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

- **Second best:** Concerns about insurance, price discrimination, negative externalities
  Hirshleifer, 1971; Taylor, 2004; Varian, 2009

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

\[ N = 5377 \text{ firms in Compustat US} \]

Finding privacy policies:
- Automated google search
- Web crawling
- Manual checking

Visibility: "Privacy" link on website

Access and Visibility

<table>
<thead>
<tr>
<th>Total</th>
<th>Found</th>
<th>Visible</th>
</tr>
</thead>
<tbody>
<tr>
<td>5377</td>
<td>4078</td>
<td>3479</td>
</tr>
</tbody>
</table>
Text Analysis Terminology

Represent policies as vectors in a term-document matrix

<table>
<thead>
<tr>
<th></th>
<th>personal information</th>
<th>third party</th>
<th>personal data</th>
<th>privacy policy</th>
<th>web site</th>
<th>personally identifiable</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.aa.com">www.aa.com</a></td>
<td>22.0</td>
<td>19.7</td>
<td>26.3</td>
<td>11.3</td>
<td>0.0</td>
<td>0.9</td>
</tr>
<tr>
<td><a href="http://www.cecoenviro.com">www.cecoenviro.com</a></td>
<td>0.7</td>
<td>10.8</td>
<td>68.6</td>
<td>10.9</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><a href="http://www.asaltd.com">www.asaltd.com</a></td>
<td>15.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><a href="http://www.pinnaclewest.com">www.pinnaclewest.com</a></td>
<td>0.0</td>
<td>3.5</td>
<td>0.0</td>
<td>7.6</td>
<td>3.2</td>
<td>5.7</td>
</tr>
<tr>
<td><a href="http://www.aarons.com">www.aarons.com</a></td>
<td>0.0</td>
<td>29.8</td>
<td>0.0</td>
<td>2.8</td>
<td>0.0</td>
<td>4.3</td>
</tr>
</tbody>
</table>

- **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:

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

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

![Graph showing cumulative share of similarity to different centroids against cosine similarity]

- Similarity to Sample Centroid
- Similarity to Sector Centroid
- Similarity to SIC2 Centroid
- Similarity to SIC3 Centroid

Imperial College Business School

Imperial means Intelligent Business
Variation: No Industry Boilerplates

Latent Semantic Analysis with 250 topics

Cumulative Share

Similarity to Sample Centroid
Similarity to Sector Centroid
Similarity to SIC2 Centroid
Similarity to SIC3 Centroid

Cosine Similarity

Imperial College Business School

Imperial means Intelligent Business
Firm Characteristics

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

Legal Clarity Index

Fog Index

Cookies

Firm Characteristics

Imperial College Business School

Imperial means Intelligent Business
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 $\phi$), 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

Share data? Choose $q$

Bargaining Firm gets $\mu \cdot$ surplus

$V(y) - L(q)$

$V(x) - \phi$

0

Imperial College Business School

Imperial means Intelligent Business
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 \( M \equiv \frac{\mu L(q^*) + \kappa(q^*)}{\mu} \) is cost-benefit trade-off in data sharing.
Partial Effects of Knowledge Capital

Controlling for firm size, market share, industry FE:

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

- We assemble comprehensive data for studying the market for privacy, focusing on the supply side

- Stylized facts on cross-firm variation
  - Clear policies ⇒ 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:

```python
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
```