The Economics of Social Data

Dirk Bergemann[1] Alessandro Bonatti[2] Tan Gan[1]

[1]Yale University

[2]MIT Sloan

Privacy Workshop
Princeton University
February 2020

Information and Data

- rise of large internet platforms, Amazon, Facebook, Google, and JD, Tencent, Alibaba, leads to unprecedented collection of individual user data
- information markets central to economic activity, \$20b to acquire/process consumer data (IAB 2018)
- selling information $\longrightarrow$ providing access to data
- consumer scores, predictions, ratings, recommendations, customized products and services

Individual and Social Data

- individual-level data allows companies to refine search results, personalize product recommendations, informative ratings, timely traffic data, targeted advertising
- central feature of individual data is its social aspect
- data captured from an individual user is informative about users similar to the individual, thus it is social data!
- social nature of data generates data externality

Objectives and Challenges

consumer data must be acquired, aggregated, packaged, and sold.

- who buys consumers' information in equilibrium? does the market enable an efficient use of individual information?

"social" dimension of the data: data about an individual consumer is informative about *similar* consumers.

- how does the social dimension of the data impact the terms of trade between consumers, data buyers, and data intermediaries?
- what determines the value of individual and aggregate data for an information intermediary?

Basic Model

- a data broker, N consumers, and a producer (merchant)
- each consumer has willingness-to-pay

$$w_i = \theta + \theta_i$$

- common and idiosyncratic demand shocks, θ and θ_i:

$$\begin{pmatrix} \theta \\ \theta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & 0 \\ 0 & \sigma_{\theta_i}^2 \end{pmatrix}\right)$$

and consumer i chooses quantity q_i

$$u\left(w_i, q_i\right) = w_i q_i - \frac{1}{2} q_i^2 - p_i q_i$$

- producer maximize revenues $p = (p_1, ..., p_N)$

$$\pi\left(p\right) = \mathbb{E} \sum_i \left(p_i - c\right) q_i.$$

Data Trade

- data broker buys data from individuals and sells to producer
- bilateral contracting
- data broker collects linear differentially private signal of w_i

$$s_i = \sum\nolimits_j \alpha_{ij} \left(w_j + \varepsilon + \varepsilon_j \right) a_j,$$

with common and idiosyncratic shock, ε and ε_j
weight $\alpha_{ij} \in \mathbb{R}$ prescribes influence data of j has on $\mathbb{E}\left[w_i \,|s_i \right]$

matched: $\alpha_{ij} = \mathbb{I}_{i=j}$; anonymized: $\alpha_{ij} = 1/N$

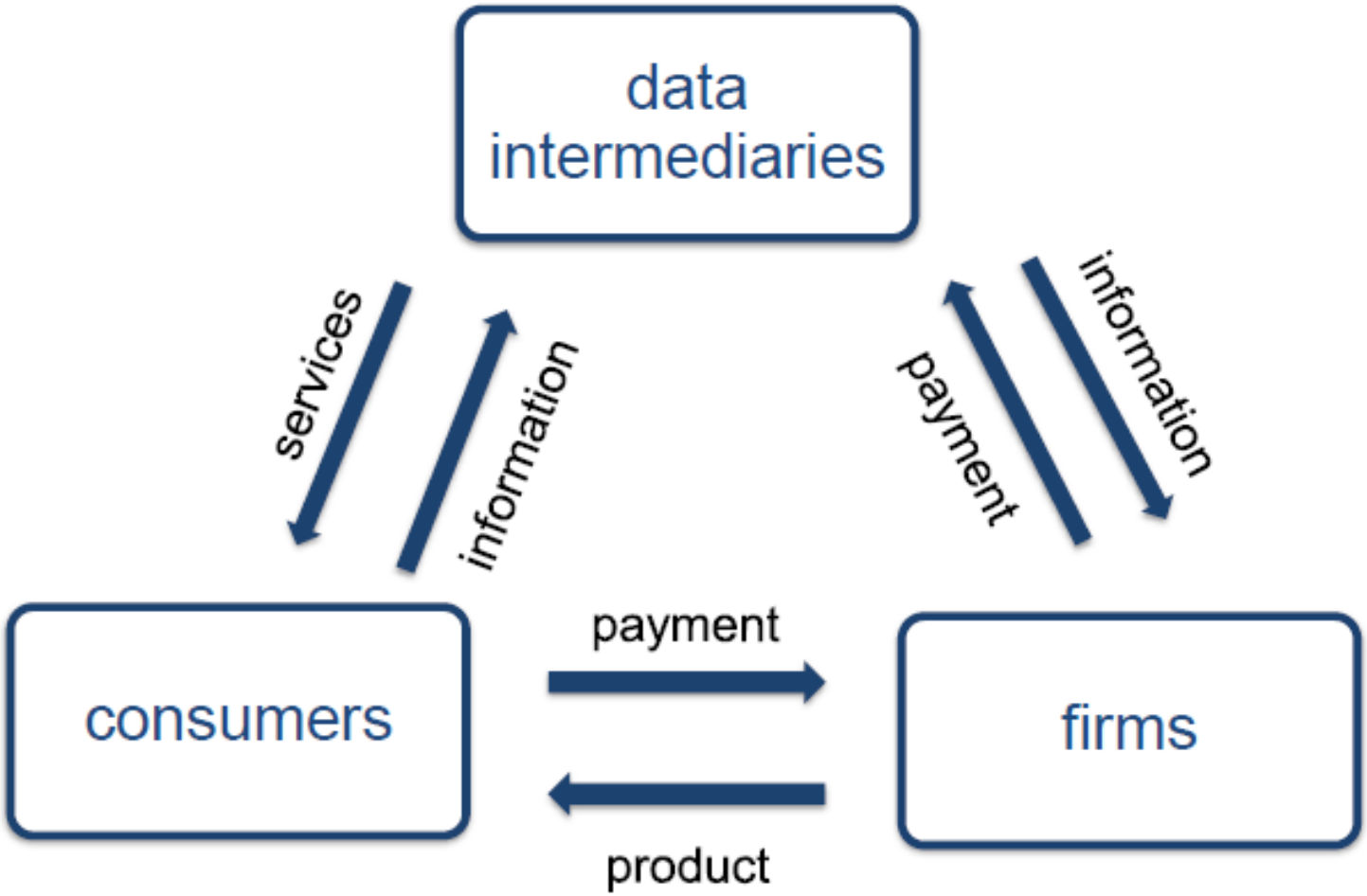
and $a_j \in \{0, 1\}$ identifies participation of consumer j

$$a_j \in \{0, 1\}$$

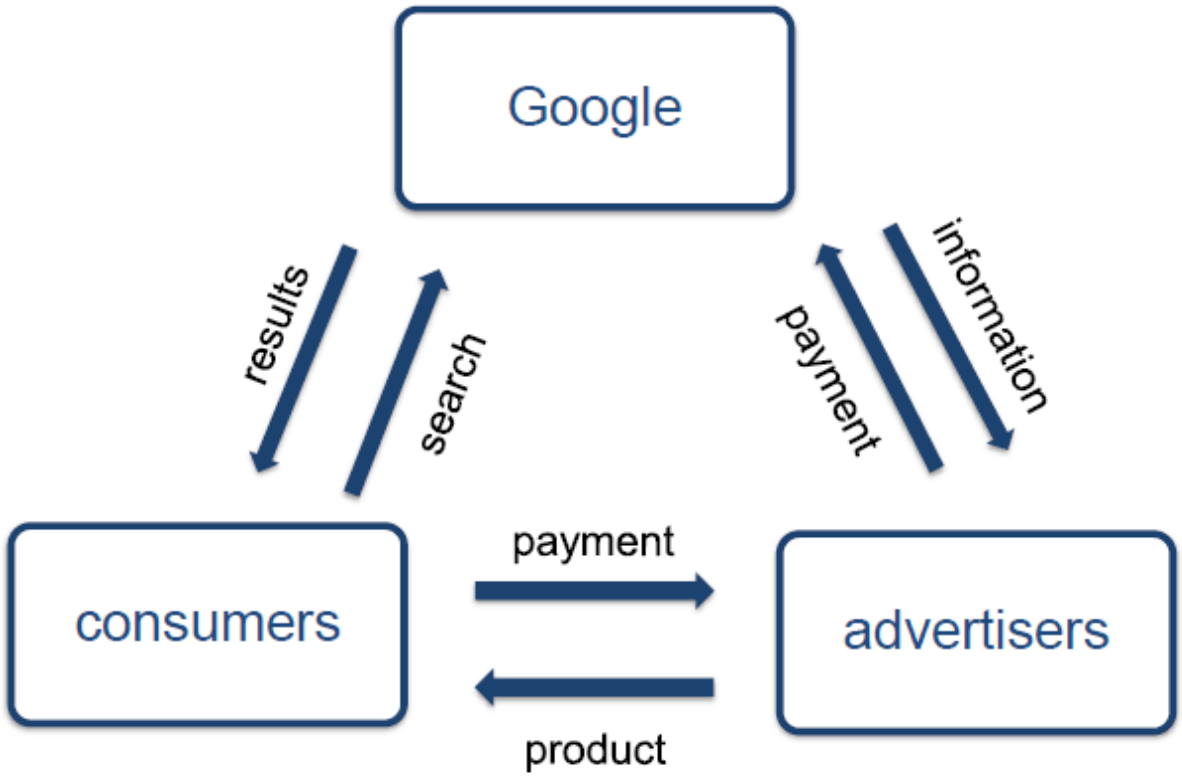
Timing

1. Data broker offers *ex ante* payment to consumer for data (signals can be *anonymized* or *matched*.)
2. Data broker sells ex-ante data to merchant
3. Data broker transmits data from consumers to merchant
4. Merchant charges uniform unit price p, or personalized price p_i; consumer i buys q_i

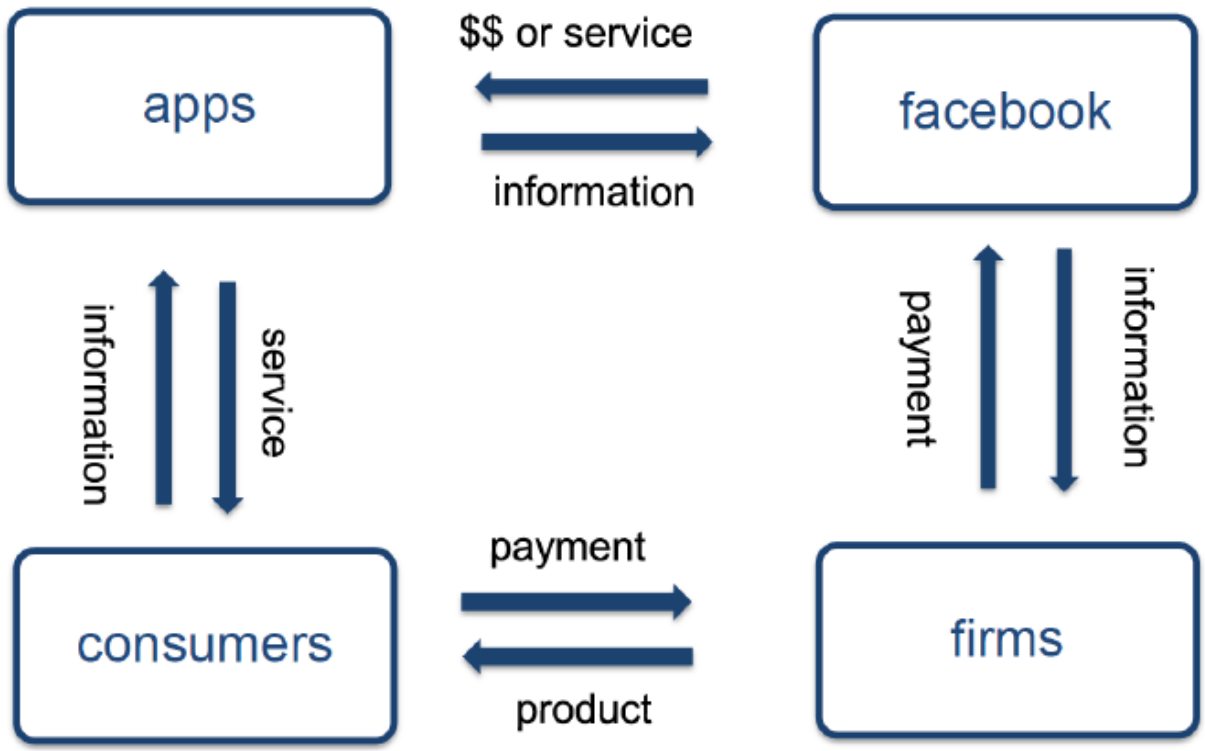
Model of Data Intermediation



Application: Google Search (Indirect Sale)



Application: Supply Chain of Data



Data in the Wild

- suppose demand information w_i were known to merchant
- offers a personalized pricing policy against demand

$$q_i^* = w_i - p_i^*$$

- personalized price:

$$p_i^* = \frac{w_i + c}{2}$$

- realized demand:

$$q_i^* = \frac{w_i - c}{2}$$

- general feature: value of match vs surplus extraction

Data and Welfare

- ex ante expected price (quantity) unaffected by information
- welfare driven by variance/covariance of surplus:

$$\Delta CS_i \triangleq CS_i(w_i, w_{-i}) - CS_i(\varnothing, \varnothing) = -\operatorname{cov}[w_i, p_i] + \frac{1}{2}\operatorname{var}[p_i]$$

$$\Delta PS_i \triangleq PS_i(w_i, w_{-i}) - PS_i(\varnothing, \varnothing) = \operatorname{cov}[w_i, p_i] - \operatorname{var}[p_i]$$

- therefore information reduces total surplus:

Proposition

Demand data increases profit of producer, decreases consumer surplus and social surplus.

- socially inefficient to trade data in downstream market

Value of Social Data

- data point s_i increases variance of individual estimate

$$\mathbb{E}\left[w_i \,|s_i\right]$$

- data point s_i increases variance of aggregate estimate

$$\mathbb{E}[\sum_j w_j \,|s_i]$$

- social nature of data: data externality (DE_i):

$$DE_i = (CS_i(\varnothing, s_{-i}) - CS_i(\varnothing, \varnothing))$$

Data Trade and Compensation

- since $\Delta CS_i < 0$, consumer i must be compensated for revealing signal s_i
- externality from information sale:

$\longrightarrow$ if sale of s_i is harmful to consumer i, i is compensated;

$\longrightarrow$ if sale of s_i helps predict $w_{j \neq i}$, i is not compensated;

$\longrightarrow$ if sale of s_i is harmful to consumer $j \neq i$,
j is not compensated

Data Intermediation: Aggregation

- should the broker collect anonymized data
- recall broker profits

$$\Pi_i = \Delta TS_i + DE_i$$

- suppose broker collects identities, considee data externality DE_i
- if i doesn't participate, p_i depends on average signal $\bar{s}_{-i}$
- unaffected by anonymous data, but less information transmitted
- therefore, the loss in TS_i is smaller

Proposition (Anonymized Data)

With ex ante homogeneous consumers, the data broker collects anonymized data iff information reduces social welfare.

- reduces consumer compensation relative to value of information

Data Intermediation: Optimality and Noise

Proposition (Optimal Data Intermediation)

1. *There exists a threshold* $\overline{N}$ *such that positive profits iff the number of consumers is* $N > \overline{N}$.
2. *Broker's profit is increasing in* σ_θ^2 *and decreasing in* $\sigma_{\theta_i}^2$.
3. *Ddata broker never adds idiosyncratic noise:* $\sigma_{\varepsilon_i}^2 = 0$.
4. *Optimal aggregate noise* $\sigma_\varepsilon^2 > 0$ *for large* $\sigma_{\theta_i}^2$ *or small* N.

- if consumers' preferences are not sufficiently correlated, broker does not trade any information
- information is traded even if it decreases social surplus
- common noise makes signals (s_i, s_j) less informative but more correlated
- correlation reduces compensation relative to value of information

First Implications

- data intermediation vs data in the wild
- uniform price rather than personalized price
- noisy transmission rather than noiseless transmission
- partial compensation of consumer:
  for individual harm, but not for social harm
- yet, far from socially efficient allocation

More Users

- as number of consumers N becomes large, individual information becomes less valuable
- let $m_i :=$ individual consumer compensation
- let $m_0 :=$ broker revenue from merchant

Growing Revenue

Proposition (Consumer Base)

1. $m_0(N)/N$ *is growing in* N*;*
2. *As* $N \to \infty$, m_0 *grows linearly in* N*.*
3. *As* $N \to \infty$, $m_i \to 0$, $N \times m_i \to k < \infty$*.*

- explains frequent absence of consumer compensation for individual data
- cost of compensation decreases with size of consumer base

More Services / More Data

- facebook connect: login tracks consumer across web, Instagram, Snapchat, Facebook Groups...
- gmail (identity), google maps, youtube...
- each source of information has idiosyncratic noise:

$$s_{i,j} = t_i + \varepsilon_{i,j}$$

- let $x =$ number of services offered to consumer i
- reducing idiosyncratic noise has a direct effect: increases the value of information
- indirect effect: lower consumer compensation as signals are more correlated

Proposition (More Data)

1. *the constrained optimal amount of common noise* $\sigma_\varepsilon^*(x)$ *is decreasing in* x;
2. *the broker's profit is convex in* x.

Concluding Thoughts

- cost of acquiring information vanishes; gains persist as markets grow large
- additional users or data sources increase broker revenue more than linearly
- value of information to intermediary $\neq$ total surplus generated

with competition:

- limited scope for increase in privacy
- implications for market structure in data intermediation sector.