The Economics of Social Data

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

[1] Yale University

[2] MIT Sloan

Pennsylvania State University

December 2020

Data and Information

- Markets for data ever more relevant to economic welfare.
  (IAB: $\sim$\$20b spent to acquire/process consumer data in 2019.)

- Rise of large internet platforms leads to unprecedented collection and commercial use of individual data.

- Amazon, Facebook, Google / JD, Tencent, Alibaba: intermediaries:

  selling information $\gg$ providing access to a database;

  consumer scores, predictions, ratings, custom audiences.

Individual and Social Data

- Central feature of individual data is its social aspect.
- “Social” dimension of the data $\triangleq$ data about an individual consumer is informative about *similar* consumers.
- Social nature of data generates a *data externality* not signed *a priori*.
- Individual data enables both *surplus creation* and *extraction*:

  product reviews, traffic data, targeted advertising;

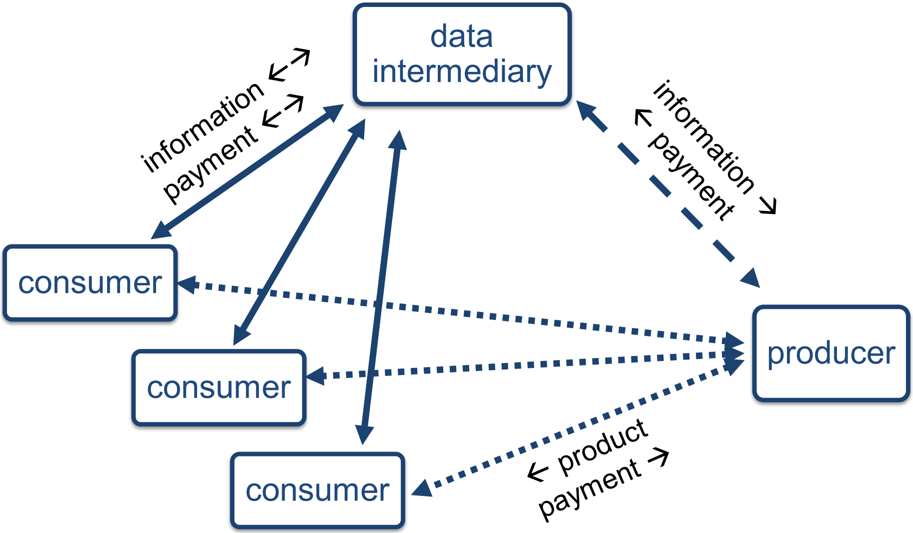
  personalized recommendations, search results, and prices.

Questions

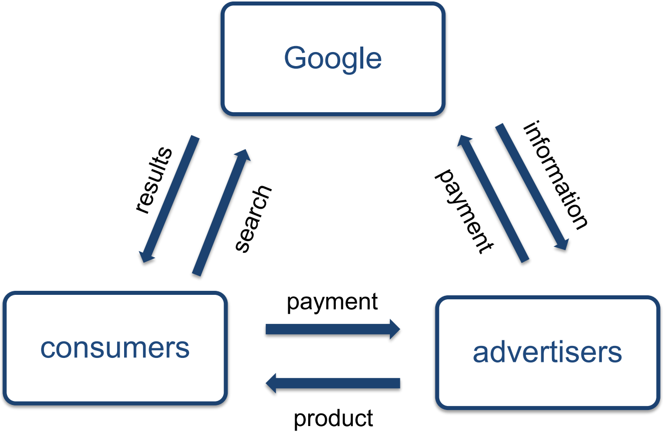
1. How does the social dimension of the data impact the terms of trade between consumers, data buyers, and data intermediaries?
2. How does the social dimension of the data magnify the value of individual data for the intermediaries?
3. How do data intermediaries choose the level of aggregation and precision of the information that they provide?

Model

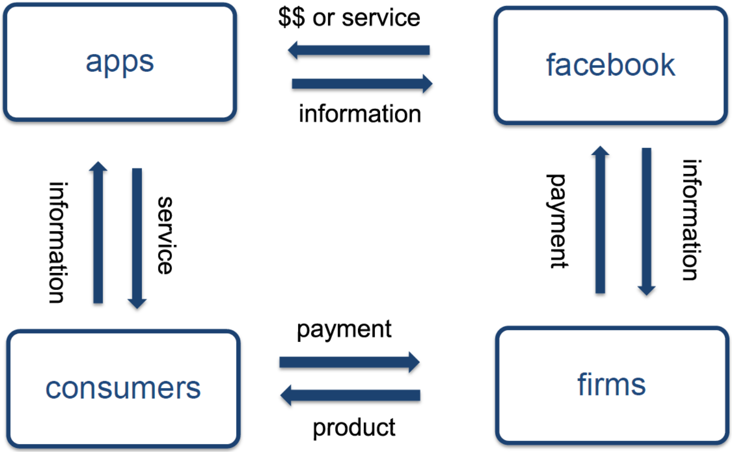
Model of Intermediation



Application: Google Search (Indirect Sale)



Application: Supply Chain of Data



Product Market

- A data broker, N consumers, and a producer (firm).
- Consumer i has baseline willingness to pay w_i.
- Consumer i chooses quantity q_i to maximize

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

- Producer chooses prices $p = \left(p_1, ..., p_N\right)$ to maximize

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

Data Environment

Joint distribution of consumers' types $w = (w_1, ..., w_N)$:

$$w \sim F_w, \text{ with } \mathbb{E}[w_i] = \mu \text{ and } \operatorname{var}[w_i] = 1 \text{ for all } i.$$

Consumer i has incomplete information about wtp w_i:

$$s_i \triangleq w_i + \sigma \cdot e_i, \quad \text{with } \sigma > 0.$$

Joint distribution of consumers' error terms $e = (e_1, ..., e_N)$:

$$e \sim F_e, \text{ with } \mathbb{E}[e_i] = 0 \text{ and } \operatorname{var}[e_i] = 1 \text{ for all } i.$$

Distributions F_w and F_e admit *symmetric* densities.

Leading Example

- Each consumer has willingness to pay

$$w_i = \theta + \theta_i$$

- Each consumer observes

$$s_i = \theta + \theta_i + \varepsilon + \varepsilon_i .$$

- Social data = common and idiosyncratic components:

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

- Common and idiosyncratic error terms ε and ε_i:

$$\begin{pmatrix} \varepsilon \\ \varepsilon_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \sigma_{\varepsilon_i}^2 \end{pmatrix} \right).$$

Key Modeling Choices

- Any information beyond common prior = consumers’ signals.
- Data sharing teaches consumers about their preferences:

  correlation in fundamental and noise terms captures social dimension;

  “common attributes” or “common experience;”
- Social data can be exploited by an adversary.
- Work-in-progress: “data for service.”

Complete Data Sharing

Data Sharing and Product Market

- All individual data $s = (s_1, ..., s_N)$ is shared completely.
- Predicted willingness to pay of i given s

$$\hat{w}_i(s) \triangleq \mathbb{E}[w_i \mid s].$$

- Realized demand function of consumer i is

$$q_i(s, p) = \hat{w}_i(s) - p.$$

- Producer charges optimal personalized price $p_i^*(s)$

$$p_i^*(s) = \frac{\hat{w}_i(s) + c}{2}.$$

Data and Welfare

- Ex ante payoffs (consumer's information, firms' information):

$$
\begin{aligned}
U_i\left(S,S\right) &\triangleq \mathbb{E}\left[u_i\left(w_i,q_i^*\left(s\right),p_i^*\left(s\right)\right)|S\right]=\frac{1}{8}\mathbb{E}\left[\left(\hat{w}_i\left(s\right)-c\right)^2|S\right],\\
\Pi_i\left(S,S\right) &\triangleq \mathbb{E}\left[\pi_i\left(q_i^*\left(s\right),p_i^*\left(s\right)\right)|S\right]=\frac{1}{4}\mathbb{E}\left[\left(\hat{w}_i\left(s\right)-c\right)^2|S\right].
\end{aligned}
$$

- Linear strategies: $\mathrm{I}^*, \mathrm{II}^*$ independent of S.
- Ex ante surplus depends on the variance of the posterior mean only.
- "Quantity" of information $(\sim R^2)$ under structure S:

$$G(S)\triangleq \mathrm{var}\left[\hat{w}_i\left(s\right)\mid S\right].$$

- Under no data sharing, consumer i has information $G(S_i)>0$.

Proposition (Value of Data Sharing)

1. *The value of complete data sharing for the producer is:*

$$\Pi_i\left(S, S\right) - \Pi_i\left(S_i, \varnothing\right) = \frac{1}{4}G(S).$$

2. *The value of complete data sharing for consumer i is:*

$$U_i\left(S, S\right) - U_i\left(S_i, \varnothing\right) = \frac{1}{2}\left(G(S) - G(S_i)\right) - \frac{3}{8}G(S).$$

3. *The social value of complete data sharing is:*

$$W_i\left(S, S\right) - W_i\left(S_i, \varnothing\right) = \frac{1}{2}\left(G(S) - G(S_i)\right) - \frac{1}{8}G(S).$$

Value of Data Sharing: Basic Properties

1. Consumers' and social welfare increase with consumers' information gains, and decrease with the firms' information gains.
2. If consumers know their types $(\sigma = 0)$, data sharing is socially harmful.
3. If consumers' types (w_i, w_j) <u>and</u> error terms (e_i, e_j) are independent, data sharing is socially harmful.
4. If consumers' don't learn anything from others' signals, data sharing is socially harmful.
5. If individual consumers are uninformed (but the complete dataset is informative), data sharing benefits consumers.

Polar Cases

1. Common type, independent errors, $s_i = w + \sigma \cdot e_i$

2. Independent types, common error term, $s_i = w_i + \sigma \cdot e$

Data Externality

Surplus of consumer i when others share their signals:

$$U_i\left(S, S_{-i}\right) \triangleq \mathbb{E}\left[u_i\left(w_i, q_i^*\left(s\right), p_i^*\left(s_{-i}\right)\right) \mid S\right].$$

Definition (Data Externality)

Data externality imposed by consumers $-i$ on consumer i,

$$DE_i \triangleq U_i\left(S, S_{-i}\right) - U_i\left(S_i, \varnothing\right)$$

Proposition (Data Externality)

The data externality DE_i is given by

$$DE_i = \frac{1}{2}\left(G(S) - G(S_i)\right) - \frac{3}{8}G(S_{-i}).$$

Data Externality: Properties

- If consumers know their types $(\sigma = 0)$, then $DE_i < 0$.
- If types (w_i, w_j) are independent, $DE_i \geq 0$.

  But if σ is small, then $DE_i > 0 > \Delta U_i$.
- $DE_i > \Delta U_i$ (the only difference is the firm observing s_i.)
- But it is possible that $\Delta W_i > 0 > DE_i$.

Data Intermediation

Data Market

Data broker buys data from each consumer and sells to producer:

1. data contract with consumer i specifies an inflow data policy

$$X_i : S_i \to \Delta\left(\mathbb{R}\right),$$

   and a fee $m_i \in \mathbb{R}$ paid to the consumer;

2. data contract with the producer specifies an outflow data policy

$$Y_0 : X \to \Delta\left(\mathbb{R}^N\right),$$

   a data sharing policy with consumers

$$Y_i : X \to \Delta\left(\mathbb{R}^N\right),$$

   and a fee $m_0 \in \mathbb{R}$ paid by the producer.

Data Market: Timing

1. Data broker offers *ex ante* payment to consumer for data.
2. Data broker offers sells available (ex ante) data to merchant.
3. Data broker transmits data from consumers to merchant.
4. Merchant charges unit price p_i; consumer i buys q_i.

Complete Data Sharing and Participation

The broker collects and shares all data with every agent $Y_0 = Y_i = X = S$.

- Producer's participation constraint

$$m_0/N \leq \Pi_i(S, S) - \Pi_i(S, \varnothing) = \Pi_i(S, S) - \Pi_i(S_i, \varnothing).$$

- Consumer i's participation constraint

$$m_i \geq U_i(S, S_{-i}) - U_i(S, S) \geq 0$$

Social nature of data: externality from information sale:

→ if sharing s_i is harmful to consumer i, consumer i is compensated;

→ if sharing s_i helps predict $w_{j \neq i}$, consumer i is not compensated;

→ if sharing s_i is harmful to $j \neq i$, consumer j is not compensated.

Data Sharing and Compensation

- Total payment from producer:

$$m_0^* = N\left(\Pi_i\left(S, S\right) - \Pi_i\left(S_i, \varnothing\right)\right).$$

- Represent consumer i's compensation as

$$\begin{aligned} m_i^* &= U_i\left(S, S_{-i}\right) - U_i\left(S, S\right) \\ &= \underbrace{U_i\left(S, S_{-i}\right) - U_i\left(S_i, \varnothing\right)}_{DE_i(S)} - \underbrace{\left(U_i\left(S, S\right) - U_i\left(S_i, \varnothing\right)\right)}_{\Delta U_i(S)}. \end{aligned}$$

- The intermediary's profit is then

$$R\left(S\right) = m_0^* - \sum_{i=1}^{N} m_i^* = \Delta W_i\left(S\right) - \sum_{i=1}^{N} DE_i\left(S\right).$$

Equilibrium with Complete Data Sharing

Proposition (Complete Data Sharing)

Complete data intermediation is profitable if and only if

$$3G\left(S_{-i}\right) \geq G\left(S\right).$$

Recall complete data sharing is efficient iff $G(S) > (4/3)G(S_i)$.

Broker profits do not depend on consumer i's initial information.

Intuitively, profits depend on signal substitutability.

Uninformative individual signals: profitable and efficient data sharing.

Market Failures

1. Type-I error: correlated fundamentals & precise individual signals.

2. Type-II error: independent fundamentals & noisy individual signals.

Gaussian Data Structures

- Common and idiosyncratic terms:

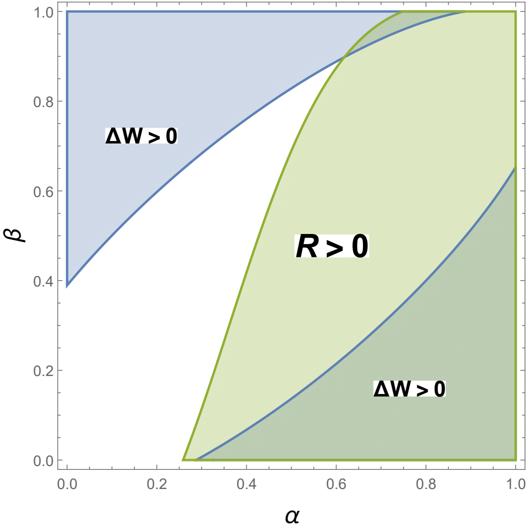
$$s_i = \theta + \theta_i + \varepsilon + \varepsilon_i.$$

- Correlation coefficients for two consumers' fundamentals and errors:

$$\alpha \triangleq \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_{\theta_i}^2}, \qquad \beta \triangleq \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_{\varepsilon_i}^2}.$$

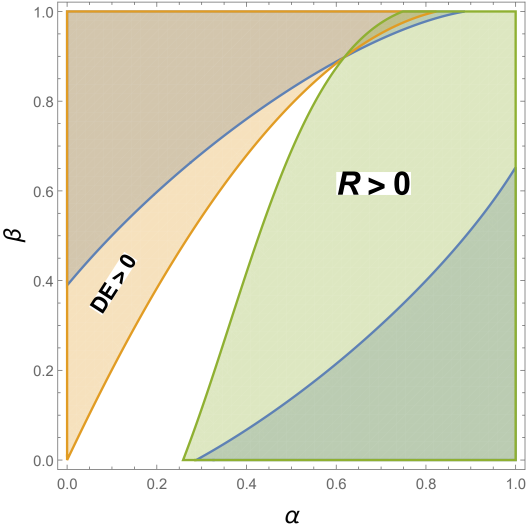
- Refer to pair $(\alpha, \beta) \in [0,1]^2$ as *data structure*.
- Data structure (α, β) captures social dimension of individual data.

Equilibrium vs Efficient Data Sharing



- Socially efficient data structures (blue) and profitable data structures (green) for $\sigma_e = 2, \sigma_w = 1, N = 10$

Equilibrium vs Efficient Data Sharing



- Socially efficient data structures (blue), profitable data structures (green), and data externality (orange) for $\sigma_e = 2, \sigma_w = 1, N = 10$

Optimal Data Sharing

Optimal Data Sharing

Design data intermediation policy along three key dimensions:

1. allow intermediary *not* to release all of the data, i.e., to introduce incomplete and possibly asymmetric information;
2. allow intermediary to choose between collecting *anonymized* or *matched* signals;
3. allow intermediary to introduce further (possibly correlated) noise terms in any (anonymized or matched) signals it collects.

Optimal Data Intermediation: Outflow

Wlog, a data inflow policy X consists of signals

$$x_i \triangleq s_i + \xi_i$$

for each consumer $i = 1, \ldots, N$ who accepts the intermediary's offer.

Proposition (Optimal Outflow)

Given any realized data inflow X, the complete data outflow policy $Y^* (X) = X$ maximizes the gross revenue of the producer among all feasible outflow data policies.

- No withholding information from the producer: sell everything.
- No superior information to the producer:
  she does not benefit from signaling ex ante.

Optimal Data Intermediation: Inflow

Theorem (Data Anonymization)

For any data inflow X, the intermediary obtains strictly greater profits by collecting anonymized rather than matched signals.

Proof Sketch

Recall the intermediary's profits:

$$R(S) = \Delta W_i(S) - \sum_{i=1}^{N} DE_i(S).$$

- Suppose broker collects matched signals, consider data externality.
- By symmetry, if consumer i does not participate, $p_i^*(s_{-i})$ is independent of other consumers' identities.
- Data externality unchanged under anonymization.
- If consumer i participates, her inference problem does not depend on the identities of $j \neq i$.
- Firm's inference problem is now harder, which improves total surplus.

Profitable Intermediation

Anonymized data sharing is profitable iff $3G(S_{-i}) \geq \tilde{G}(S)$.

If types are independent, still no profitable intermediation.

If matched sharing is profitable AND efficient, so is anonymized sharing.

Intuition (linear estimators): anonymized signals are closer substitutes.

Large Markets

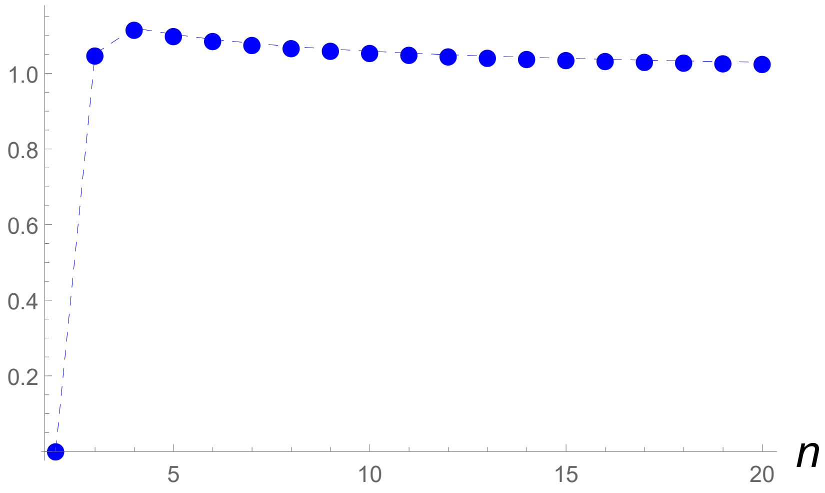
- "Digital privacy paradox:" negligible compensation for individual data.
- Compensation decreases with size of consumer base.

Theorem (Large Markets (Gaussian Case))

1. *As $N \to \infty$, the individual consumer's compensation goes to zero, and the total compensation converges to a finite number.*
2. *For sufficiently correlated fundamentals the total compensation is asymptotically decreasing in N.*
3. *As $N \to \infty$, the intermediary's revenue and profit grow linearly in N.*

Large Markets

Total Consumer Compensation



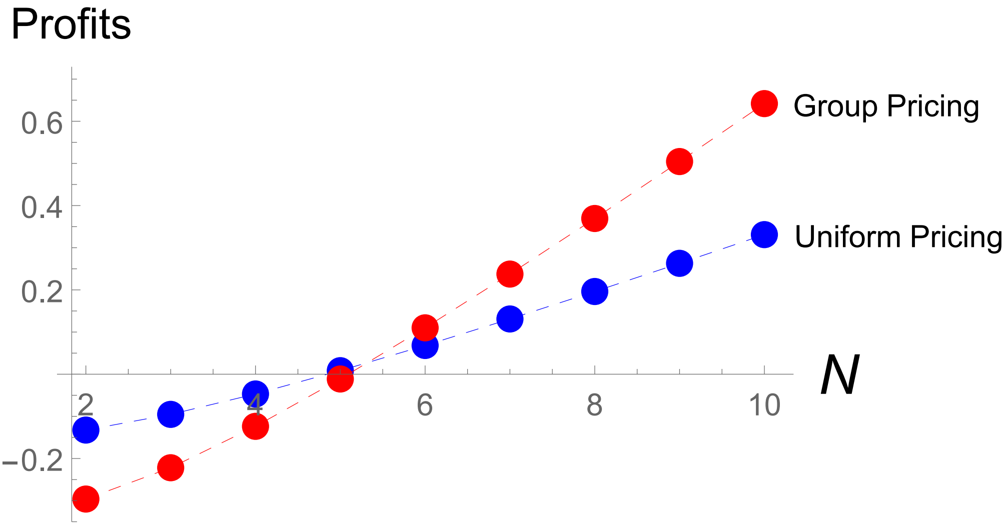
Limits to Anonymization

Proposition (General Anonymization)

Suppose consumers are ex ante homogeneous. The data broker collects anonymized data if and only if information reduces social welfare.

- With multiple consumer segments, the intermediary reveals (at most) each consumer's *group* identity.
- Profitability of group vs. uniform price depends on N, degree of within-group and across-group correlation.

Gaussian Case: Multiple Segments



Recommender Systems

Consumer i's utility function is given by

$$u_i\left(w_i, q_i, p_i, y_i, t_i\right) = \left(w_i - \left(y_i - t_i\right)^2 - p_i\right) q_i - q_i^2/2,$$

- w_i is willingness to pay, t_i is consumer's ideal location.
- y_i is the product's characteristic.
- Location $t_i \in \mathbb{R}$ of each consumer i is

$$t_i \triangleq \tau + \tau_i.$$

Proposition (Optimal Recommendation)

The intermediary's optimal policy collects anonymized data on the vertical component w_i and matched data on the horizontal component t_i.

Concluding Thoughts

Optimal data sharing vs complete data sharing:

- uniform price rather than personalized prices;
- personalized recommendations.

Far from socially efficient allocation of data:

- consumers compensated for individual harm, but not for social harm;
- socially efficient anonymization, not intermediation decisions;
- cost of acquiring information vanishes, gains persist as market grows.