Discussion of Demirer, Jiménez-Hernández, Li, and Peng (2023):

Data, Privacy Laws and Firm Production: Evidence from the GDPR

Yizhou Jin

UToronto

AEA
Jan 2024

Paper Summary

Strategy

- Anonymized data of *all* customers of a large global cloud service provider
- Incorporate data and computation into production function estimation by modeling latent "information" formation
- Event study of GDPR implementation in 2018

Results

- data and computation are complements, esp. for services industries
- variation in *computation* productivity is a key driver for information differences across firms
- direct evaluation of GDPR impact
  - event study on quantity: data storage -26%; computation -15%;
  - GDPR "wedge": estimated MC of cloud data storage +20%
  - cost of information +4% (cost of production +0.5%): data is far cheaper than computation

Unprecedented Data Access

- Firms (proprietary + Aberdeen, 60% match)
  - *all* customers (focus on stable cloud users pre-GDPR)
  - SIC industry codes, HQ location, start-up status
  - single- or multi-cloud
  - employment (80% European firms)
- Cloud Usage
  - storage in Gb and computation in core-hours
  - price + $ spend in high (monthly) frequency (although can't publish price stats)
  - broken down by firm HQ and *servers location*
    - control group: US-HQ firms with only US servers
    - treatment group: All EU-HQ firms + US firms with EU servers

Core Insights

1. Information Production Model + Cost Minimization Approach
   - Focus on information (I) production, which mediates how data (D) and computation (C) affect output $\implies$ we can study the cost and production of I regardless of how it enters the (output) production function or other distortions such as market power!
   - Elasticity of substitution between C and D (σ) + productivity of computation (ω^C) pins down (and can be identified from) the elasticity of D/C ("data intensity").

Core Insights

1. Information Production Model + Cost Minimization Approach
   - Focus on information (I) production, which mediates how data (D) and computation (C) affect output $\implies$ we can study the cost and production of I regardless of how it enters the (output) production function or other distortions such as market power!
   - Elasticity of substitution between C and D (σ) + productivity of computation (ω^C) pins down (and can be identified from) the elasticity of D/C ("data intensity").
   - Link to the production function literature
     - Olley and Pakes (1996): we can "invert" out firm-level ω from their investment and exit choices, addressing the simultaneity issue with productivity shocks
     - Levinsohn and Petrin (2003): use input choices instead of investment/exit since they are nonzero and less lumpy
     - Ackerberg, Caves, and Frazer (2006): structural models of input choices introduce collinearity.
     - Gandhi, Navarro, and Rivers (2013): a cost minimization approach can avoid structurally modeling input choices. With **Hicks neutral** productivity shocks, input demand elasticity pins down shares (σ and ω are reparameterization of the demand elasticities of D and C)

Core Insights

1. Information Production Model + Cost Minimization Approach
   - ▷ Focus on information (I) production, which mediates how data (D) and computation (C) affect output $\Longrightarrow$ we can study the cost and production of I regardless of how it enters the (output) production function or other distortions such as market power!
   - ▷ Elasticity of substitution between C and D (σ) + productivity of computation (ω^C) pins down (and can be identified from) the elasticity of D/C ("data intensity").

2. Shift-share IV to Get Input Demand Elasticity
   - ▷ A key challenge in B2B/enterprise sale: even listed prices are negotiated
   - ▷ Shift-share IV exploiting very sticky lagged data center locations
   - ▷ But the results are similar to OLS $\rightarrow$ are pricing and negotiation actually strategic?

Comment 1: Rationality / Interpretation of the "Wedge"

- Firms are far from rational about their cloud usage
  - Self-report cloud waste averaged 32% of cloud budgets (Flexera 2021)
  - Most companies do not understand how to measure cloud ROI (PwC 2021)
  - Employees frequently misused company resources (imperfect monitoring/bitcoin mining)

Comment 1: Rationality / Interpretation of the "Wedge"

- Firms are far from rational about their cloud usage
  - Self-report cloud waste averaged 32% of cloud budgets (Flexera 2021)
  - Most companies do not understand how to measure cloud ROI (PwC 2021)
  - Employees frequently misused company resources (imperfect monitoring/bitcoin mining)
- A key impact of GDPR is to trigger awareness and audit of cloud usage
  - demand elasticity $\uparrow$, and probably differentially so since D cost much more than C.
  - the post-GDPR demand elasticities might better reflect cost minimization
  - relatedly, how should we think about GDPR as potentially triggering non-Hicks neutral shock?

Comment 1: Rationality / Interpretation of the "Wedge"

- Firms are far from rational about their cloud usage
  - Self-report cloud waste averaged 32% of cloud budgets (Flexera 2021)
  - Most companies do not understand how to measure cloud ROI (PwC 2021)
  - Employees frequently misused company resources (imperfect monitoring/bitcoin mining)
- A key impact of GDPR is to trigger awareness and audit of cloud usage
  - demand elasticity $\uparrow$, and probably differentially so since D cost much more than C.
  - the post-GDPR demand elasticities might better reflect cost minimization
  - relatedly, how should we think about GDPR as potentially triggering non-Hicks neutral shock?
- Recommendation
  - Regulatory "wedge" is OK, but perhaps de-emphasize privacy/misallocation?
  - Provide aggregate measures of actual computation usage in event studies (e.g. data center operation costs/energy consumed)
  - Follow-on papers examining subsequent privacy regulations / enforcement actions

Comment 2: Selection and Dynamics

- A key issue for production function estimation is selection $\rightarrow$ censored ω distribution
  - labor and capital are always positive – not necessarily for D and C.
  - extensive margin #1: $D = 0$ post GDPR (even among historically stable cloud users)
  - "extensive" margin #2: implementation delays post GDPR
- The dynamics are a bit hard to understand
  - analysis cut off at COVID with no slowdown of D and C reduction...
  - ...but we saw tapering on extensive margin #1 and for $\frac{D}{C}$, why?
- Recommendation
  - Quantile FE regressions to better understand compositional vs dynamic effects
  - Discuss more about dynamics and what you think the "longer-term" effect would be
  - Consider extensive margin more formally (I don't think dynamics is needed, just selection)

Overall

1. This is one of the most important recent papers in digitization and IO
   - ▷ Both the measurements and the empirical modeling of information production are sorely needed to drive consensus
   - ▷ One of the few cases in which cross-industry analysis is actually appropriate
   - ▷ Among modern production inputs, D and especially C are as important as they are opaque
   - ▷ Pervasive descriptive evidence of huge markups $\rightarrow$ really need to understand input demand
2. The structural model
   - ▷ Of course there are many simplifications and debatable assumptions
   - ▷ But beyond the model and GDPR estimates, it captures much richer and higher-dimensional information about the proprietary datasets than RF results
   - ▷ e.g. can we calibrate markups given realistic assumptions of scale economy and switching costs for cloud services?