

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

Tarun Ramadorai Antoine Uettwiller Ansgar Walther

Workshop on Artificial Intelligence in Finance

March 20th 2025

Motivation

Growing interest in consumers' data privacy and firms ability to track.

- ▶ Demand: Many consumers are passive, “consent fatigue”.
 - Goldfarb and Tucker (2012), Acquisti et al. (2015).
- ▶ Privacy paradox: stated preferences vs. behavior and willingness to pay.
 - Athey et al. (2017), Tang (2019).
- ▶ Reassurance by mere presence of legal text.
 - Acquisti et al. (2016).

Lately:

- ▶ Firms increasingly rely on consumer data for insights.
- ▶ Privacy and security risks have become more significant challenges for firms.

Understanding the *supply* of privacy and how firms balance data utilization with risk management is important.

Data Collection: For a comprehensive set of US firms, we measure:

1. What they say: Privacy policy text.
2. What it means: What kind of data they gather on their consumers.
3. What they do: Data extraction via cookies.

Stylized facts using variation across firms:

- ▶ Most of the variation is within industries: No boilerplates.
- ▶ Data extraction is associated with long and visible policies.
- ▶ Heightened cybersecurity risks for extracting firms (both perceived ex-ante and ex-post).
- ▶ Systematic variation across firm characteristics: Size and technical sophistication.

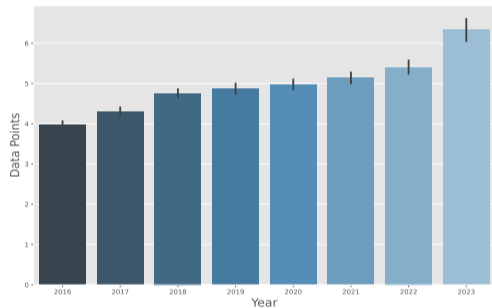
Models of Consumer Data:

1. Firms that do not participate in data extraction.
2. Firms that rely on a “collect and share” model.
3. Firms that rely on a “receive and process” model.

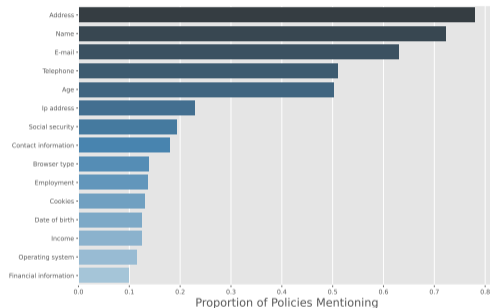
Data

Privacy Policies: Data Collected

(a) Mean Data Points Collected per Year



(b) Most Recurring Data Points

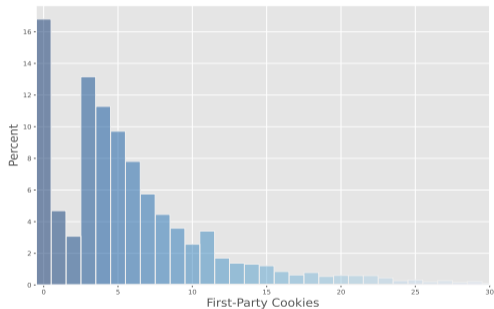


We use OpenAI and GPT-4 to extract the data point firms collect on consumers:

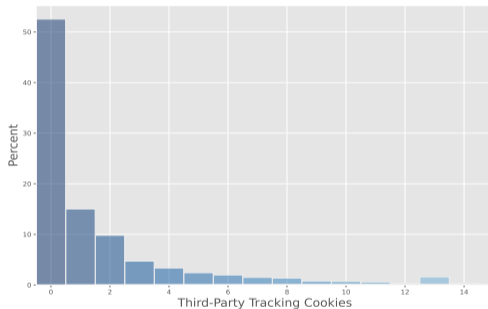
- ▶ “Can you give me list of information they collect?”

Data Extraction: Cookies

(a) First-Party Cookies



(b) Third-Party Tracking

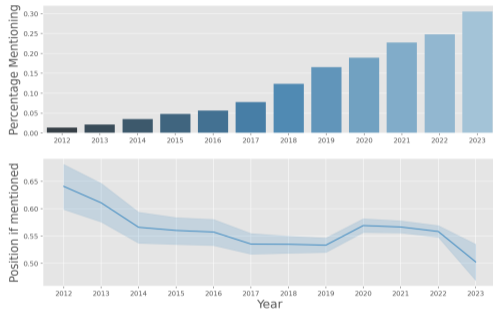


We use the OpenWPM crawler developed by Englehardt and Narayanan (2016).

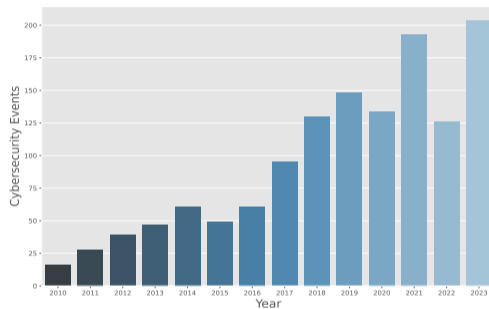
- We ran the crawler twice. Once in 2019, another in 2023.

Cybersecurity Risks

(a) Data Breaches 10-K Positional



(b) Cybersecurity Events



We use two different measures of cybersecurity risks:

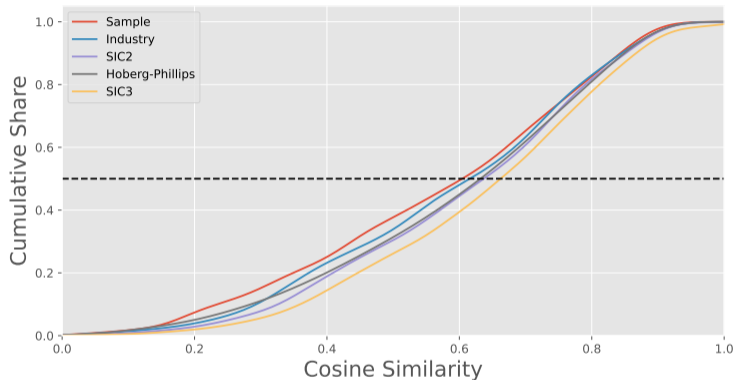
1. **Perceived:** Position within the 10-K risk factors.

- SEC: “Companies generally list the risk factors in order of their importance”.

2. **Ex-post:** Cybersecurity incidents and costs.

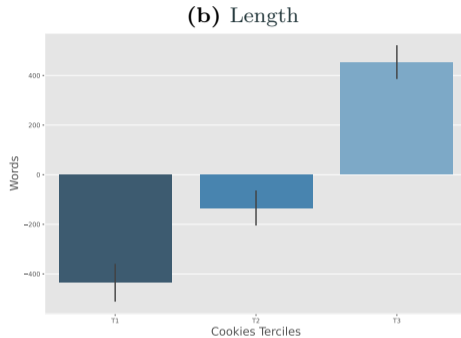
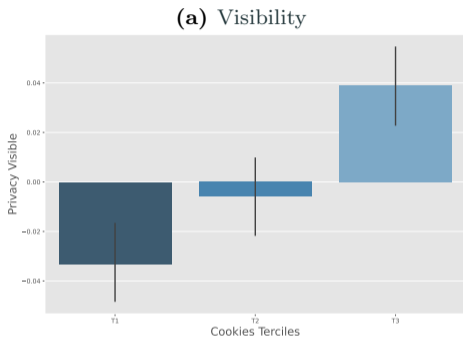
Stylized Facts

Variation: No Industry Boilerplates



- ▶ No boilerplates.
- ▶ Most of the variation is within industries, rather than between industries.
- ▶ Even when looking at the closest competitors (Hoberg and Phillips, 2010).

Tracking is Associated with Visible Yet Longer Policies



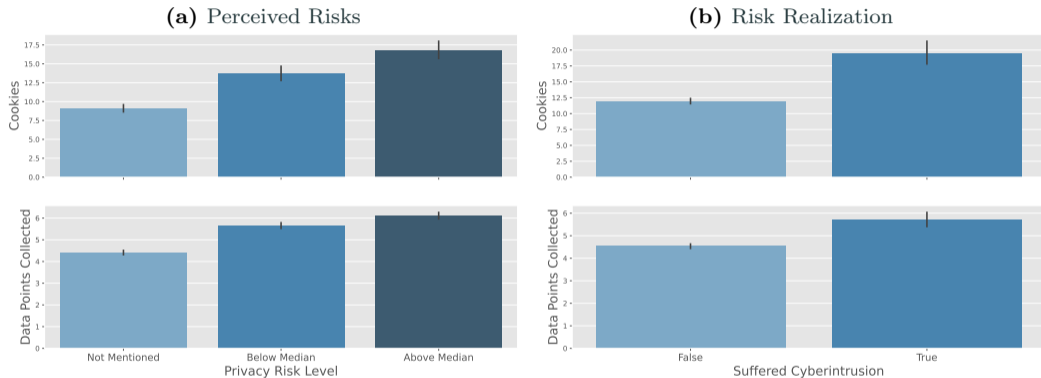
Similar patterns for the number of data points collected.

Plots

Economic interpretation:

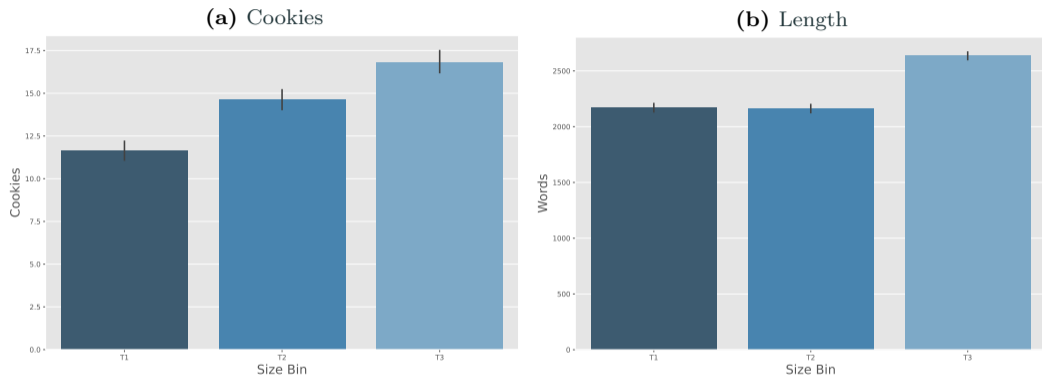
- ▶ Appear to be used by firms as a way to hedge legal risk.
- ▶ Rather than protect consumers.

Data Extraction and Cybersecurity Risks



- ▶ Data collection increases the likelihood of suffering a cyber-intrusion.
- ▶ Cybersecurity risk appears to be more about data than trade secrets (Florackis et al., 2023). Information accessed in hacks

Firm Size, Policies and Behavior

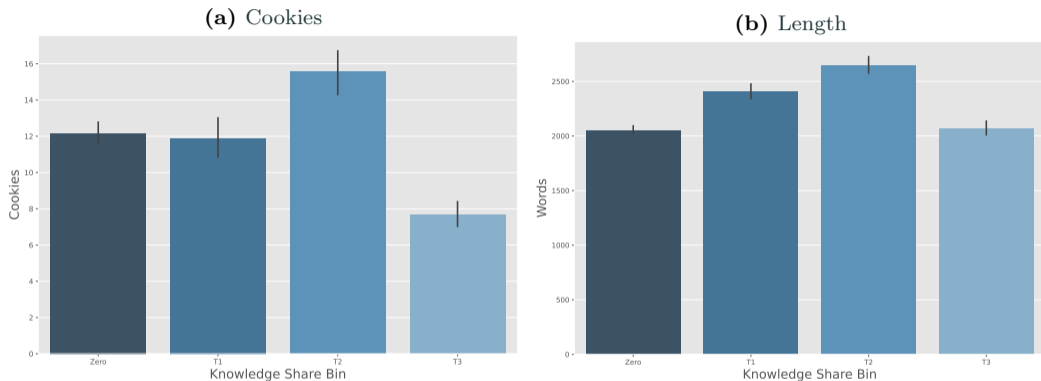


Similar patterns for visibility and the number of data points collected.

Larger firms:

- ▶ Collect more data.
- ▶ Have longer policies.

Knowledge Share, Policies and Behavior



Similar patterns for visibility and the number of data points collected.

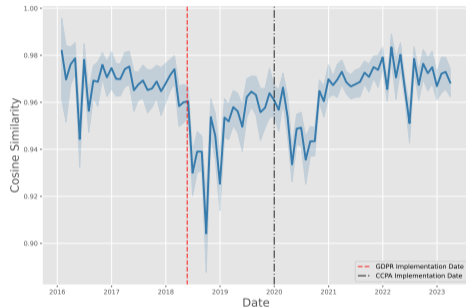
Plots

Firms of intermediate sophistication: Regressions

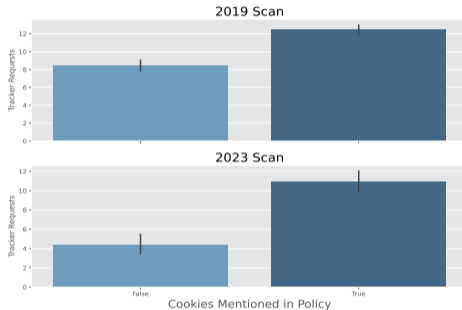
- ▶ Collect more data, especially through cookies.
- ▶ Have longer policies.

Regulation: GDPR and CCPA

(a) Policy Cosine



(b) Truthfulness over Time



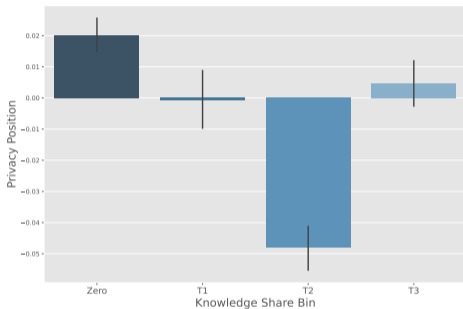
GDPR: Policies must be:

- *“Concise, transparent, intelligible and easily accessible form, using clear and plain language”.*
 - ▶ Firms have become more truthful about their tracking.
 - ▶ Policies have become lengthier and more visible while clearer.

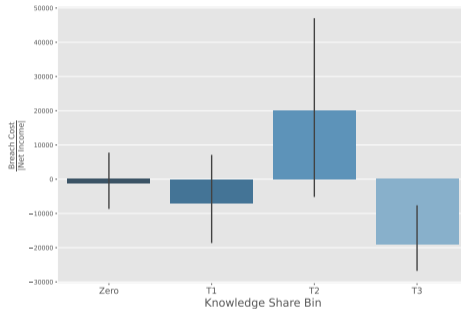
Two Business Models of Consumer Data

Privacy Risks by Knowledge Share

(a) 10-K Privacy Positional Risk



(b) Cybersecurity Failure Costs



Firms of intermediate sophistication:

- ▶ Mention risks coming from privacy sooner within their risk factors.
- ▶ Have greater cybersecurity failure costs.

Regressions

Three Types of Firms

Those results might suggest that:

- ▶ Firms in the 3rd tercile of knowledge share are less reliant on consumer data.
- ▶ We argue this is not the case.

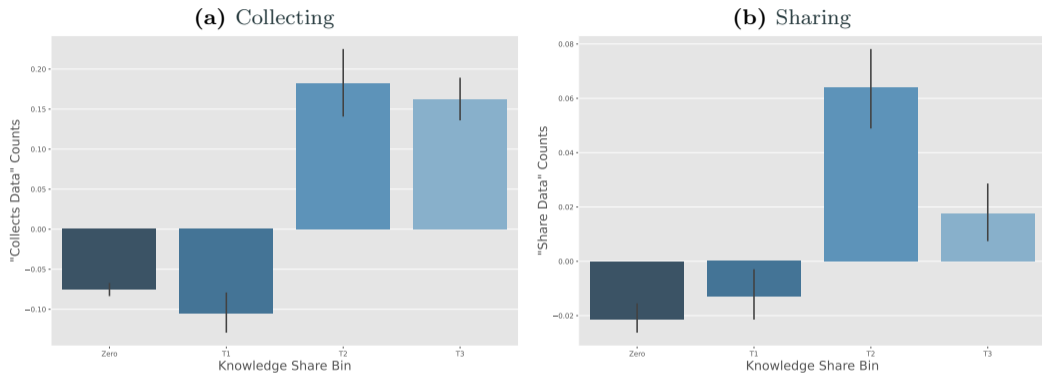
Instead, we argue that two-tier market exists:

1. The “collect and share” model.
2. The “receive and process” model.

To do so, we:

- ▶ Use the risk factors within the annual reports.
- ▶ Count the number of times we find “data“ to be preceded, within 50 characters, by:
 - “collect”, “process”, “shar[e]” and “receiv[e]”

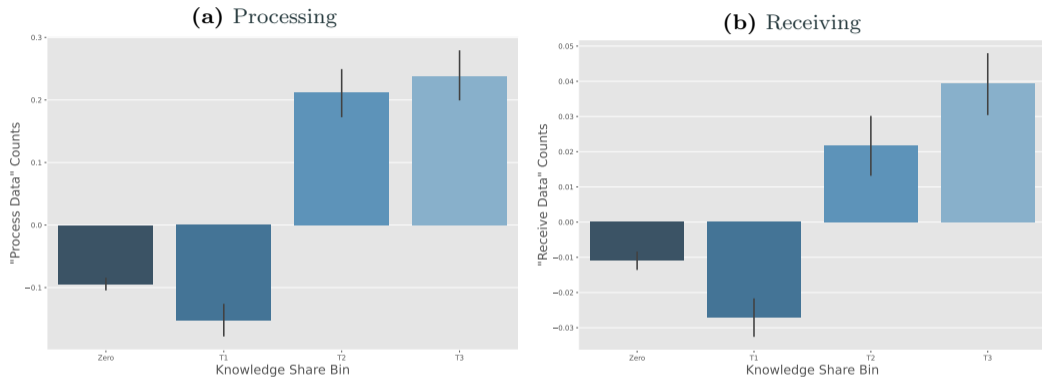
Risk Factors: “Collect” and “Share” model



Firms with an intermediate level of sophistication:

- ▶ Those firms extract and rely on consumer data.
- ▶ Monetise it via data sharing with third-parties.
- ▶ Do not have the expertise to extract insights in-house.

Risk Factors: “Process” and “Receive” model



Firms with high technical sophistication:

- ▶ Gather comparatively less consumer data than their less sophisticated counterparts.
- ▶ Receive data from third-parties.
- ▶ Have the expertise to extract insights in-house.

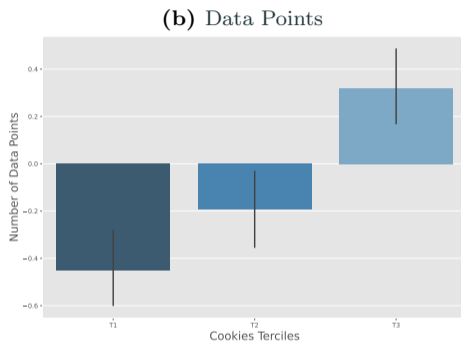
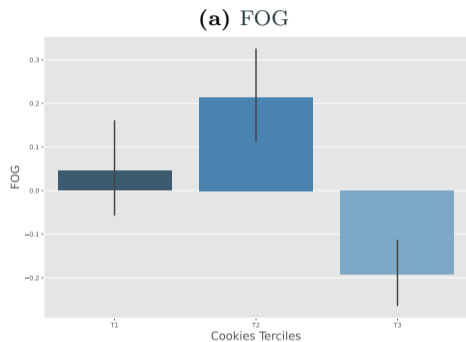
Conclusion

Conclusion

- ▶ We assemble comprehensive data for studying the market for privacy, focusing on the supply side.
 - \approx 6,000 firms between 2016 and 2023.
- ▶ Stylized facts on cross-firm variation.
 - No industry boilerplate
 - Privacy policies are a way to hedge legal risk rather than protect consumers.
 - Systematic variation across firm characteristics (size and technical sophistication).
- ▶ Evidence of a two-tier market for data:
 - “Collect and share”.
 - “Receive and process”.
- ▶ Cybersecurity risk is associated with the “collect and share” model.
 - Less about trade secrets than data.

Appendix

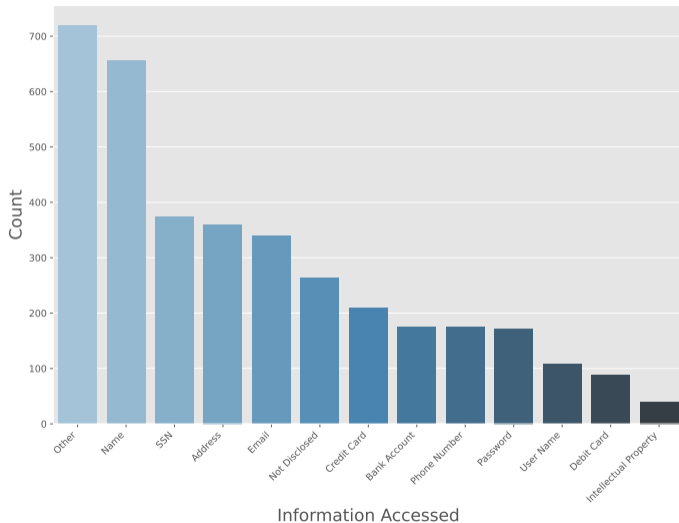
Tracking is Associated with Visible Yet Longer Policies



The “intelligible” aspect of privacy policies, as required by regulators:

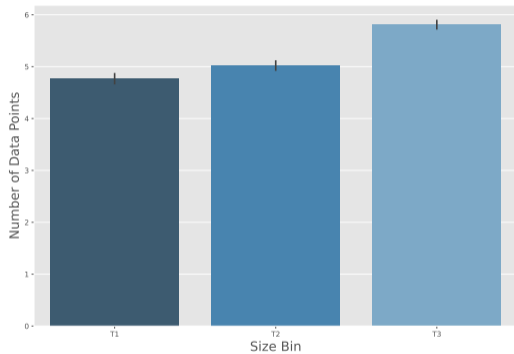
- ▶ Is more commonly implemented by firms that engage in extensive tracking,
- ▶ While companies that engage in minimal tracking have little reason to make their privacy policies opaque.

Data Extraction and Cybersecurity Risks

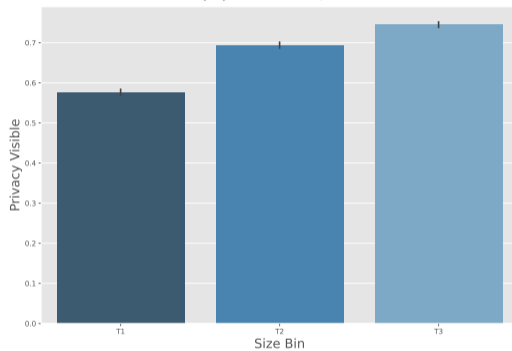


Firm Size, Policies and Behavior

(a) Data Points

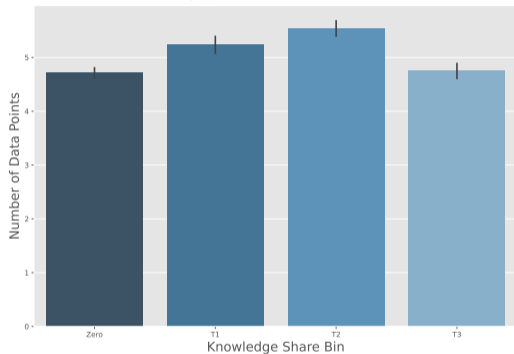


(b) Visibility

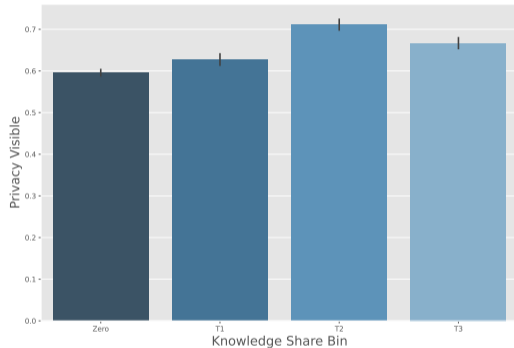


Knowledge Share, Policies and Behavior

(a) Data Points



(b) Visibility



Knowledge Share, Policies and Behavior

Table 1: Policy Attributes: Regressions 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.

Privacy Risks and Knowledge Share

Table 2: Privacy Risks: Regressions

	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 clusters at the Sector level.

References

- Acquisti, A., Brandimarte, L., and Loewenstein, G. (2015) Privacy and human behavior in the age of information, *Science* 347, 509–514.
- Acquisti, A., Taylor, C., and Wagman, L. (2016) The economics of privacy, *Journal of Economic Literature* 54, 442–492.
- Athey, S., Catalini, C., and Tucker, C. (2017) The digital privacy paradox.
- Englehardt, S., and Narayanan, A. (2016) Online tracking: A 1-million-site measurement and analysis, in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1388–1401.
- Florackis, C., Louca, C., Michaely, R., and Weber, M. (2023) Cybersecurity risk, *Review of Financial Studies* 36, 351–407.
- Goldfarb, A., and Tucker, C. (2012) Shifts in privacy concerns, *American Economic Review* 102, 349–353.
- Hoberg, G., and Phillips, G. (2010) Product market synergies and competition in mergers and acquisitions: A text-based analysis, *Review of Financial Studies* 23, 3773–3811.
- Tang, H. (2019) The value of privacy: Evidence from online borrowers.