has been a rising concern for both individuals and regulators over the period in our sample, which captures three important regulatory changes. The first is the General Data Protection Regulation (GDPR), the second is the California Consumer Privacy Act (CCPA) and finally the Virginia Consumer Data Protection Act (VCDPA). Those regulatory frameworks, along with the now defunct EU–US Privacy Shield are prominently mentioned in the privacy policies, as shown in Figure 9a. Furthermore, policies change following new regulations. Figure 9b plots the average cosine similarity of policies compared to their previous snapshot. It shows, quite clearly a drop following both the implementation of GDPR and CCPA.



Figure 9: Policy Changes over Time

Note: These Figures present a series of time-series analyses on privacy policy attributes and data extraction practices in response to major regulatory changes, including the General Data Protection Regulation (GDPR), implemented on May 25th, 2018, the California Consumer Privacy Act (CCPA), implemented on January 1st, 2020, and the Virginia Consumer Data Protection Act (VCDPA), implemented on January 1st, 2023. For Figures 9a, 9c, 9d, and 9e, we keep at most one policy per firm per month. Figure 9b plots the cosine average cosine per month of a policy compared to its previous version in our sample. For Figure 9f we keep the closest policy to our scan dates (February 28th, 2019 and May 10th, 2023).

In light of these large regulatory changes, we might expect policies and cookies to have adapted to regulations. Most notably, GDPR requires privacy information notices to be “concise, transparent, intelligible and easily accessible; written in clear and plain language”. Given those stated goals, we should first expect that more companies and websites have their privacy policies displayed prominently. Figure 9c shows that this is the case: while in 2016, 55% of the websites we scanned featured the word privacy on their home page, this number increased to over 80% by the end of our sample.

Second, despite the conciseness requirement imposed by GDPR, we should expect privacy policies to have gotten lengthier over time as per the transparency requirements. Once again, this is very much the case, as shown in Figure 9d rising from 2,500 words in 2017 to over 4,000 on average in 2023. Given intelligibility requirements, privacy policies should have gotten simpler over time. This is reflected in Figure 9e with a significant decrease in FOG.

Finally, the transparency requirement can be tested by using the cookies data. We do find an increase in truthfulness, shown in Figure 9f: in our 2019 scan, we found that firms that did not mention cookies in their privacy policies had an average of 8. As of our 2023 scan, this number decreased to a little over 4.

4. Two Business Models of Consumer Data

Following the previous findings, in particular those in Sections 3.3 and 3.4, we might expect firms of intermediate sophistication to acknowledge the cybersecurity risks associated with their data extraction and sharing practices. Figure 10a plots the positional risk of “privacy” by terciles of knowledge share, after removing industry-year fixed effects. It shows that firms in the second tercile of knowledge share mention those risks earlier in their 10-Ks. If, as the SEC dictates, these risks are listed in order of importance, then these firms acknowledge the risks they face arising from their business practices.

A question remains whether firms of intermediate sophistication are disproportionately affected by cybersecurity events (i.e. ex-post). On the one hand, gathering consumer data makes a firm a more valuable target to malicious actors, and sharing data with third parties creates more attack vectors. On the other hand, since these firms acknowledge these risks, they likely implement defensive measures and, in particular, spend more on cybersecurity thus reducing the likelihood of a cybersecurity incident.

Additionally, and more unambiguously, when a cyberintrusion happens, however rare, firms in the second tercile of knowledge share should have, compared to other firms, greater costs associated with such incidents as they (1) gather more data on their consumers and would therefore face more extensive fines but also (2) face potentially greater disruptions when losing access to those data, either temporarily (e.g., ransomware) or permanently.

Figures 10b and 10c tackle these questions. Figure 10b plots the number of cyberbreaches by knowledge share terciles between 2015 and 2023 after removing industry-year fixed effects and shows that firms in the first and second tercile have had more cyberbreaches compared to those in the third tercile. This suggests that risk mitigation efforts engaged in by these firms do not fully mitigate cyberthreats. Figure 10c plots breach costs, which include fines as well as remediation costs, when those are disclosed, divided by the absolute value of net income by terciles of knowledge share, after removing industry fixed effects²¹. Since breaches are rare, we average the knowledge share and sum the costs between 2015 and 2023. The figure shows that firms in the second tercile of knowledge share are disproportionately affected by fines and remediation costs. We confirm these findings in Table B.3 of the appendix which regresses the positional risk found in the 10-K²² and the breach costs divided by net income against knowledge share.

²¹It is reasonable to divide by net income for two reasons. First, it stands to reason that fines should be commensurate with the amount of data a firm has. Hence we need a proxy for this number. Second, privacy regulations typically set guidance values for fines via accounting calculations. For example, Article 83(5) of GDPR states that the fine can be up to 20 million euros, or up to 4% of the total global turnover of the preceding fiscal year. Here, we divide by net income to not penalize low-margin industries.

²²Filling in a "1" 10-Ks that do not mention the word "privacy", meaning that it is a low ranked concern



Figure 10: Privacy Risks by Knowledge Share

Note: Figure 10a plots the positional risk of “privacy” mentioned in firms’ annual reports by terciles of knowledge share for that same year, after removing industry-year fixed effects. Given the rarity of cybersecurity incidents, Figures 10b and 10c sum events and fines over 2015-2023 while knowledge share bins are based on the average for firms over the sample period. Industry-year fixed effects are also removed.

While results in Figure 8 and 10, might suggest firms in the third tercile of knowledge share are less reliant on consumer data, we argue that this is not the case. Figure 11 counts the occurrences of different word patterns within the 10-K risk factors: the number of times we find “data” to be preceded, within 50 characters, by “collect”, “process”, “shar[e]”, or “receiv[e]”. We then plot those occurrences by knowledge share, removing for such firms.

industry-year fixed effects. The resulting Figures 11a and 11b show that firms in the second and third terciles of knowledge share mention collecting and processing data at the same rate: both are reliant on consumer data.

Our preferred explanation of these results is that there are two principal business models relating to consumer data extraction. The first, which we name “collect and share”, is common among firms with an intermediate level of sophistication that do not have the expertise to extract insights in-house. Those firms extract (Figures 8a and 8b) and rely on consumer data (Figures 11a and 11b), monetise it via data sharing with third-parties (Figure 11c), have higher cybersecurity risks, of which they are aware of (Figure 10a), and which are not fully offset by firm investments in protection, as they have greater remediation costs (Figure 10c). This suggests that cybersecurity risks are concentrated amongst firms that are reliant on data sharing, which creates vulnerabilities and risks, through different attack vectors²³. Finally, such firm legally “hedge” these risks by creating relatively more sophisticated privacy policies which create a presumption of caveat emptor (Figure 8d).

The second model which we name “receive and process” is more prevalent among firms with high technical sophistication that have the expertise to extract insights in-house. Those firms gather comparatively less consumer data than their less sophisticated counterparts (Figures 8a and 8b) although they remain reliant on it (Figures 11a and 11b), process it in-house rather than sharing it (Figure 11c), indeed receiving data from third-parties for such processing (Figure 11d). As a result, they have lower cybersecurity risks (Figure 10) and a lower need for sophisticated privacy policies as a means to hedge those risks (Figure 8d).

²³This was recently proven true with the 2024 AT&T hack in which a data breach exposed phone call and text message records for 110 million people via a cloud partner, Snowflake: “Earlier this year, malicious hackers figured out that many major companies have uploaded massive amounts of valuable and sensitive customer data to Snowflake servers, all the while protecting those Snowflake accounts with little more than a username and password.” Source: krebsonsecurity.com/2024/07/hackers-steal-phone-sms-records-for-nearly-all-att-customers

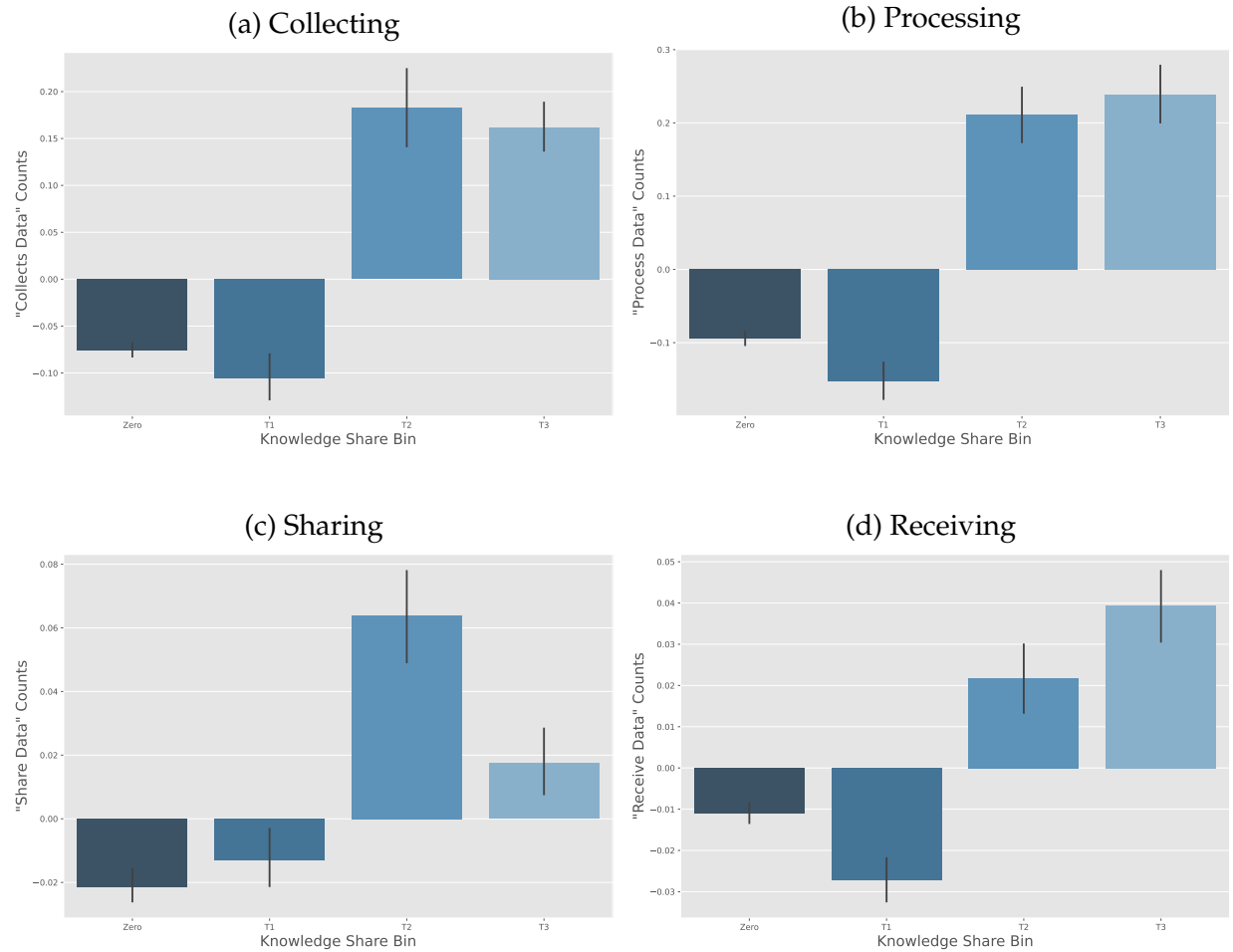


Figure 11: 10-K Risk Factors Counts

Note: Figures 11a, 11b, 11c and 11d plot respectively the counts of “collect”, “process”, “shar[e]”, and “receiv[e]” when followed by the word “data” within 50 characters in the 10-K risks factors by categories of knowledge share bins for the same year. Industry fixed effects calculated at the yearly level are removed.

This finding on cybersecurity risks is of particular interest, given that the existing literature has primarily associated cybersecurity with growth opportunities, R&D expenditures, and trade secrets (Florackis et al., 2023). As investors tend to demand higher risk premiums from firms vulnerable to cyber threats, understanding this variation is especially critical in a context where these risks are expected to escalate. In this case, we find the focus to be on consumer data, particularly the sharing of such data. This conclusion

is further supported by Appendix Figure A.5 which plots counts of reported stolen information following a hack²⁴. It shows that intellectual property ranks last, while consumer data (names, social security numbers, addresses, etc.) are among the most prevalent types of information accessed by malevolent actors.

5. Conclusion

By analyzing a large set of US firms' privacy policies through time, we uncover new facts about firms' supply of data privacy. There is significant variation in the ease of acquiring and finding firms' privacy policies, and when found, these policies do not follow a standard boilerplate. Instead, their text varies substantially both within and across industries and is clearly shaped by regulations. Perhaps surprisingly, the visibility, length, and legibility of policies are all correlated with the extent to which firms extract data from consumers browsing their websites and using their services. We find that the variation in policy text is systematic, with large firms and those with intermediate levels of knowledge capital exhibiting longer and more visible policies—correlated with their greater extraction of data on their users. These data in turn create risks for the collecting firms, particularly for those with intermediate levels of knowledge capital which have, relative to their net income, larger costs associated with cybersecurity incidents.

Our preferred interpretation of these facts is one in which data-using firms fall into one of two business models with respect to their consumer data: “collect and share” versus “receive and process”. We find that the first model is common amongst firms with intermediate technical sophistication which do not have the expertise to extract insights in-house, forcing them to rely on data sharing with third-parties, which subsequently creates cybersecurity risks which they do not—or cannot—fully offset. The second model, “receive and process”, is more prevalent among firms with higher technical sophistication.

²⁴The data is from Audit Analytics, which is described in Section 2.2.3

These companies have the in-house expertise to derive insights from data, and extract comparatively less data from their consumers, as they primarily data from third-parties. Such firms have lower cybersecurity risks and therefore less need to hedge these risks via sophisticated privacy policies.

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 firms' data extraction behaviour, which we hope our work will help to spur. Future research can build on these insights in several key directions. First, on the empirical side, granular data on firms' privacy policy updates and their interactions with regulatory changes could shed light on the causal impact of evolving privacy laws on firms and consumer outcomes. For example, what downstream effects do regulation shifts have on firms' market power and competitive dynamics? Further exploration of the relationship between data extraction and one's ability to compete would help clarify trade-offs faced by firms of intermediate sophistication.

Theoretically, there is room to develop models that integrate firm heterogeneity, regulatory pressures, and cybersecurity risks to understand further the trade-offs firms face when designing their data strategies. Such models could then explore the welfare implications of these strategies and regulations for consumers. Another promising avenue lies in analysing how advancements in AI and machine learning might alter the two business models of consumer data we document, perhaps reducing the sophistication gap and the need for sharing.

Data availability

The data used to create the figures and tables in this paper have been made available.

References

- Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar, 2022, “Too much data: Prices and inefficiencies in data markets”, *American Economic Journal: Microeconomics* 14, 218–256.
- Acquisti, Alessandro, Laura Brandimarte, and George Loewenstein, 2015, “Privacy and human behavior in the age of information”, *Science* 347, 509–514.
- Acquisti, Alessandro, Leslie K. John, and George Loewenstein, 2013, “What is privacy worth?”, *Journal of Legal Studies* 42, 249–274.
- Adjerid, Idris, Alessandro Acquisti, Laura Brandimarte, and George Loewenstein, 2013, “Sleights of privacy: Framing, disclosures, and the limits of transparency”, in *Proceedings of the ninth symposium on usable privacy and security* pp. 1–11.
- Admati, Anat R., and Paul Pfleiderer, 1986, “A monopolistic market for information”, *Journal of Economic Theory* 39, 400–438.
- Athey, Susan, Christian Catalini, and Catherine Tucker, 2017, “The digital privacy paradox”, Working Paper.
- Bergemann, Dirk, and Alessandro Bonatti, 2019, “Markets for information: An introduction”, *Annual Review of Economics* 11, 85–107.
- Bui, Duc, Kang G. Shin, Jong-Min Choi, and Junbum Shin, 2021, “Automated extraction and presentation of data practices in privacy policies”, *Proceedings on Privacy Enhancing Technologies* 2021, 88–110.
- Chen, Long, Yadong Huang, Shumiao Ouyang, and Wei Xiong, 2021, “The data privacy paradox and digital demand”, Working Paper.
- Eisfeldt, Andrea L., and Dimitris Papanikolaou, 2014, “The value and ownership of intangible capital”, *American Economic Review* 104, 189–194.
- Englehardt, Steven, and Arvind Narayanan, 2016, “Online tracking: A 1-million-site measurement and analysis”, in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* pp. 1388–1401.
- Fabian, Benjamin, Tatiana Ermakova, and Tino Lentz, 2017, “Large-scale readability analysis of privacy policies”, in *Proceedings - 2017 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2017* pp. 18–25. Association for Computing Machinery, Inc.
- Finkel, Jenny Rose, Trond Grenager, and Christopher Manning, 2005, “Incorporating non-local information into information extraction systems by gibbs sampling”, in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* pp. 363–370.
- Florackis, Chris, Christodoulos Louca, Roni Michaely, and Michael Weber, 2023, “Cybersecurity risk”, *Review of Financial Studies* 36, 351–407.

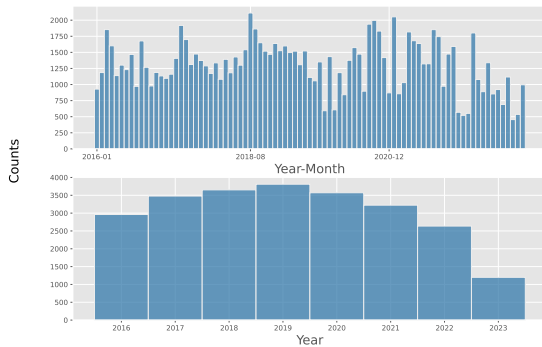
- Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, "Text as data", *Journal of Economic Literature* 57, 535–574.
- Goldfarb, Avi, and Catherine Tucker, 2012, "Shifts in privacy concerns", *American Economic Review* 102, 349–353.
- Gunning, Robert, 1952, *The Technique of Clear Writing* (McGraw-Hill).
- Hoberg, Gerard, and Gordon Phillips, 2010, "Product market synergies and competition in mergers and acquisitions: A text-based analysis", *Review of Financial Studies* 23, 3773–3811.
- Hoofnagle, Chris Jay, and Jennifer M. Urban, 2014, "Alan westin's privacy homo economicus", *Wake Forest Law Review* 261.
- Ichihashi, Shota, 2021, "Competing data intermediaries", *The RAND Journal of Economics* 52, 515–537.
- Liu, Zhuang, Michael Sockin, and Wei Xiong, 2020, "Data privacy and temptation", Working Paper.
- Mcdonald, Aleecia M., and Lorrie Faith Cranor, 2008, "The cost of reading privacy policies", *I/S: A Journal of Law and Policy for the Information Society* 4.
- U.S. Department of the Treasury, 2024, "Managing artificial intelligence-specific cybersecurity risks in the financial services sector", Report.
- Peters, Ryan H., and Lucian A. Taylor, 2017, "Intangible capital and the investment-q relation", *Journal of Financial Economics* 123, 251–272.
- Rajaraman, Anand, and Jeffrey Ullman, 2011, *Mining of Massive Datasets* (Cambridge University Press).
- Spence, A Michael, 1975, "Monopoly, quality, and regulation", *The Bell Journal of Economics* 6, 417–429.
- Tang, Huan, 2019, "The value of privacy: Evidence from online borrowers", Working Paper.

Appendix for

“Privacy Policies and Consumer Data Extraction: Evidence From U.S. Firms”

A. Figures

(a) Histogram of Privacy Policies



(b) Histogram of Homepages

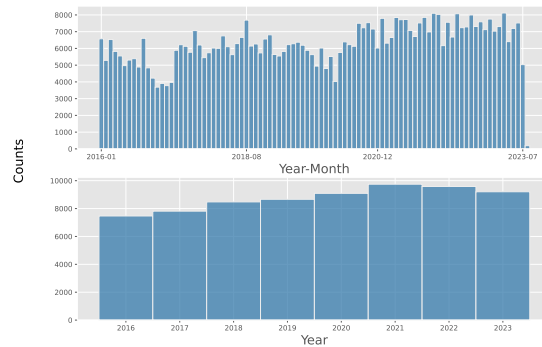


Figure A.1: Wayback Machine Screenshots Dates

Note: Figure A.1a plots counts of the number privacy policies available through the Wayback Machine per month for the first panel and year for the second. Figure A.1b plots counts of the number of homepages available through the Wayback Machine per month for the first panel and year for the second.

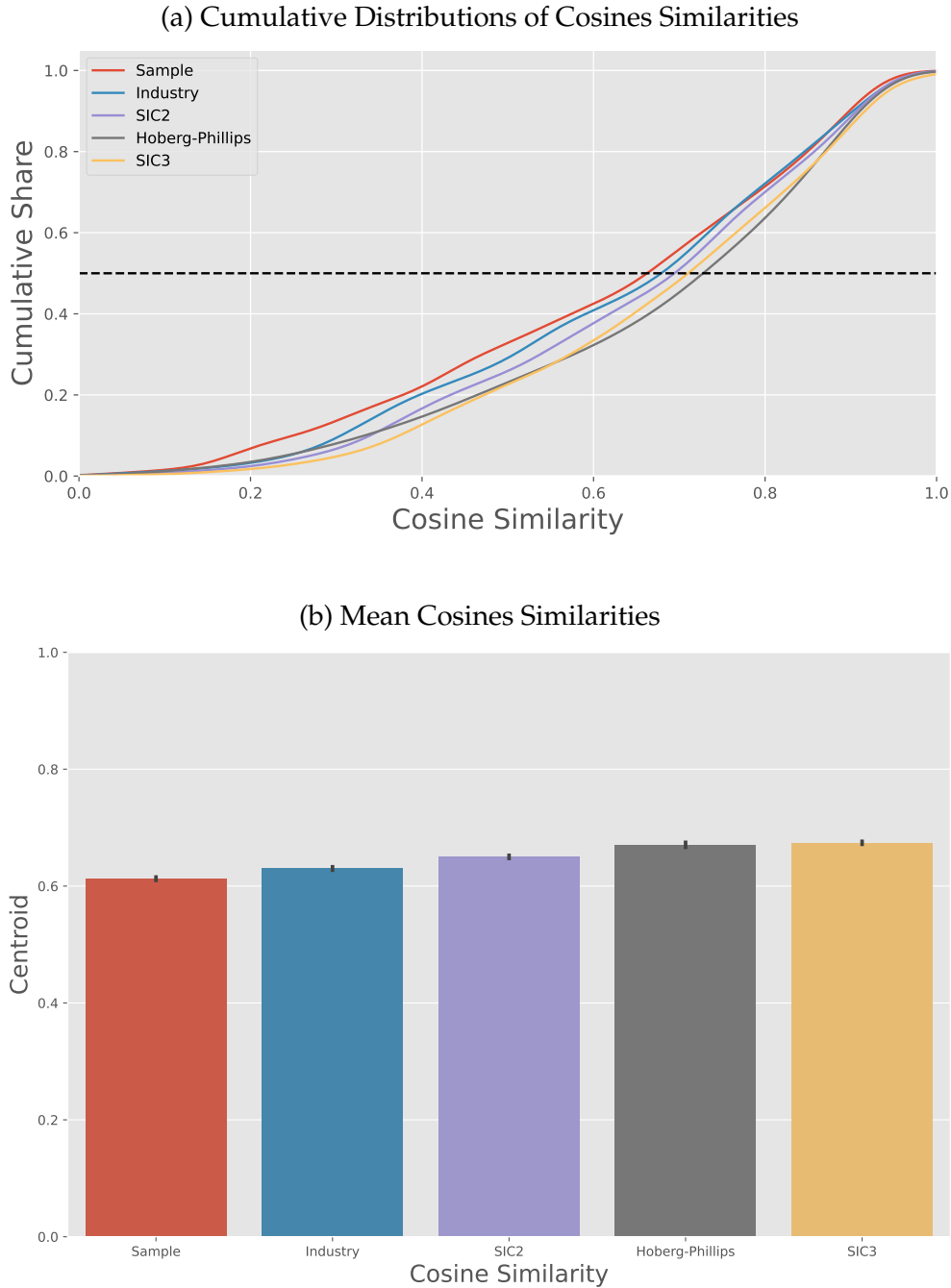


Figure A.2: Cosine Similarities of Privacy Policies with LSA

Note: The Sample centroid is the mean LSA vector across our privacy policies, keeping one policy per year per firm. Industry centroids are the mean LSA 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 Nonclassifiable (9900,9999). SIC2 and SIC3 centroids are mean frequencies at the 2-digit and 3-digit SIC code levels, respectively. Hoberg-Phillips centroids are based on firms competitors as defined by [Phillips \(2010\)](#) (data available at hoberg-phillips.tuck.dartmouth.edu).

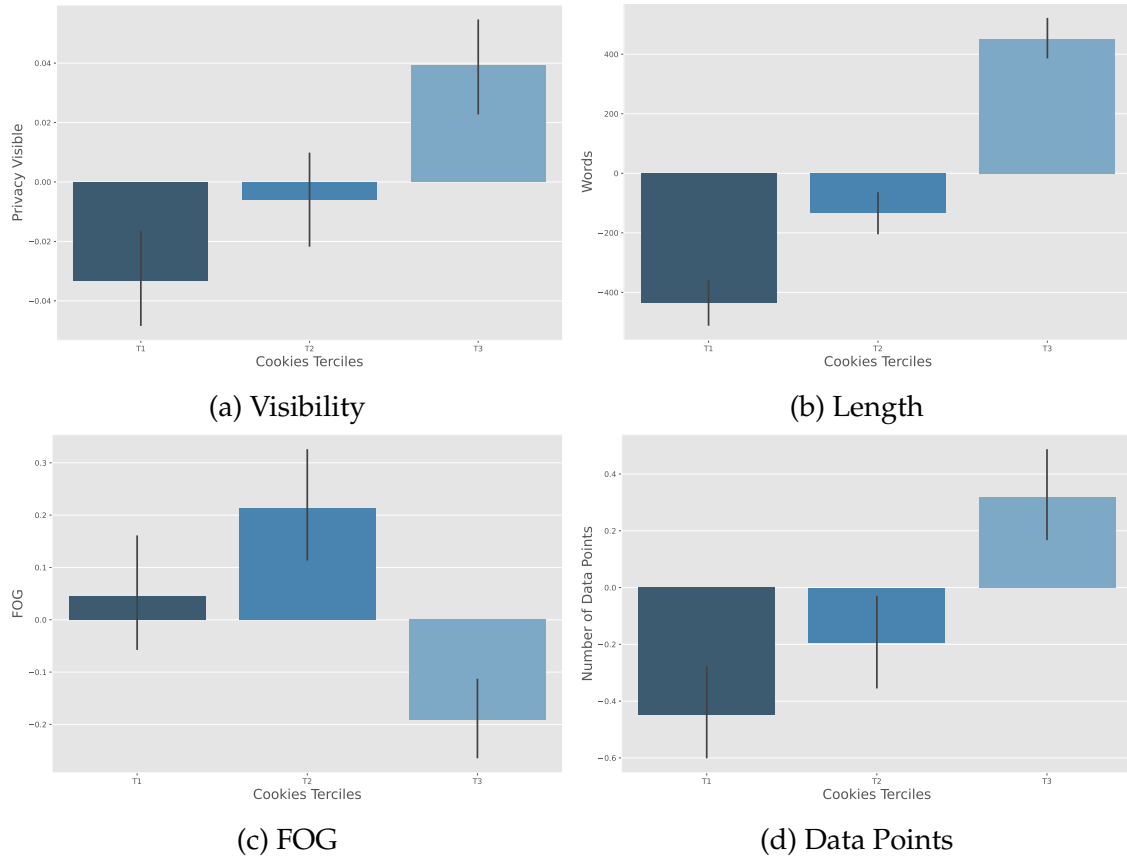


Figure A.3: Privacy Policies and Total Cookies, removing fixed effects

Note: These figures present the average policy visibility, length, fog and number of data points collected for firms sorted by terciles of the total number of cookies they collect, removing industry-year fixed effects. For each firm in our sample, we keep the home page (Figure 5a) or privacy policy (Figures 5b, 5c and 5d) closest to the two scan dates performed through privacyscore.org: February 28th, 2019 and May 10th, 2023.

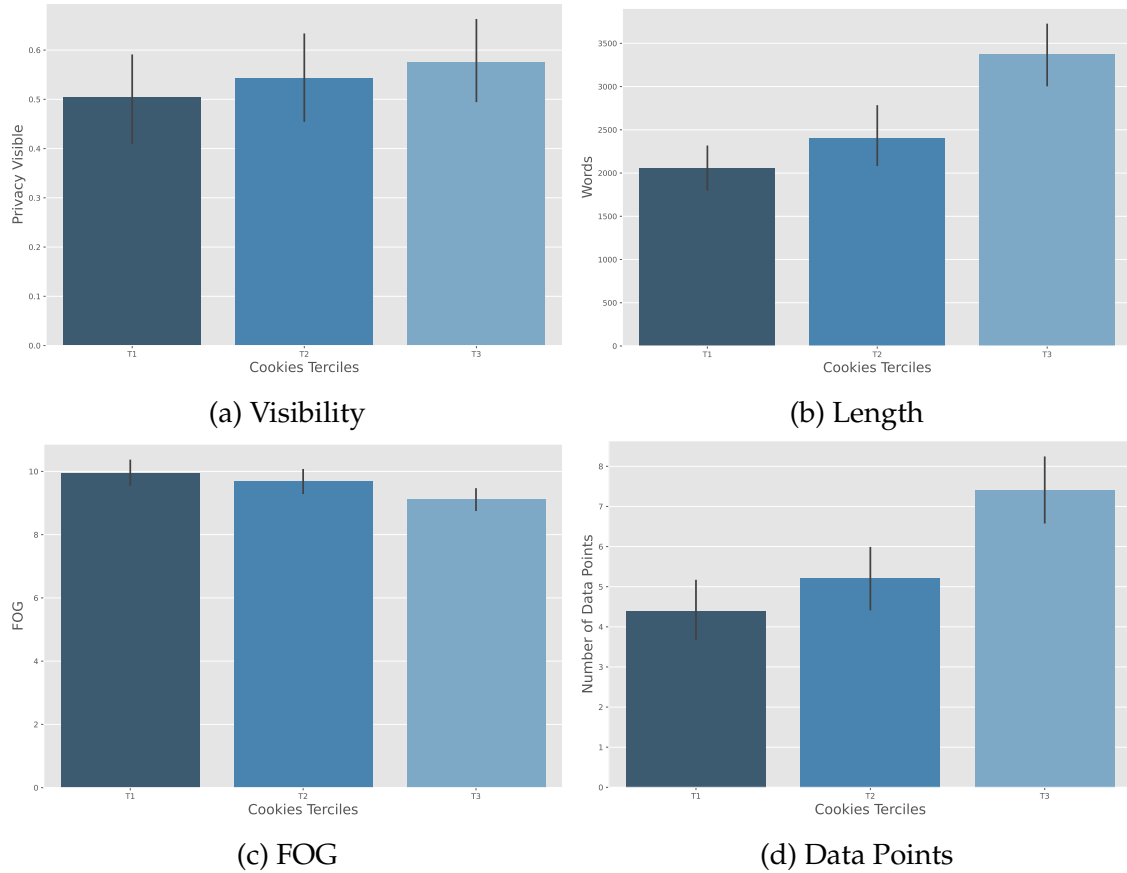


Figure A.4: Privacy Policies and Total Cookies for Data Intensive Firms

Note: These figures present the average policy visibility, length, fog and number of data points collected for firms sorted by terciles of the total number of cookies they collect, for data-intensive firms. For each firm in our sample, we keep the home page (Figure 5a) or privacy policy (Figures 5b, 5c and 5d) closest to the two scan dates performed through privacyscore.org: February 28th, 2019 and May 10th, 2023.

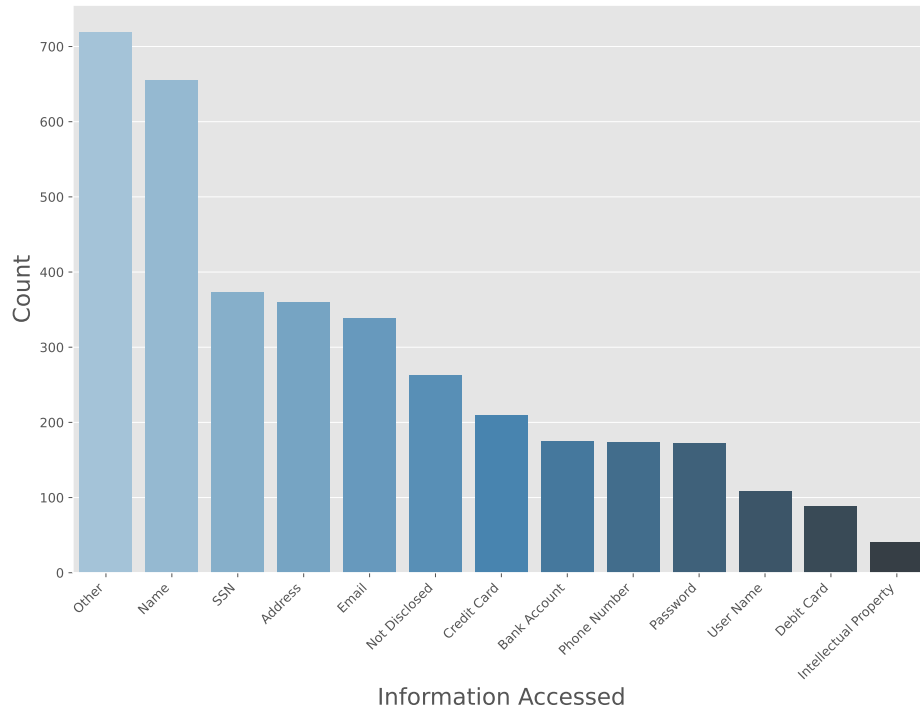


Figure A.5: Information Accessed in Hacks

Note: Figure A.5 plots counts of reported stolen information following a hack, in which attackers may access multiple types of information. In a number of cases, the accessed information was not disclosed (6th column).

B. Tables

Table B.1: Policy Attributes: Regressions

This table presents multivariate regression results examining the relationship between firm characteristics and both data extraction practices and privacy policy attributes. The table is divided into two panels: Panel (a) reports regressions without fixed effects, while Panel (b) includes year (when possible) and sector fixed effects at the SIC division level. The dependent variables represent different privacy policy attributes, while the independent variables include firm size, knowledge share (along with its square and whether it is equal to zero), market share, market to book (and whether it is equal to zero), marketing to assets (and whether it is equal to zero).

(a) Without Fixed Effects

	Policy Visible	Log Words	Fog Index	Cookies	Data Collected
Log Market Value	0.044*** (0.005)	0.039*** (0.011)	-0.119** (0.058)	1.276*** (0.365)	0.279*** (0.034)
Log Market Share	-0.005 (0.010)	0.053*** (0.013)	0.045 (0.077)	0.518** (0.209)	0.123*** (0.032)
Knowledge Share	0.608*** (0.110)	1.207*** (0.304)	0.187 (0.619)	20.286* (10.485)	2.925*** (0.892)
Knowledge Share ²	-0.466*** (0.117)	-1.231*** (0.425)	0.551 (0.490)	-25.835 (15.856)	-2.196* (1.179)
Zero Knowledge Capital	0.036 (0.040)	-0.095** (0.039)	-0.043 (0.244)	1.738*** (0.600)	-0.073 (0.217)
Market To Book	-0.000*** (0.000)	0.000*** (0.000)	-0.000 (0.000)	0.001*** (0.000)	-0.000*** (0.000)
Market To Book Missing	-0.050*** (0.015)	-0.016 (0.030)	0.064 (0.130)	3.186*** (0.688)	0.398* (0.219)
Marketing / Assets	-0.135 (0.098)	0.134 (0.305)	0.592 (0.619)	4.967 (7.172)	2.567*** (0.491)
Marketing Missing	-0.140*** (0.039)	-0.178*** (0.035)	0.957*** (0.354)	-6.878*** (0.903)	-0.749*** (0.209)
Intercept	-0.220 (0.177)	7.258*** (0.335)	12.313*** (1.748)	-5.597 (8.298)	0.768 (0.981)
Obs	25892	12660	12660	4826	12660

Note: Standard Errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.
Standard Errors are clustered at the Sector level.

(b) With Fixed Effects

	Policy Visible	Log Words	Fog Index	Cookies	Data Collected
Log Market Value	0.037*** (0.004)	0.040*** (0.008)	-0.067 (0.049)	1.308*** (0.449)	0.227*** (0.027)
Log Market Share	0.000 (0.009)	0.047*** (0.010)	0.015 (0.068)	0.440* (0.251)	0.148*** (0.035)
Knowledge Share	0.571*** (0.093)	1.050*** (0.230)	0.175 (0.522)	17.750** (7.198)	2.560*** (0.654)
Knowledge Share ²	-0.414*** (0.078)	-1.046*** (0.289)	0.701 (0.437)	-21.389** (10.208)	-1.371*** (0.530)
Zero Knowledge Capital	0.043 (0.030)	-0.056* (0.030)	0.076 (0.200)	1.620** (0.809)	-0.088 (0.176)
Market To Book	-0.000*** (0.000)	0.000*** (0.000)	-0.000 (0.000)	0.000 (0.000)	-0.000*** (0.000)
Market To Book Missing	-0.044*** (0.014)	-0.041 (0.030)	0.053 (0.112)	2.490*** (0.786)	0.301 (0.235)
Marketing / Assets	-0.137 (0.099)	-0.011 (0.250)	0.565 (0.512)	2.346 (7.800)	2.085*** (0.408)
Marketing Missing	-0.100*** (0.035)	-0.136*** (0.023)	0.755** (0.334)	-5.781*** (0.574)	-0.439*** (0.125)
Sector Fixed Effects	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	No	Yes
Obs	25892	12660	12660	4826	12660

Note: Standard Errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.
Standard Errors are clusters at the Sector level.

Table B.2: Policy Attributes: Regressions for Data Intensive Firms

This table presents multivariate regression results examining the relationship between firm characteristics and both data extraction practices and privacy policy attributes, including year (when possible) and sector fixed effects at the SIC division level, focusing on data-intensive firms. The dependent variables represent different privacy policy attributes, while the independent variables include firm size, knowledge share, market share, market.

	Policy Visible	Log Words	Fog Index	Cookies	Data Collected
Log Market Value	0.016*** (0.006)	-0.013 (0.030)	-0.029 (0.021)	1.695 (1.230)	0.015 (0.150)
Log Market Share	0.010 (0.009)	0.082*** (0.008)	0.065 (0.143)	0.177 (1.088)	0.326*** (0.106)
Knowledge Share	0.551*** (0.171)	2.184*** (0.553)	0.375 (1.684)	28.415** (12.609)	6.592*** (0.882)
Knowledge Share ²	-0.671*** (0.163)	-2.425*** (0.589)	0.873 (1.123)	-36.862*** (11.200)	-4.973*** (1.181)
Sector Fixed Effects	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	No	Yes
Obs	1704	838	838	354	838

Note: Standard Errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.
Standard Errors are clusters at the Sector level.

Table B.3: Privacy Risks: Regressions

The table presents multivariate regression results examining the relationship between firm characteristics and privacy risks, including year (when possible) and sector fixed effects at the SIC division level. The table includes two main dependent variables: positional risk, which captures how prominently privacy risks are mentioned in firms' 10-K reports, and cost-to-net-income ratios, which measure the financial impact of cybersecurity breaches, while the independent variables include firm size, knowledge share, market share, market.

	Positional Risk	Positional Risk	Cost/ NI	Cost/ NI
Log Market Value	-0.026*** (0.005)	-0.022*** (0.006)	-1120.004 (1232.907)	-408.849 (1010.347)
Log Market Share	0.004 (0.003)	0.001 (0.003)	-3151.186*** (1076.021)	-4415.160*** (1026.923)
Knowledge Share	-0.581** (0.229)	-0.499*** (0.057)	38665.329* (21639.317)	19246.492 (16798.556)
Knowledge Share ²	0.626* (0.356)	0.467*** (0.092)	-121597.495*** (34010.488)	-102326.082*** (14445.635)
Intercept	1.432*** (0.098)		15679.835 (35189.906)	
Sector Fixed Effects	No	Yes	No	Yes
Year Fixed Effects	No	Yes	No	Yes
Obs	59940	59940	784	784

Note: Standard Errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.
Standard Errors are clustered at the Sector level.