The Market For Data Privacy

Tarun Ramadorai

Antoine Uettwiller

Ansgar Walther

Data Privacy in the Internet Era

Firms collect, share and aggregate data about a wide range of consumers' online and offline activities

Varian, 2009; Krishnamurthy and Wills, 2009; FTC, 2014

Economics principles are subtle:

- Classical: Consumer data improves efficiency of allocations

  Stigler, 1980; Posner, 1981; Goldfarb and Tucker, 2011

- Second best: Concerns about insurance, price discrimination, negative externalities

  Hirshleifer, 1971; Taylor, 2004; Varian, 2009

Cambridge Analytica

How does the market for data privacy operate?

The Market for Data Privacy

Demand: Many consumers are passive, "consent fatigue"

Goldfarb and Tucker; 2012; Acquisti et al., 2015; Campbell et al., 2018

- Privacy paradox: stated preferences vs. behavior and WTP
- Reassurance by mere presence of legal text

  Norberg et al., 2007; Acquisti, 2016; Athey et al., 2017

Understanding *supply of privacy* is important in this context

This paper: What determines firms' privacy contracts and data sharing policies?

This Paper

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

1. What they say: Privacy policy text
2. What it means: Evaluation of these policies by a legal expert
3. What they do: Third party cookies on websites

Stylized facts using variation across firms:

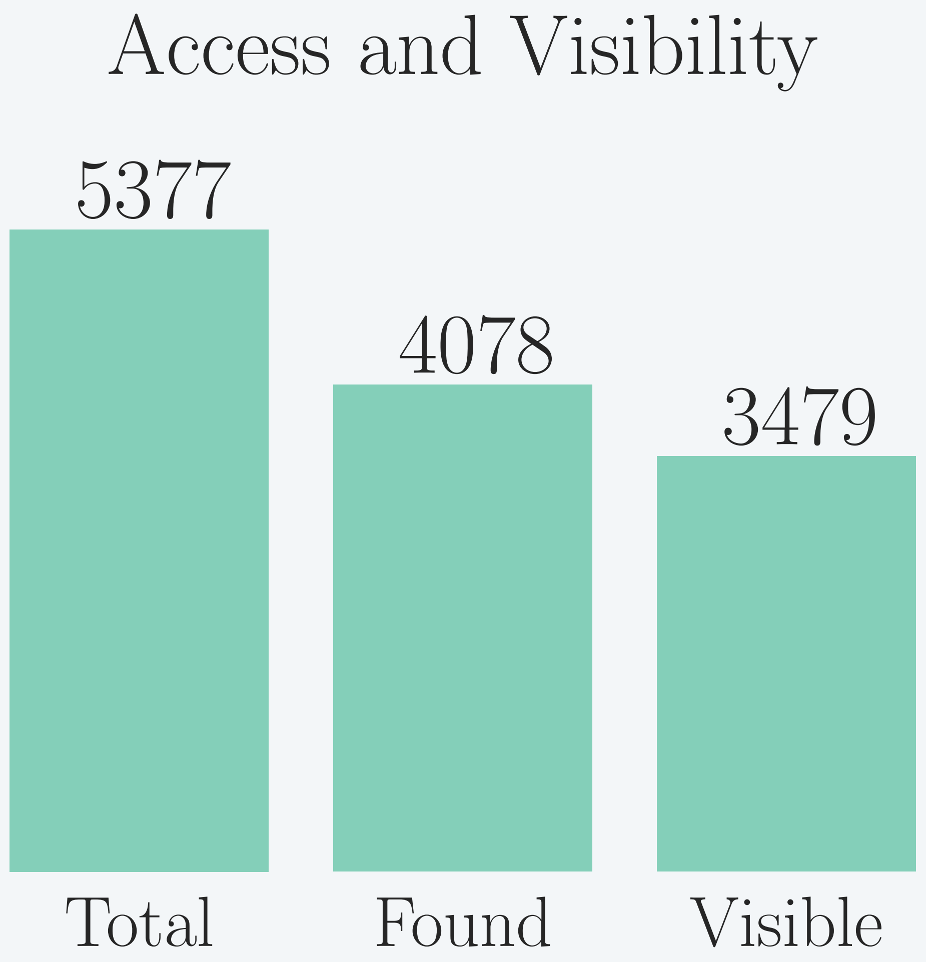
- No standard industry-level boilerplate
- Detailed policies are **associated with more sharing** (fig leaves?)
- Systematic variation across firm characteristics
  - Size and technical sophistication

Theory: Determinants of firms' data sharing and privacy policies

Data

Privacy Policies

N =5377 firms in Compustat US



Finding privacy policies:

- Automated google search
- Web crawling
- Manual checking

Visibility: "Privacy" link on website

Text Analysis Terminology

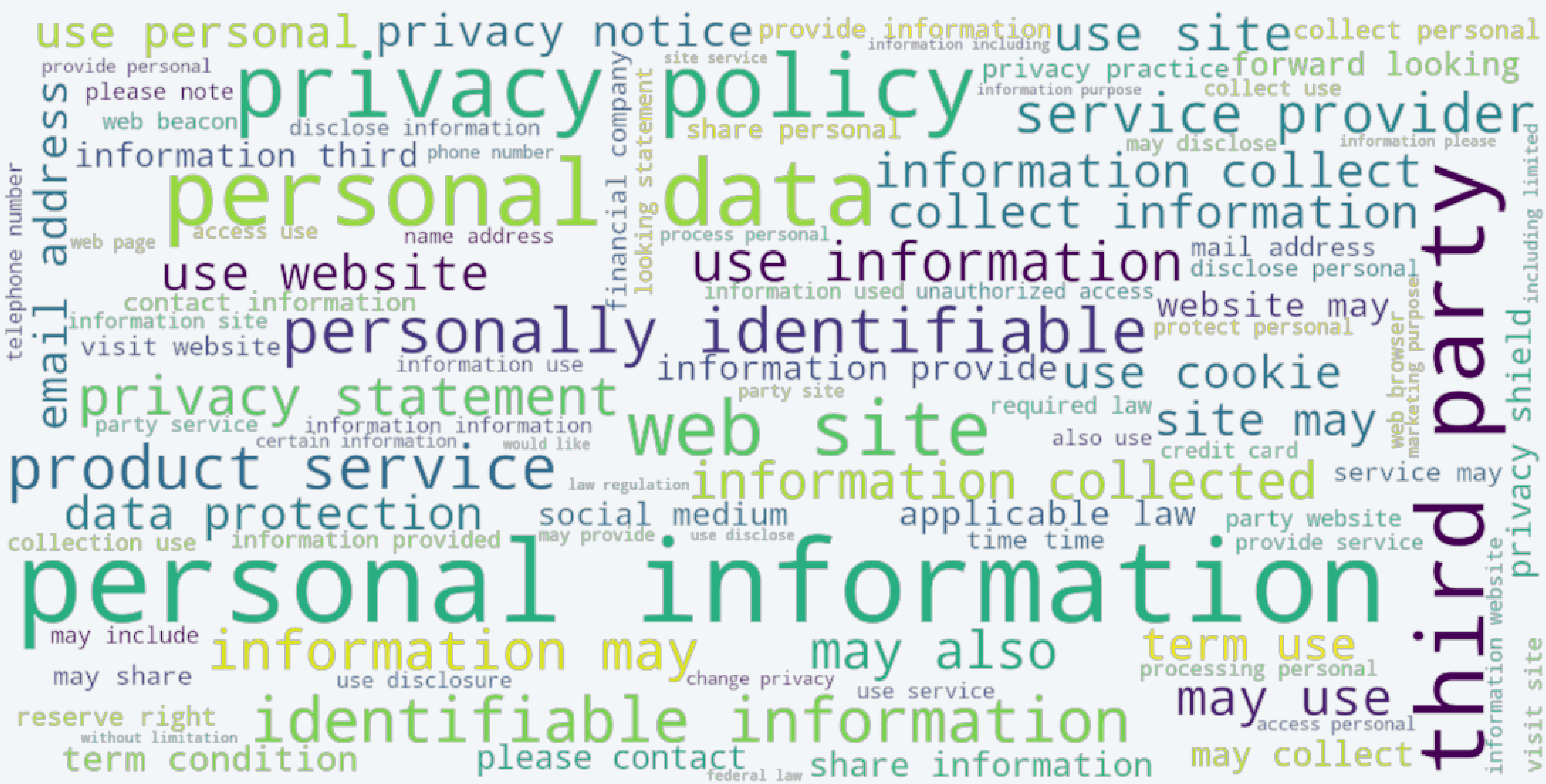
Represent policies as vectors in a term-document matrix

	personal information	third party	personal data	privacy policy	web site	personally identifiable
www.aa.com	22.0	19.7	26.3	11.3	0.0	0.9
www.cecoenviro.com	0.7	10.8	68.6	10.9	0.0	0.0
www.asaltd.com	15.3	0.0	0.0	0.0	0.0	0.0
www.pinnaclewest.com	0.0	3.5	0.0	7.6	3.2	5.7
www.aarons.com	0.0	29.8	0.0	2.8	0.0	4.3

- **TF.IDF**: Transformation rewards frequency, penalizes genericity
- **Cosine similarity:** Angle between two policy vectors (rows)
- **Latent semantic analysis**: Reduce of high-dimensional term-document matrix to loadings on a smaller number of principal components ("topics")

Privacy Policies: Word Cloud

Bigrams, TF.IDF Transformation



Expert Evaluation

We sent 10% of the sample to a legal expert for evaluation

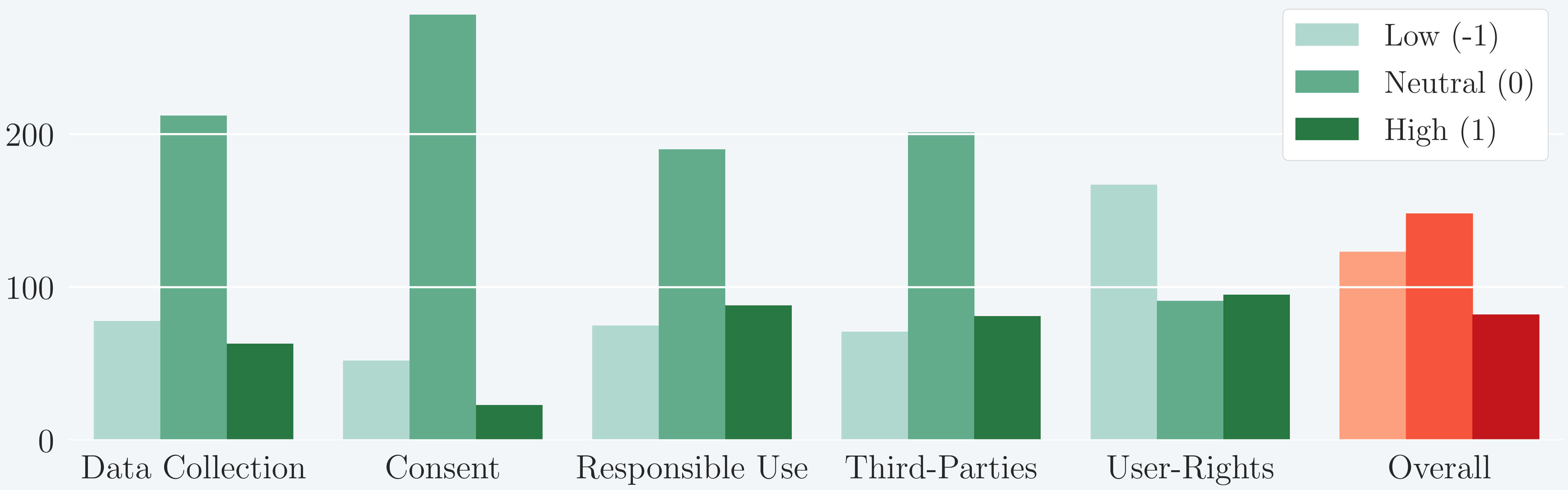
Expert assigned scores (high, neutral, low) along 6 dimensions:

1. **Data Collection:** Clear needs for collection, not excessive
2. **Consent:** Consent *sought* not *presumed*, notified of policy changes
3. **Responsible use:** Clear benefits and robust assurances
4. **Third party sharing:** Clearly explained and legitimate sharing
5. **User rights:** Comprehensive and simple to exercise
6. **Overall**

Emphasis is on legal clarity

Expert Evaluation

10% Sample of Policies



Overall score has strong association with Third Parties ($\rho = 0.68$) and User Rights ($\rho = 0.75$)

Legal Clarity Index

High and low score policies look different, so we construct:

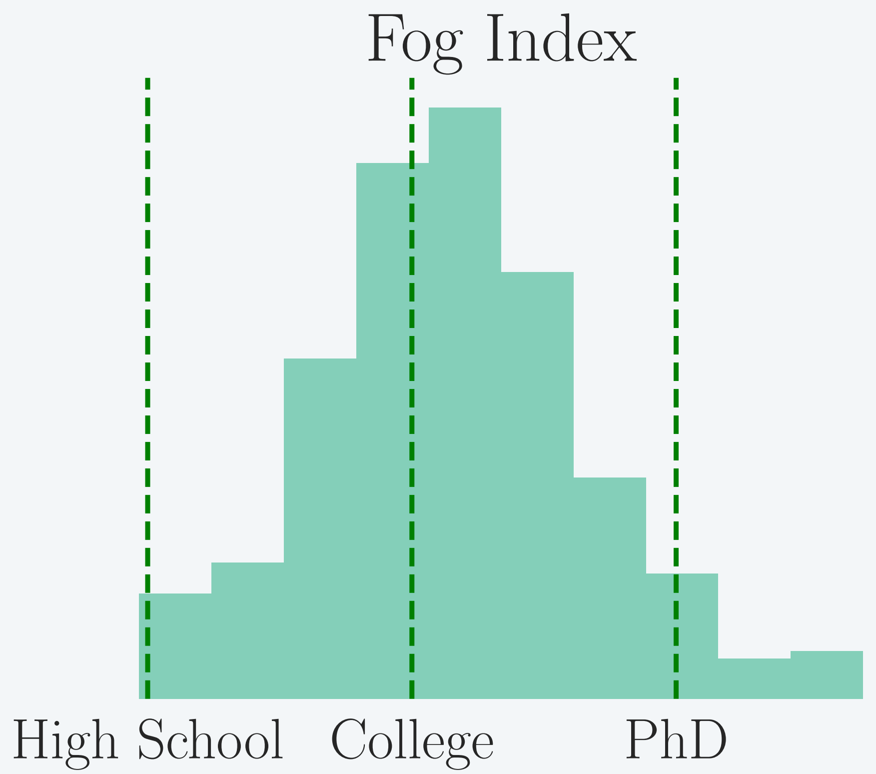
$$\textit{Legal Clarity Index} = \text{Frequency of top 100 "High" bigrams}$$
$$- \text{Frequency of top 100 "Low" bigrams}$$

Similar results with an index that uses supervised machine learning

Readability and Third Party Data Sharing

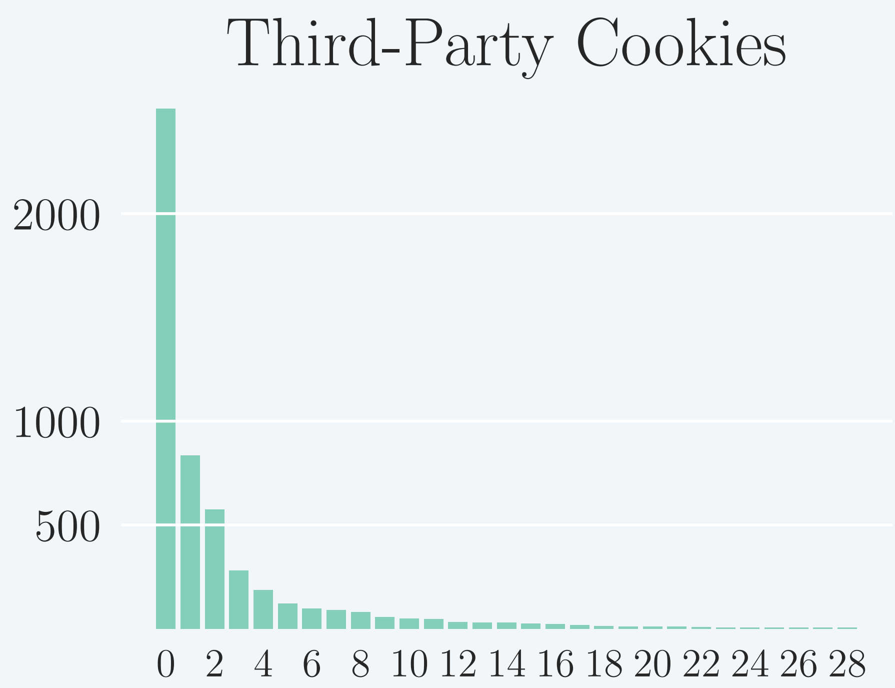
"Fog" readability index: Years of formal education needed to read a document

Gunning, 1952



OpenWPM scraper counts cookies on firm's website

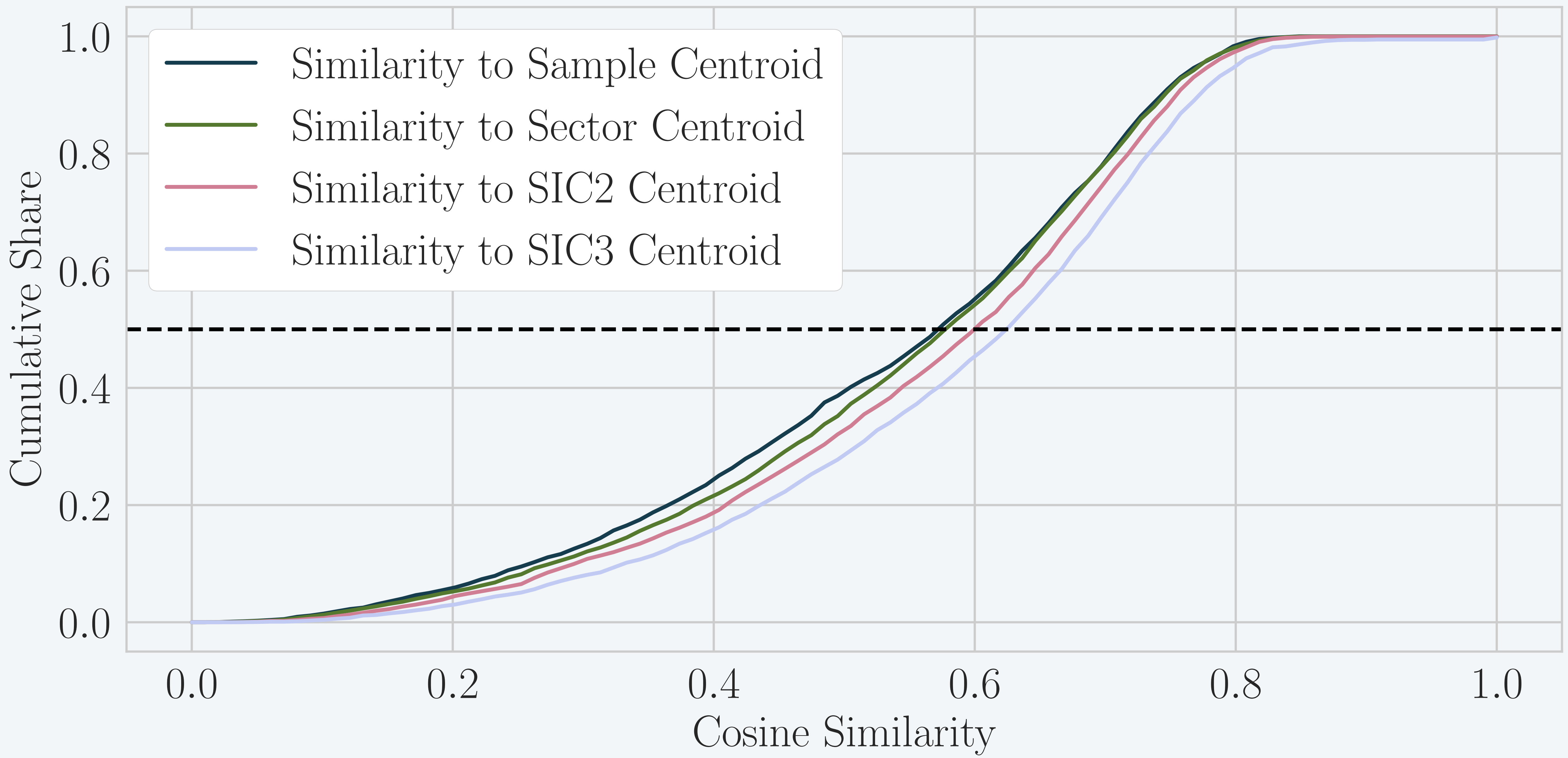
Englehardt and Narayanan, 2016



Stylized Facts

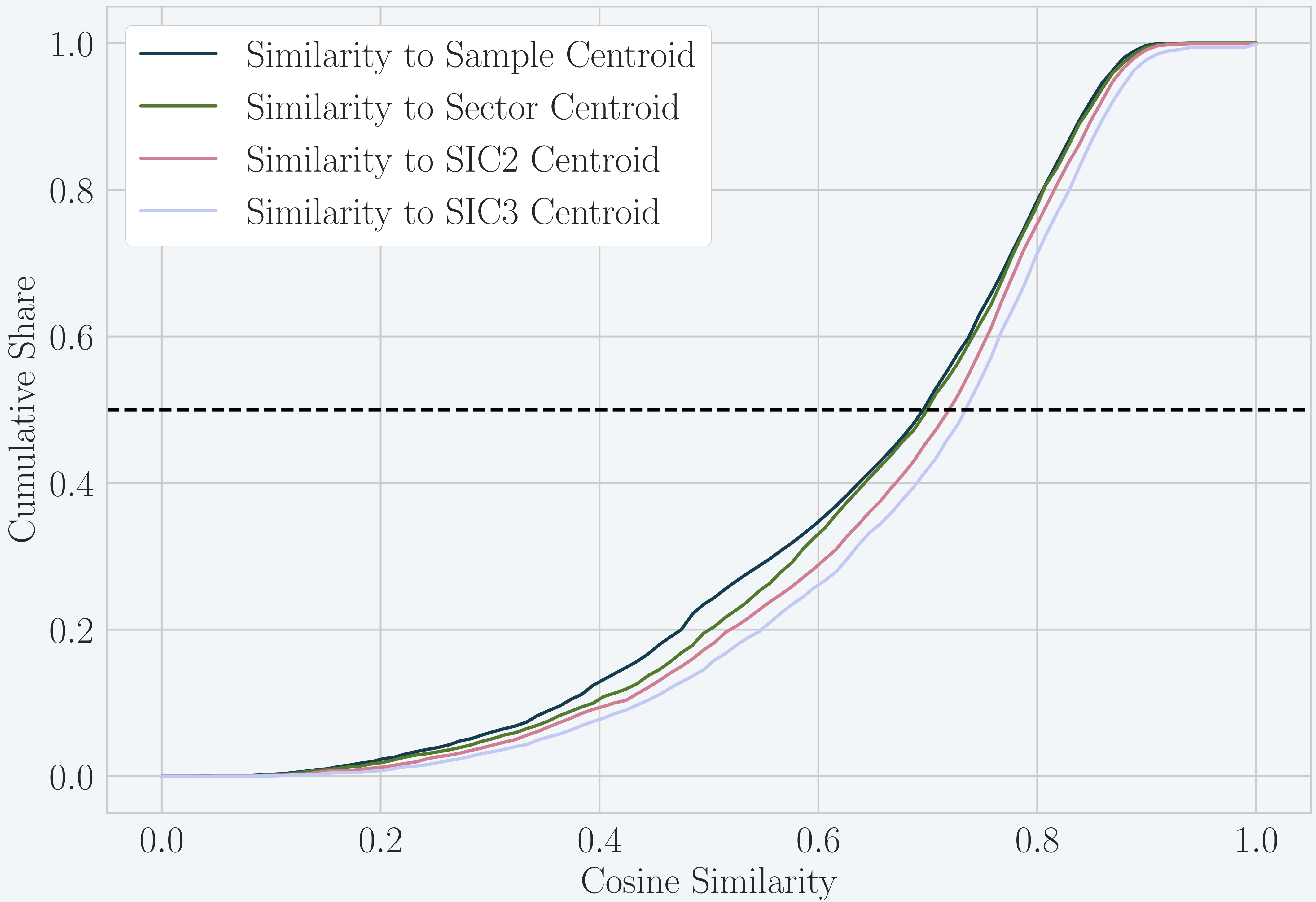
Variation: No Industry Boilerplates

Similarity of Word Frequency Vectors Across Policies

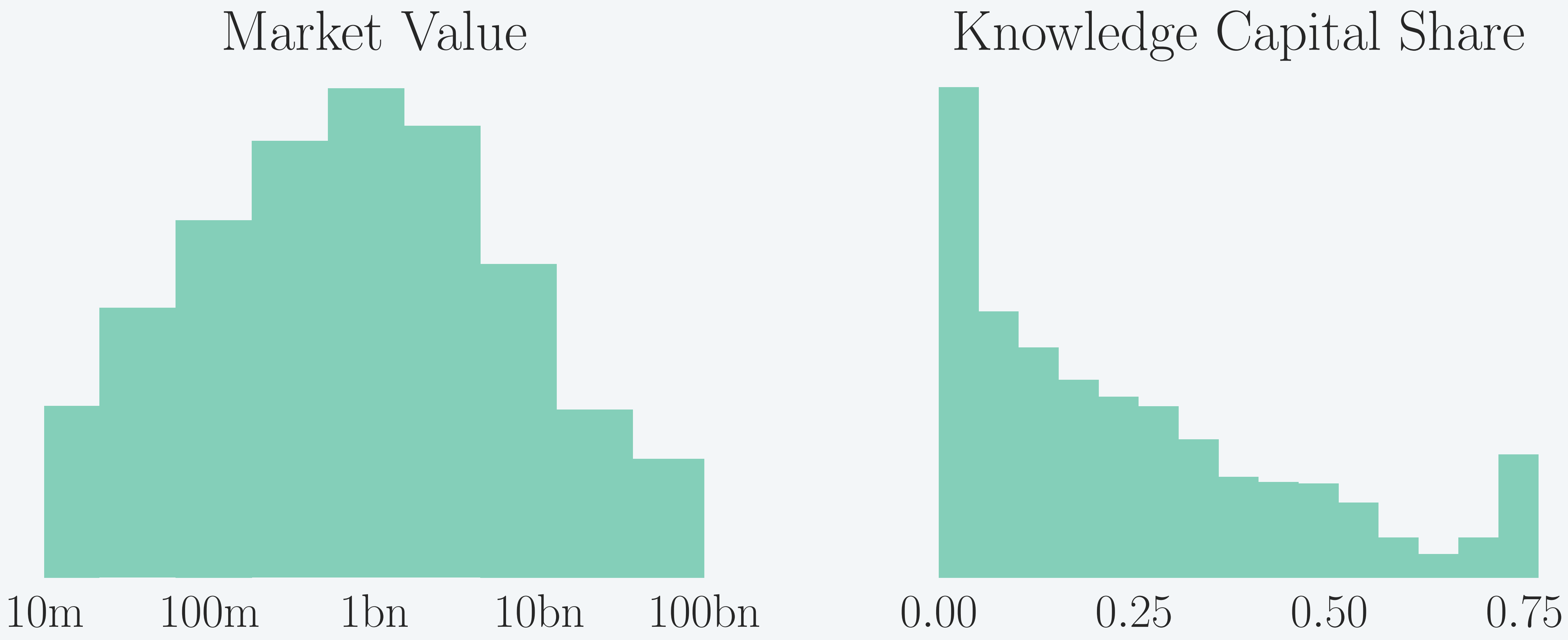


Variation: No Industry Boilerplates

Latent Semantic Analysis with 250 topics



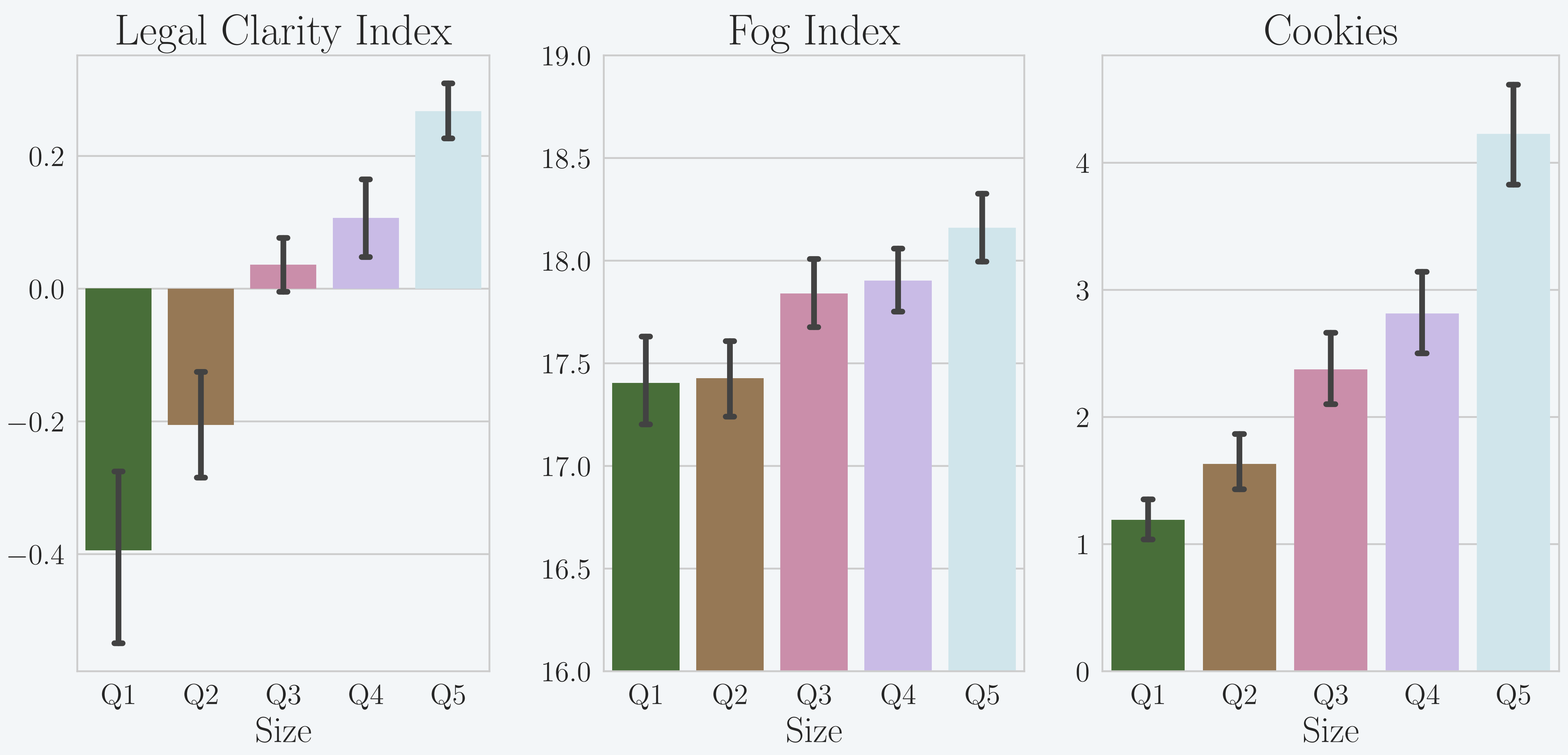
Firm Characteristics



$$\text{Knowledge Share} = \frac{\text{Capital accumulated through R\&D}}{\text{Total Assets}}$$

Peters and Taylor, 2017

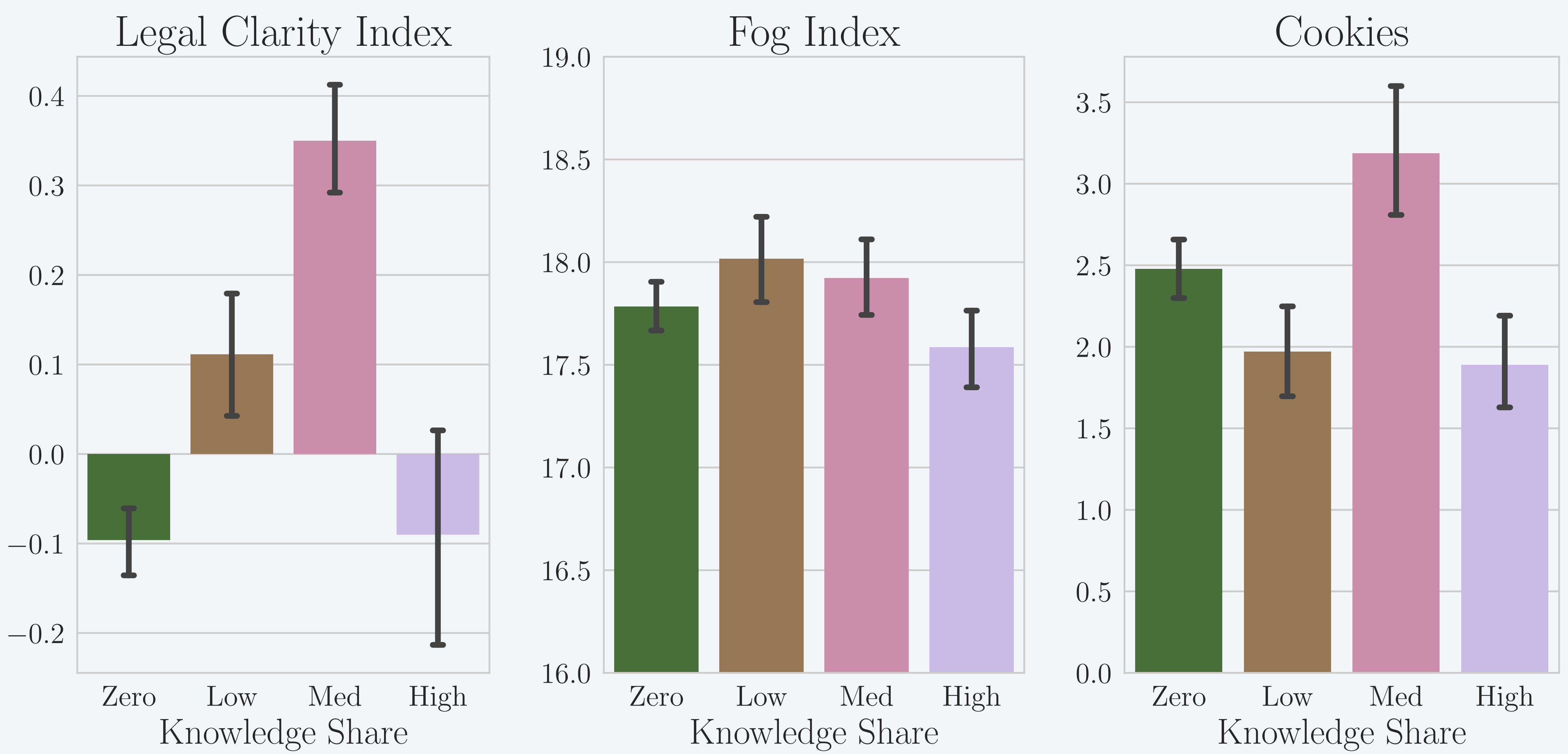
Firm Size, Policies and Behavior



Large firms also have longer policies which are easier to find

Knowledge Share, Policies and Behavior

Capital Accumulated through R&D / Total Assets



Firm Characteristics

Theory of Data Sharing

Theory of Data Sharing

Firm has data about its consumers, input into production of signals (e.g., about consumers' preferences)

Firm choices: Discard the data, process in house (signal x, cost ϕ), or share with specialized intermediary (better signal y, cost 0)

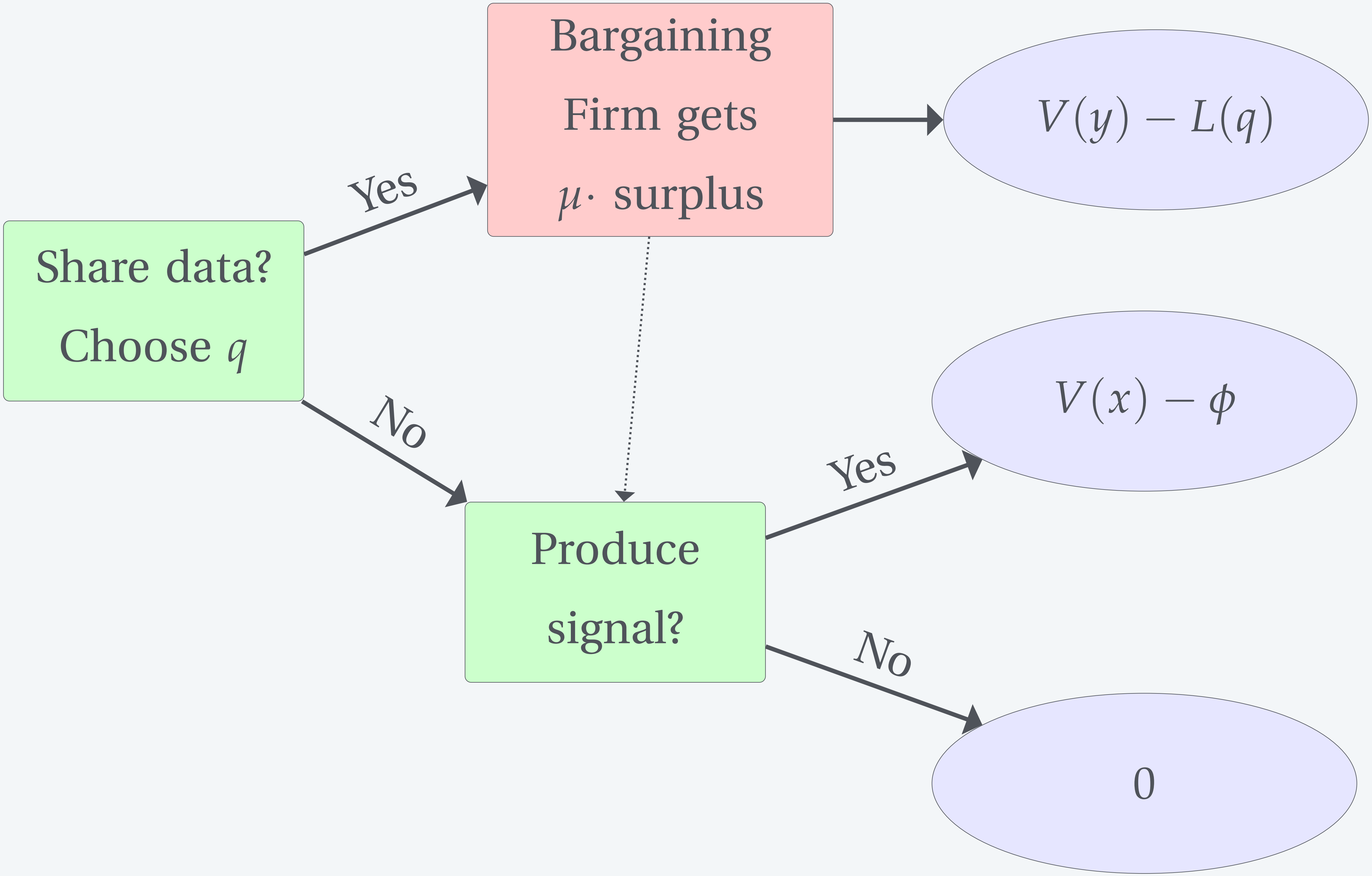
- Cost of data sharing: Future litigation risk $L(q)$, where q is clarity of legal text, cost $\kappa(q)$

Data valuation: Maximized value $V(x)$ from selling or using signal x

Horner-Skrzypacz, 2016; Bergemann and Bonatti, 2018

$$V(y) \geq V(x) > 0$$

Timing


Bargaining
Firm gets
$\mu\cdot$ surplus
$V(y) - L(q)$
Yes
Share data?
Choose q
No
$V(x) - \phi$
Yes
Produce
signal?
No
0

Data Sharing Condition

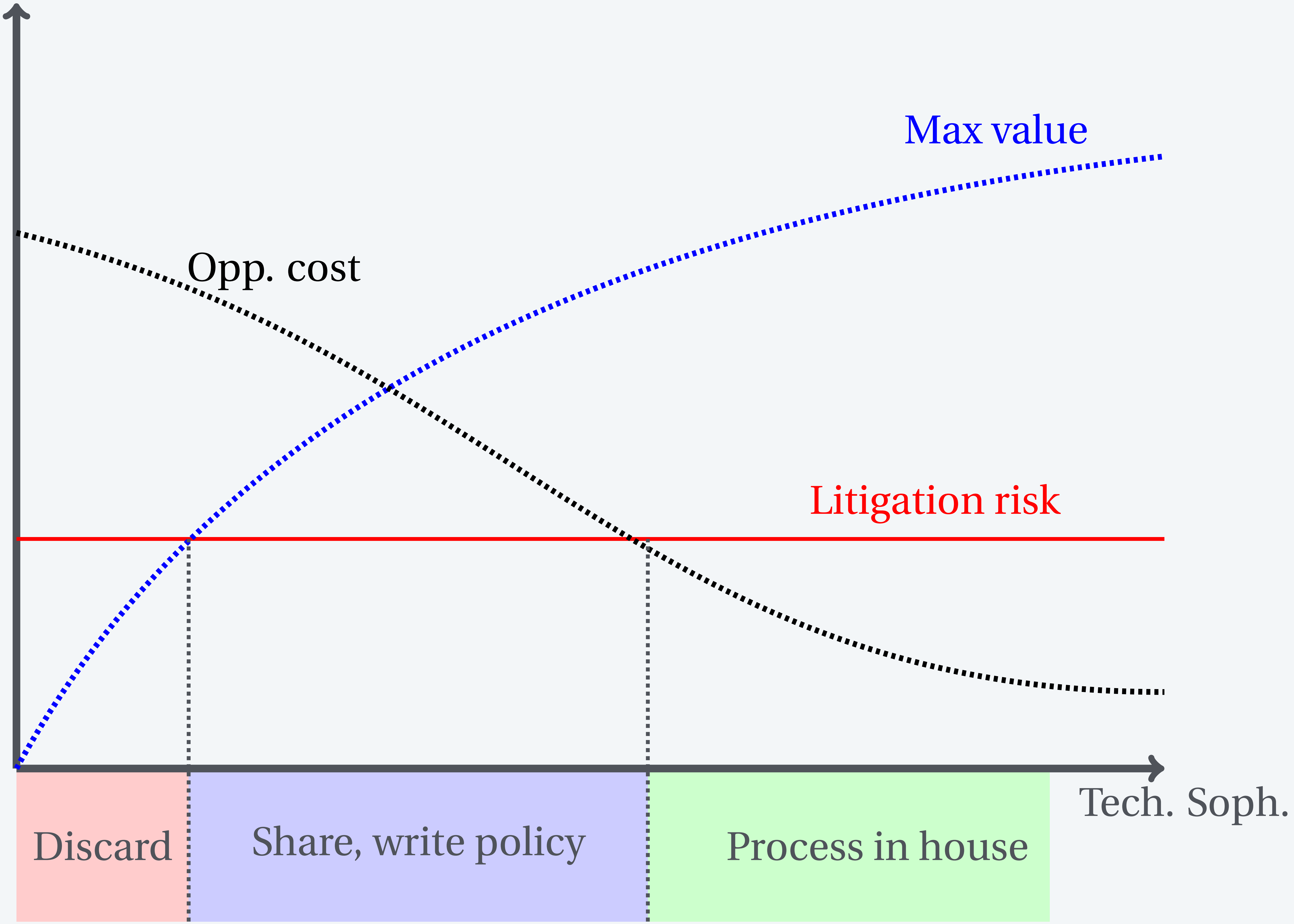
Opportunity cost of in house processing:

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

Proposition: *The firm shares data if and only if*

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

where $M \equiv \frac{\mu L(q^\star)+\kappa(q^\star)}{\mu}$ *is cost-benefit trade-off in data sharing*


Max value
Opp. cost
Litigation risk
Tech. Soph.
Discard
Share, write policy
Process in house

Partial Effects of Knowledge Capital

Controlling for firm size, market share, industry FE:

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

Conclusions

- We assemble comprehensive data for studying the market for privacy, focusing on the supply side
- Stylized facts on cross-firm variation
  - Clear policies $\Rightarrow$ more sharing
- Simple testable theory of data sharing

Public Resources

www.github.com/ansgarw/privacy

- All our data for work with Compustat US firms
- Python code, demos and documentation
- Get policies and their attributes for *any* sample of firms or websites

Simplest Example

Here are 5 lines of code that find the policy for American Airlines:

```
from src.urls import crawlPrivacy, filterPrivacy
from src.text import findPolicy
status, urls = crawlPrivacy('www.aa.com',clicks=2) # crawls candidate URLs
ranked = filterPrivacy(sum(urls,[])) # filter and rank by likelihood of being privacy policy
status, policy, url = findPolicy(ranked) # scrape highest ranked page that contains 'privacy'
```


```
Legal clarity of www.aa.com: 1.136
Legal clarity of www.ba.com: 1.691
```

Imperial College
Business School

Imperial means
Intelligent Business