Data as a networked asset

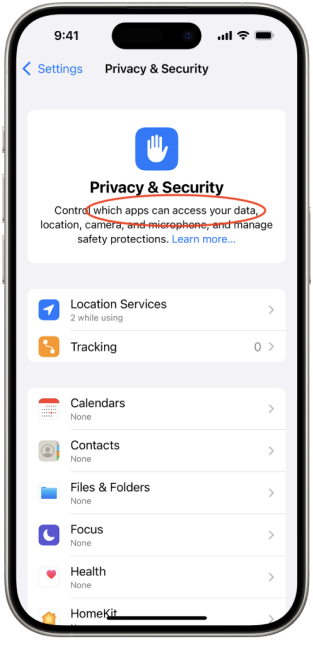
Bo Bian (UBC), Qiushi Huang (SAIF), Ye Li (Foster) and Huang Tang (Wharton)

Discussion by Nicolas Crouzet (Kellogg)

Duke/UNC corporate finance conference, Spring 2025

What is this paper about?

Question: Does consumer data sharing across firms impact their performance?


9:41
Settings
Privacy & Security
Privacy & Security
Control which apps can access your data, location, camera, and microphone, and manage safety protections. Learn more...
Location Services
2 while using
Tracking
0
Calendars
None
Contacts
None
Files & Folders
None
Focus
None
Health
None
HomeKit

What is this paper about?

Question: Does consumer data sharing across firms impact their performance?

Do certain firms play an outsize role in data sharing?

Are firm-specific shocks propagated through data sharing linkages?

(Above and beyond supply chain or financial linkages)

What is this paper about?

Question: Does consumer data sharing across firms impact their performance?

Do certain firms play an outsize role in data sharing?

Are firm-specific shocks propagated through data sharing linkages?

(Above and beyond supply chain or financial linkages)

Answer: Yes.

Empirics: measure of pairwise "data connectedness" between firms

Higher connectedness is associated with higher comovement in financial and real outcomes

Restrictions in data sharing reduces comovement

Model: Q-theory model investment in data subject to positive network externalities

Quantify which firms contribute most to the overall value of data assets

What's special about this paper?

This paper uses a dataset on mobile apps to construct its measure of "data connectedness"

Mobile apps are built using SDKs

SDK = "Software Development Kit"

- Open source
- Created by Google, Meta, Apple, Microsoft ...
- They are (I think?) building tools (or blocks) for apps
- But I am not a software engineer! (Maybe the paper should add one as co-author.)

The paper's data records which SDKs are embedded in each app.

This data source is really new (at least to me). And super interesting.

Roadmap

1. What are we trying to measure?
2. What are we actually measuring?
3. How does the model relate to the measurement?

1. What are we trying to measure?

An example with three firms: Google, GM, Amazon

An example with three firms: Google, GM, Amazon

Google: designs and makes an SDK available for interacting with Google Ads

Building block in GM and Amazon's apps

Collects and standardizes user data

An example with three firms: Google, GM, Amazon

Google: designs and makes an SDK available for interacting with Google Ads

Building block in GM and Amazon's apps

Collects and standardizes user data

GM: collects customer data on driving habits, car accessories, financing through its app

Shares it with Google through the SDK

Google aggregates it with other data on auto purchases, and shares that with Amazon

Amazon uses this data to improve ad targeting for e.g. car accessories

An example with three firms: Google, GM, Amazon

Google: designs and makes an SDK available for interacting with Google Ads

Building block in GM and Amazon's apps

Collects and standardizes user data

GM: collects customer data on driving habits, car accessories, financing through its app

Shares it with Google through the SDK

Google aggregates it with other data on auto purchases, and shares that with Amazon

Amazon uses this data to improve ad targeting for e.g. car accessories

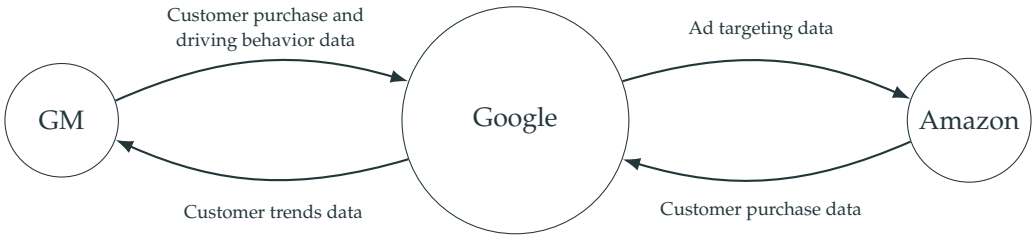
Amazon: collects customer data on purchasing habits

Shares it with Google through the SDK

Google aggregates it with other data on car accessory purchases and shares it with GM

GM uses this data to adapt product design

An example with three firms: Google, GM, Amazon



A lazy attempt at formalism

GM's (and Amazon's) revenue functions

$$\begin{pmatrix} \overbrace{y_{i,t}}^{\#\text{units sold}} \\ \\ \underbrace{o_{i,t}}_{\text{Consumer data gathered}} \end{pmatrix} = F\Big(\overbrace{k_{i,t}}^{\text{Physical K}} \;,\; \underbrace{b_{i,t}}_{\text{Intangible K}} \;,\; \overbrace{d_{i,t}}^{\text{Consumer data used}} \Big)$$

Google's revenue function

$$\forall i \text{ using Google's SDK}, \quad d_{i,t} = G\left(k_{g,t} \;,\; b_{g,t} \;,\; \overbrace{H_i\big(\{o_{j,t}\}_{j \neq i} \big)}^{\text{Data from other firms}} \right)$$

What we would like to measure: $o_{i,t}$ (data produced by firm i); $d_{i,t}$ (data used by firm i)

[sidenote: units of s and d? In this paper, average monthly users]

Where is the network in this lazy example?

Core-periphery (or intermediation) structure

Google = core

GM, Amazon = periphery

The distinction is probably more fuzzy in reality ...

Google's data aggregation technology

$$H_i\left(\left\{o_{j,t}\right\}_{j\neq i}\right) = \prod_{j\neq i} o_{j,t}^{\gamma_{i,j}}$$

$\gamma_{i,j}$ = "Data connectedness" between firm i and firm j

("How much" firm i relies on the data produced by firm j)

Maybe we would like to measure $\left\{\gamma_{i,j}\right\}$ as well!

2. What are we actually measuring?

Measuring "data connectedness"

1. Firm i, app a, data-sharing SDK k, quarter t

$$
\begin{aligned}
m_{i,a,t} &= \text{average number of daily active users on app } (a,i) \\
d_{i,a,k,t} &= \mathbf{1}\{\text{app } (a,i) \text{ uses data-sharing SDK } k\} \\
s_{i,a,k,t} &\equiv m_{i,a,t} \times d_{i,a,k,t} = \text{"amount of data" shared by firm } i \text{ through SDK } k \text{ embedded in app } a
\end{aligned}
$$

2. Firm i, SDK k, quarter t

$$
\begin{aligned}
S_{i,k,t} &\equiv \sum_a s_{i,a,k,t} = \text{"amount of data" shared by firm } i \text{ through SDK } k \\
S_{i,t} &\equiv (S_{i,1,t}, ..., S_{i,K,t})' \qquad [K \times 1]
\end{aligned}
$$

3. "Data connectedness" between firm i and j in quarter t

$$
\rho_{i,j,t} \equiv \text{cosine similarity } (S_{i,t}, S_{j,t})
$$

Interpreting the measure

Are all data-sharing SDKs the same?

$$s_{i,a,k,t} \equiv m_{i,a,t} \times d_{i,a,k,t} \times (\text{"how much data" SDK } k \text{ extracts from each customer})$$

[e.g. a payments SDK might share less data than an consumer demographics SDK]

Is this measuring "how much data" firm i shares with firm j?

No.

It's measuring a (user-weighted) overlap in data-sharing functionalities between apps of firm i and firm j.

Answers the question: do firm i and firm j use **similar software** to collect, analyze and share data?

Does it matter for the paper?

Maybe.

At least: don't interpret the vertices on the paper's network graphs as bilateral data flows.

An example where the interpretation of the measure matters

Regression on effects of cyberattacks:

$$Y_{i,t} = \alpha + \beta \mathbf{1}\{\text{cyberattack to firm } i_0 \text{ at time } k \leq t\} \times (\text{data connectedness } i \to i_0) + \text{f.e.} + \varepsilon_{i,t}$$

$\beta > 0$ interpreted as spillovers/network effects.

But subject to "reflection problem": common exposure to shock might be driving $\beta > 0$.

In this case:

data connectedness $\approx$ software similarity

software similarity creates common exposure to cyberattacks

Could we measure things differently?

Certain firms — GM, Expedia, Capital One — are "**end-users**" of data

- Use consumer data to improve operational performance (i.e., sell more cars)
- Data is obtained from own apps — but also from third parties, who can process and augment it

Other firms — Meta, Google — operate as "**data intermediaries**"

- Collect and analyze data from "end-users" (i.e. GM)
- Connect "end-users" together (i.e. third-party websites and GM for advertising)

Can the data be used to explore this distinction?

- Study the contents of the SDKs? (i.e, the code!)
- Which **"data intermediaries"** does each SDK typically route the data to?
- Focus on links between core ("**data intermediaries**") and periphery ("**end-users**")

3. How does the model relate to the measurement?

Key components of the model

Q-theory model (with capital = customer data), with two main twists

1. Data has positive externalities across firms

   When GM produces more data, it benefits Amazon as well

2. Firms face an intertemporal trade-off btw.

   "Investing in data" $\rightarrow$ increases profits in the future

   "Monetizing the customer base" $\rightarrow$ increases profits today

Twist 1: externalities

$$\text{Profits}_{i,t} = \overbrace{\delta_{i,t}}^{\text{Profits from existing data stock}} + F_{i,t}$$

$$\frac{d\delta_{i,t}}{\delta_{i,t}} = \alpha \sum_j \gamma_{i,j} \frac{\delta_{j,t}}{\delta_{i,t}} + \underbrace{x_{i,t}}_{\text{Investment}} - \mu_\delta$$

- α determines the overall strength of externalities

  $\alpha = 0$: standard AK model

- $\gamma_{i,j}$ represents pairwise "data connectedness"

  $\Gamma = (\gamma_{i,j})$ represents network structure; analog to the $(\rho_{i,j,t})$ network from the empirics

Twist 1: externalities

$$\text{Profits}_{i,t} = \overbrace{\delta_{i,t}}^{\text{Profits from existing data stock}} + F_{i,t}$$

$$\frac{d\delta_{i,t}}{\delta_{i,t}} = \alpha \sum_j \gamma_{i,j} \frac{\delta_{j,t}}{\delta_{i,t}} + \underbrace{x_{i,t}}_{\text{Investment}} - \mu_\delta$$

- α determines the overall strength of externalities
  - $\alpha = 0$: standard AK model
- $\gamma_{i,j}$ represents pairwise "data connectedness"
  - $\Gamma = (\gamma_{i,j})$ represents network structure; analog to the $(\rho_{i,j,t})$ network from the empirics
- No "data intermediaries"; only bilateral exchanges w/ some more central firms
- Other firms' data production is always good for my own profits

Twist 2: intertemporal trade-off

$$\text{Profits}_{i,t} = \delta_{i,t} + \overbrace{F_{i,t}}^{\substack{\text{Profits from} \\ \text{monetizing the customer base}}}$$

$$F_{i,t} = F\left(\overbrace{\sum_j \gamma_{i,j}\delta_{j,t}}^{(+)} , \underbrace{x_{i,t}}_{(-)} \right)$$

- Narrative:

  Firm can acquire paying customers today

  Harder to do so when investing more in data ($x_{i,t}$)

Twist 2: intertemporal trade-off

$$\text{Profits}_{i,t} \quad = \quad \delta_{i,t} \quad + \quad \overbrace{F_{i,t}}^{\substack{\text{Profits from} \\ \text{monetizing the customer base}}}$$

$$F_{i,t} = F\left(\overbrace{\sum_j \gamma_{i,j}\delta_{j,t}}^{(+)} \ , \ \underbrace{x_{i,t}}_{(-)} \right)$$

- Narrative:

  Firm can acquire paying customers today

  Harder to do so when investing more in data $(x_{i,t})$

- Why do this? Because it leads to a standard Q-theory condition of the type:

$$-F_x \ = \ V_{\delta_i}$$

Twist 2: intertemporal trade-off

$$\text{Profits}_{i,t} \quad = \quad \delta_{i,t} \quad + \quad \overbrace{F_{i,t}}^{\substack{\text{Profits from} \\ \text{monetizing the customer base}}}$$

$$F_{i,t} = F\left(\overbrace{\sum_j \gamma_{i,j}\delta_{j,t}}^{(+)} \ , \ \underbrace{x_{i,t}}_{(-)} \right)$$

- Narrative:

  Firm can acquire paying customers today

  Harder to do so when investing more in data $(x_{i,t})$

- Why do this? Because it leads to a standard Q-theory condition of the type:

$$-F_x \ = \ V_{\delta_i}$$

- But what is the economic mechanism?

  Why do firms trade-off learning about their customers vs. monetizing them?

Implications for valuations

In general,

$$V_{i,t} = \kappa_i + \left\{ \left(\left((\rho + \mu_\delta) I - \beta \Gamma \right) \delta_t \right)^{-1} \right\}_i$$

When there are no externalities ($\Gamma = I$),

$$V_{i,t} = \kappa_i + \frac{\delta_{i,t}}{\rho + \mu_\delta - \beta}$$

- Paper: measures of the contribution of each firm to overall stock market value

  Can be $>$ firm's own market value, because of externalities

  Depends on degree of centrality

- Another simple application — valuation of data assets in the presence of externalities

  Enterprise value = PV of cash flows from firm's own data + PV of externalities from other firms' data

  Can this help rationalize $Q >> 1$ for data-intensive firms?

Conclusion

Conclusion

Very exciting paper, particularly on the data front

Conclusion

Very exciting paper, particularly on the data front

Main suggestion

Does data connectedness reflect data flows or software similarity?

Conclusion

Very exciting paper, particularly on the data front

Main suggestion

Does data connectedness reflect data flows or software similarity?

Research going forward

Is a “data intermediation” model a more useful way to think about this market?