The Value of Data Records

Simone Galperti
UC San Diego

Aleksandr Levkun
UC San Diego

Jacopo Perego
Columbia University

Motivation

Personal data has become a key input in the modern economy

- Search and social media platforms use it to sell targeted ads
- E-commerce platforms use it to intermediate buyers and sellers
- Matching platforms use it decrease search frictions

In each case, personal data fuels a multi-billion dollars industry

Motivation

Personal data has become a key input in the modern economy

- Search and social media platforms use it to sell targeted ads
- E-commerce platforms use it to intermediate buyers and sellers
- Matching platforms use it decrease search frictions

In each case, personal data fuels a multi-billion dollars industry

This Paper: How much of this value is generated by the data of a single individual?

This question is at the core of some recent debates on data markets:

- Compensate individuals for their data (Seim et al., ’22, PW’18)
- Conduct demand analysis in data markets (FTC ’14)
- Data as a source of market power (Stiegler Report ’19)

example

A Simple Example

A two-sided market:

- An e-commerce platform
- Many buyers
- A firm (third-party seller)

A Simple Example

A monopolist **firm** sells its product through an e-commerce **platform**

The platform is used by group of **buyers**, each with independent valuation for the product

A Simple Example

A monopolist **firm** sells its product through an e-commerce **platform**

The platform is used by group of **buyers**, each with independent valuation for the product

For each buyer, platform owns a **record** of her personal characteristics, which is informative about her valuation

A Simple Example

A monopolist **firm** sells its product through an e-commerce **platform**

The platform is used by group of **buyers**, each with independent valuation for the product

For each buyer, platform owns a **record** of her personal characteristics, which is informative about her valuation

Only two **types** of records:

- ω_L reveals buyer has valuation 1
- ω_H reveals buyer has valuation 2

A Simple Example

A monopolist **firm** sells its product through an e-commerce **platform**

The platform is used by group of **buyers**, each with independent valuation for the product

For each buyer, platform owns a **record** of her personal characteristics, which is informative about her valuation

Only two **types** of records:

- ω_L reveals buyer has valuation 1
- ω_H reveals buyer has valuation 2

Platform's **database** contains:

3 million such records

6 million such records

A Simple Example

A monopolist **firm** sells its product through an e-commerce **platform**

The platform is used by group of **buyers**, each with independent valuation for the product

For each buyer, platform owns a **record** of her personal characteristics, which is informative about her valuation

Only two **types** of records:

- ω_L reveals buyer has valuation 1
- ω_H reveals buyer has valuation 2

Platform's **database** contains:

3 million such records

6 million such records

Seller knows database composition but ignores each specific ω

A Simple Example

Platform is an **intermediary** that provides the firm with **information** about each buyer, and thus can influence the price it charges to them

Firm chooses prices to maximizes profits (MC = 0)

Suppose platform choose information to maximizes buyer's surplus

A Simple Example

Platform is an **intermediary** that provides the firm with **information** about each buyer, and thus can influence the price it charges to them

Firm chooses prices to maximizes profits (MC = 0)

Suppose platform choose information to maximizes buyer's surplus

Question: How much value does platform derive from each record?

A Simple Example

An optimal information policy for the platform: (as in BBM '15, *AER*)

Record of type $\omega_L \longrightarrow s^L$

$$\begin{tikzcd}
 & s^L \\
\text{Record of type } \omega_H \arrow[ur, "\frac{1}{2}"] \arrow[dr, "\frac{1}{2}"'] & \\
 & s^H
\end{tikzcd}$$

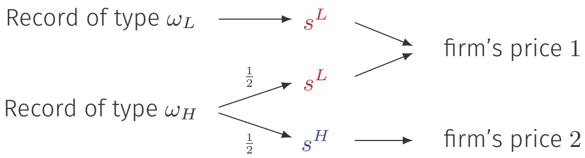
A Simple Example

An optimal information policy for the platform: (as in BBM '15, *AER*)

$$
\begin{tikzcd}
\text{Record of type } \omega_L \arrow[r] & s^L \arrow[dr] & \\
& & \text{firm's price } 1 \\
& s^L \arrow[ur] & \\
\text{Record of type } \omega_H \arrow[ur, "\frac{1}{2}"] \arrow[dr, "\frac{1}{2}"'] & & \\
& s^H \arrow[r] & \text{firm's price } 2
\end{tikzcd}
$$

A Simple Example

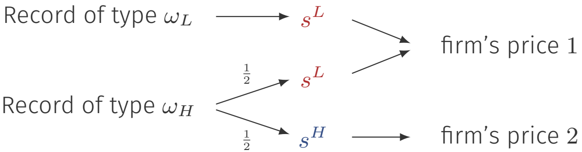
An optimal information policy for the platform: (as in BBM '15, *AER*)



Thus, platform's **payoff** is $u^*(\omega) = \begin{cases} 0 & \text{if } \omega_L \\ \frac{1}{2} & \text{if } \omega_H \end{cases}$

A Simple Example

An optimal information policy for the platform: (as in BBM '15, *AER*)

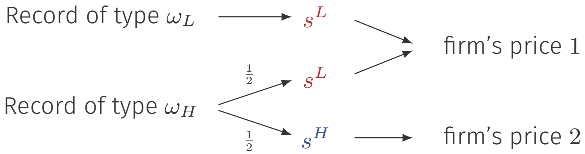


Thus, platform's **payoff** is $u^*(\omega) = \begin{cases} 0 & \text{if } \omega_L \\ \frac{1}{2} & \text{if } \omega_H \end{cases}$

Are ω_L records really worthless?

A Simple Example

An optimal information policy for the platform: (as in BBM '15, *AER*)

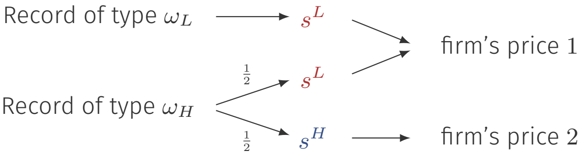


Thus, platform's **payoff** is $u^*(\omega) = \begin{cases} 0 & \text{if } \omega_L \\ \frac{1}{2} & \text{if } \omega_H \end{cases}$

Are ω_L records really worthless? No! $v^*(\omega) = \begin{cases} 1 & \text{if } \omega_L \\ 0 & \text{if } \omega_H \end{cases}$

A Simple Example

An optimal information policy for the platform: (as in BBM '15, *AER*)



Thus, platform's **payoff** is $u^*(\omega) = \begin{cases} 0 & \text{if } \omega_L \\ \frac{1}{2} & \text{if } \omega_H \end{cases}$

Are ω_L records really worthless? No! $\quad v^*(\omega) = \begin{cases} 1 & \text{if } \omega_L \\ 0 & \text{if } \omega_H \end{cases}$

1. Most valuable records are those yielding lowest payoff
2. ω_L generates no payoff but "helps" ω_H earn positive surplus
3. Payoff u^* gives *biased* account of the value created by a record

An optimal information policy for the platform: (as in BBM '15, *AER*)

$$
\begin{tikzcd}
\text{Record of type } \omega_L \arrow[r] & s^L \arrow[dr] & \\
 & & \text{firm's price } 1 \\
\text{Record of type } \omega_H \arrow[r, "\frac{1}{2}"] \arrow[dr, "\frac{1}{2}"'] & s^L \arrow[ur] & \\
 & s^H \arrow[r] & \text{firm's price } 2
\end{tikzcd}
$$

Thus, platform's **payoff** is $u^*(\omega) = \begin{cases} 0 & \text{if } \omega_L \\ \frac{1}{2} & \text{if } \omega_H \end{cases}$

Are ω_L records really worthless? No! $v^*(\omega) = \begin{cases} 1 & \text{if } \omega_L \\ 0 & \text{if } \omega_H \end{cases}$

1. Most valuable records are those yielding lowest payoff
2. ω_L generates no payoff but "helps" ω_H earn positive surplus
3. Payoff u^* gives *biased* account of the value created by a record

A Simple Example

An optimal information policy for the platform: (as in BBM '15, *AER*)

$$
\begin{tikzcd}
\text{Record of type } \omega_L \arrow[r] & s^L \arrow[dr] & \\
& & \text{firm's price } 1 \\
& s^L \arrow[ur] & \\
\text{Record of type } \omega_H \arrow[ur, "\frac{1}{2}"] \arrow[dr, "\frac{1}{2}"'] & & \\
& s^H \arrow[r] & \text{firm's price } 2
\end{tikzcd}
$$

Thus, platform's **payoff** is $u^*(\omega) = \begin{cases} 0 & \text{if } \omega_L \\ \frac{1}{2} & \text{if } \omega_H \end{cases}$

Are ω_L records really worthless? No! $v^*(\omega) = \begin{cases} 1 & \text{if } \omega_L \\ 0 & \text{if } \omega_H \end{cases}$

1. Most valuable records are those yielding lowest payoff
2. ω_L generates no payoff but "helps" ω_H earn positive surplus
3. Payoff u^* gives *biased* account of the value created by a record

this paper

Questions

We address two sets of questions

1. What is the value of a data record for the platform? What are its properties?

Questions

We address two sets of questions

1. What is the value of a data record for the platform? What are its properties?

2. What is the platform's WTP for "more data"?

Questions

We address two sets of questions

1. What is the value of a data record for the platform? What are its properties?

2. What is the platform's WTP for "more data"?

– For *more* data records

  E.g., Platform obtains access to *new* buyers

– For *better* data records

  E.g., Platform obtains more characteristics about *existing* buyers

Questions

We address two sets of questions

1. What is the value of a data record for the platform? What are its properties?

2. What is the platform's WTP for "more data"?

- For *more* data records — Marketing Lists

  E.g., Platform obtains access to *new* buyers

- For *better* data records — Data Appends

  E.g., Platform obtains more characteristics about *existing* buyers

Approach

Formulate the platform's problem as an information-design problem

We interpret it as a physical production problem:

- Platform uses **inputs** (data records) to produce **outputs** (signals/recommendations)

Use LP **duality** to characterize the **unit value** of these inputs in this "production" problem (Dorfman et al. '87 and Gale '89)

Contributions

In this paper:

1. We show how to compute the value of data for an intermediary and study the properties of the demand for data
2. We uncover a new **data externality** that, if ignored, can bias our assessment of the value of data
3. We offer a benchmark for how to compensate consumers for their data

Literature

This paper connects two fast-growing literatures:

1. Data Markets — Bergemann and Bonatti (2019) Bergemann and Ottaviani (2021)

2. Information design — Bergemann and Morris (2019), Kamenica (2019)

Literature

This paper connects two fast-growing literatures:

1. Data Markets — Bergemann and Bonatti (2019) Bergemann and Ottaviani (2021)

2. Information design — Bergemann and Morris (2019), Kamenica (2019)

Literature

This paper connects two fast-growing literatures:

1. Data Markets Bergemann and Bonatti (2019) Bergemann and Ottaviani (2021)
   - Use of a database — how to design and sell information e.g., Admati and Pfleiderer (86, 90), Bergemann and Bonatti (15), Bergemann et al. (18), Yang (20)
     - Our focus is not on use; but on inputs affect use (“upstream”)

2. Information design Bergemann and Morris (2019), Kamenica (2019)

Literature

This paper connects two fast-growing literatures:

1. Data Markets — Bergemann and Bonatti (2019) Bergemann and Ottaviani (2021)
   - Use of a database — how to design and sell information e.g., Admati and Pfleiderer (86, 90), Bergemann and Bonatti (15), Bergemann et al. (18), Yang (20)
     - Our focus is not on use; but on inputs affect use (“upstream”)
   - Consumers’ incentives to disclose data — learning externalities e.g., Choi et al. (19), Acemoglu et al. (21), Ichihashi (21), Bergemann et al. (22)
     - Ann’s record uninformative about Bob’s $\rightsquigarrow$ new data externality
     - No disclosure, platform already owns the database
2. Information design — Bergemann and Morris (2019), Kamenica (2019)

Literature

This paper connects two fast-growing literatures:

1. Data Markets — Bergemann and Bonatti (2019) Bergemann and Ottaviani (2021)
   - Use of a database — how to design and sell information e.g., Admati and Pfleiderer (86, 90), Bergemann and Bonatti (15), Bergemann et al. (18), Yang (20)
     - Our focus is not on use; but on inputs affect use (“upstream”)
   - Consumers’ incentives to disclose data — learning externalities e.g., Choi et al. (19), Acemoglu et al. (21), Ichihashi (21), Bergemann et al. (22)
     - Ann’s record uninformative about Bob’s $\rightsquigarrow$ new data externality
     - No disclosure, platform already owns the database
2. Information design — Bergemann and Morris (2019), Kamenica (2019)
   - See paper for connections to mecha/info design and duality

Plan for Rest of the Talk

1. Model
2. Value of Data Records & their externalities
3. Properties of the demand for data

model

Model

For the talk, model simply generalizes the example

For the talk, model simply generalizes the example

Denote **platform** by $i = 0$; denote **firm** by $i = 1$ and his action $a \in A$

A **buyer's** preference pinned down by independently distributed θ

For the talk, model simply generalizes the example

Denote **platform** by $i = 0$; denote **firm** by $i = 1$ and his action $a \in A$

A **buyer's** preference pinned down by independently distributed θ

A buyer's record is of type $\omega \in \Omega$ and is (partially) informative about her θ

Database composition $q \in \mathbb{R}_+^\Omega$ is common knowledge

For $i \in \{0, 1\}$, $u_i : A \times \Omega \to \mathbb{R}$ denotes i's **expected** payoff function

Model

Platform intermediates the interaction between the firm and the buyers

Specifically, platform acts as an **information designer**:

- It sends the firm information about each record ω so as to influence the firm's action a (price, discount, features, ect.)

Remark. WLOG to focus on "recommendation" mechanisms like

$$x : \Omega \to \Delta(A)$$

Model

The platform's problem is then: as in Bergemann-Morris '16

$$\mathcal{U}_q: \qquad \max_x \sum_{\omega,a} u_0(a,\omega) x(a|\omega) q(\omega)$$

s.t. for all a, a',

$$\sum_\omega \Big(u_1(a,\omega) - u_1(a',\omega) \Big) x(a|\omega) q(\omega) \geq 0$$

The platform's problem is then: as in Bergemann-Morris '16

$$\mathcal{U}_q: \quad \max_x \sum_{\omega,a} u_0(a,\omega)x(a|\omega)q(\omega)$$

s.t. for all a, a',

$$\sum_\omega \Big(u_1(a,\omega) - u_1(a',\omega)\Big)x(a|\omega)q(\omega) \geq 0$$

We let x_q^* be an optimal solution and define

- **direct payoff** of type-ω record: $u_q^*(\omega) \triangleq \sum_a u_0(a,\omega)x_q^*(a|\omega)$
- **total payoff** of database: $U^*(q) \triangleq \sum_\omega u_q^*(\omega)q(\omega)$

Today's results immediately extend to more general settings

- Multiple agents: E.g., competing firms
- Platform does more than information: E.g., allocations and transfers
- Agents have some of the principal's data

Why caring about this generality? More applications besides e-commerce: sponsored-search auctions, rideshare, navigation value of data records

Value Problem

Platform uses records as **inputs** to produce **output** in the form of informative recommendations $\rightsquigarrow$ linear program $\mathcal{U}_q$

We use **duality** to reveal value of each input Dorfman et al. ’87, Gale ’89

Platform uses records as **inputs** to produce **output** in the form of informative recommendations $\rightsquigarrow$ linear program $\mathcal{U}_q$

We use **duality** to reveal value of each input Dorfman et al. '87, Gale '89

Let $v : \Omega \to \mathbb{R}$ and $\lambda : A \times A \to \mathbb{R}_+$

The Data-Value Problem:

$$\mathcal{V}_q : \quad \min_{\lambda, v} \sum_{\omega} v(\omega) q(\omega)$$

s.t. for all $\omega \in \Omega$,

$$v(\omega) = \max_{a \in A} \Big\{ u_0(a,\omega) + t(a,\omega) \Big\} \qquad \text{(value formula)}$$

where $t(a,\omega) \triangleq \sum_{a' \in A} \Big(u_1(a,\omega) - u_1(a',\omega) \Big) \lambda(a'|a)$

Lemma 1 (Duality)

$\mathcal{V}_q$ is equivalent to the dual of $\mathcal{U}_q$. For every optimal solution v_q^* and x_q^*,

$$\sum_{\omega \in \Omega} v_q^*(\omega) q(\omega) = U^*(q) \triangleq \sum_{\omega \in \Omega} u_q^*(\omega) q(\omega)$$

Lemma 1 (Duality)

$\mathcal{V}_q$ is equivalent to the dual of $\mathcal{U}_q$. For every optimal solution v_q^* and x_q^*,

$$\sum_{\omega \in \Omega} v_q^*(\omega) q(\omega) = U^*(q) \triangleq \sum_{\omega \in \Omega} u_q^*(\omega) q(\omega)$$

Lemma 1 (Duality)

$\mathcal{V}_q$ is equivalent to the dual of $\mathcal{U}_q$. For every optimal solution v_q^* and x_q^*,

$$\sum_{\omega \in \Omega} v_q^*(\omega) q(\omega) = U^*(q) \triangleq \sum_{\omega \in \Omega} u_q^*(\omega) q(\omega)$$

- $v_q^*(\omega)$ is multiplier of feasibility constraint $\rightsquigarrow$ captures the effect on $U^*(q)$ of a change in $q(\omega)$
- $v_q^*(\omega)$ is the **unit value** of a record of type ω (Gale '89)
- We characterize the properties of $v_q^*(\omega)$

Decomposing the Value of a Record

What determines the value of a record?

Decomposing the Value of a Record

What determines the value of a record?

Proposition (Decomposition)

The value of a record ω can be decomposed as

$$v_q^*(\omega) = u_q^*(\omega) + t_q^*(\omega) \qquad \text{where:}$$

Decomposing the Value of a Record

What determines the value of a record?

Proposition (Decomposition)

The value of a record ω can be decomposed as

$$v_q^*(\omega) = u_q^*(\omega) + t_q^*(\omega) \qquad \text{where:}$$

$$u_q^*(\omega) \quad \triangleq \quad \sum_a u_0(a,\omega)x_q^*(a|\omega) \qquad \text{direct payoff}$$

- $u_q^*(\omega)$ captures the payoff that platform earns directly from record

What determines the value of a record?

Proposition (Decomposition)

The value of a record ω can be decomposed as

$$v_q^*(\omega) = u_q^*(\omega) + t_q^*(\omega) \qquad \text{where:}$$

$$u_q^*(\omega) \triangleq \sum_a u_0(a,\omega) x_q^*(a|\omega) \qquad \textbf{direct payoff}$$

$$t_q^*(\omega) \triangleq \sum_a t^*(a,\omega) x^*(a|\omega) \overset{\text{a.e.}}{=} \sum_{\omega'} q(\omega') \frac{\partial}{\partial q(\omega)} u_q^*(\omega') \qquad \textbf{externality}$$

- $t_q^*(\omega)$ **externality** that ω exerts on payoffs generated by other records

What determines the value of a record?

Proposition (Decomposition)

The value of a record ω can be decomposed as

$$v_q^*(\omega) = u_q^*(\omega) + t_q^*(\omega) \qquad \text{where:}$$

$$u_q^*(\omega) \triangleq \sum_a u_0(a,\omega) x_q^*(a|\omega) \qquad \text{direct payoff}$$

$$t_q^*(\omega) \triangleq \sum_a t^*(a,\omega) x^*(a|\omega) \overset{\text{a.e.}}{=} \sum_{\omega'} q(\omega') \frac{\partial}{\partial q(\omega)} u_q^*(\omega') \qquad \text{externality}$$

- $t_q^*(\omega)$ **externality** that ω exerts on payoffs generated by other records
- Externality relates to seller's **incentives** to disobey recommendations

What determines the value of a record?

Proposition (Decomposition)

The value of a record ω can be decomposed as

$$v_q^*(\omega) = u_q^*(\omega) + t_q^*(\omega) \qquad \text{where:}$$

$$u_q^*(\omega) \triangleq \sum_a u_0(a,\omega) x_q^*(a|\omega) \qquad \text{direct payoff}$$

$$t_q^*(\omega) \triangleq \sum_a t^*(a,\omega) x^*(a|\omega) \overset{\text{a.e.}}{=} \sum_{\omega'} q(\omega') \frac{\partial}{\partial q(\omega)} u_q^*(\omega') \qquad \text{externality}$$

- $t_q^*(\omega)$ externality that ω exerts on payoffs generated by other records
- This result clarifies why/when $u_q^*(\omega)$ is biased measure of value

We characterize when records exert positive vs negative externalities

Externality

We characterize when records exert positive vs negative externalities

Recall notation:

- $u_q^*(\omega) \rightsquigarrow$ direct payoff the platform obtains from record
- $\bar{u}(\omega) \rightsquigarrow$ payoff the platform could obtain with "full disclosure"

We characterize when records exert positive vs negative externalities

Recall notation:

- $u_q^*(\omega) \rightsquigarrow$ direct payoff the platform obtains from record
- $\bar{u}(\omega) \rightsquigarrow$ payoff the platform could obtain with "full disclosure"

Corollary

$$\text{If} \quad u_q^*(\omega) < \bar{u}(\omega), \quad \text{then} \quad t_q^*(\omega) > 0$$

Idea:

- $u_q^*(\omega) < \bar{u}(\omega)$ implies platforms withhold some information from firm

We characterize when records exert positive vs negative externalities

Recall notation:

- $u_q^*(\omega) \rightsquigarrow$ direct payoff the platform obtains from record
- $\bar{u}(\omega) \quad \rightsquigarrow$ payoff the platform could obtain with “full disclosure”

Corollary

$$\begin{array}{llll} \text{If} & u_q^*(\omega) < \bar{u}(\omega), & \text{then} & t_q^*(\omega) > 0 \\ \text{If} & t_q^*(\omega) < 0, & \text{then} & u_q^*(\omega) > \bar{u}(\omega) \end{array}$$

Moreover, $t_q^*(\omega) < 0$ for some ω if and only if $t_q^*(\omega') > 0$ for some ω'

Idea:

- $u_q^*(\omega) < \bar{u}(\omega)$ implies platforms withhold some information from firm

- This externality arises when platform **withholds info** from firms by pooling data records
    - Intermediation problems, as opposed to decision problems
        - Intermediation may involve balancing conflicting interests
    - Ubiquitous due to rise of "info-mediaries" Acquisti et al. (16)

Takeaways

- This externality arises when platform **withholds info** from firms by pooling data records
  - Intermediation problems, as opposed to decision problems
    - Intermediation may involve balancing conflicting interests
  - Ubiquitous due to rise of "info-mediaries" Acquisti et al. (16)
- The externality arises even when records are statistically **independent**
  - Thus, unrelated to "learning" externalities, (vs Choi et al. (19), Bergemann et al. (20), Acemoglu et al. (21), Ichihashi (21))

Example

Back to our introductory example:

Suppose $\Omega = \{\omega_1, \ldots, \omega_K\}$ and record of type ω_k fully reveals that $\theta = \omega_k$

Example

Back to our introductory example:

Suppose $\Omega = \{\omega_1, \ldots, \omega_K\}$ and record of type ω_k fully reveals that $\theta = \omega_k$

Suppose platform objective is:

$$u_0(a, \omega) = \beta\Big(\underbrace{a\mathbb{1}\{\omega \geq a\}}_{\text{seller's profit}}\Big) + (1-\beta)\Big(\underbrace{\max\{\omega - a, 0\}}_{\text{buyer's surplus}}\Big) \qquad \text{for } \beta \in [0,1]$$

Example

Proposition (Single-Crossing)

If $\beta < 1/2$, then $t_q^*(\omega) > 0$ for all $\omega < a_q$ and $t_q^*(\omega) \leq 0$ for all $\omega \geq a_q$

When β small, u_q^* provides biased account of the value of each record

Ignoring this externality may lead to:

- over-compensate higher-valuation buyers for their data $u_q^*(\omega) > v_q^*(\omega)$
- under-compensate lower-valuation buyers for their data $u_q^*(\omega) < v_q^*(\omega)$

Example

Proposition (Single-Crossing)

If $\beta < 1/2$, then $t_q^*(\omega) > 0$ for all $\omega < a_q$ and $t_q^*(\omega) \leq 0$ for all $\omega \geq a_q$

If $\beta \geq 1/2$, then $t_q^*(\omega) = 0$ for all ω

When β small, u_q^* provides biased account of the value of each record

Ignoring this externality may lead to:

- over-compensate higher-valuation buyers for their data $u_q^*(\omega) > v_q^*(\omega)$
- under-compensate lower-valuation buyers for their data $u_q^*(\omega) < v_q^*(\omega)$

When β large, interests are "sufficiently" aligned: externality disappears demand for data

Having More Data

What is the platform's **willingness to pay** for more data?

We study two cases ($\approx$ kinds of information products):

1. The platform obtains *more* records
2. The platform obtains *better* records

We can study both cases by exploring how v_q^* depends on q

Having More Data

What is the platform's willingness to pay for more data?

We study two cases ($\approx$ kinds of information products):

1. The platform obtains *more* records
2. The platform obtains *better* records

We can study both cases by exploring how v_q^* depends on q

Platform as a “consumer” of data records:

- $U^*(q)$ is (indirect) utility of a bundle q (i.e. the database)

Therefore,

- **WTP** for type-ω records is revealed by marginal utility: $v_q^*(\omega)$
- **Substitutability** between records: $MRS_q(\omega, \omega') \overset{\text{a.e.}}{=} -\frac{v_q^*(\omega)}{v_q^*(\omega')}$

Thus, v_q^* characterizes the platform’s preferences over databases

Use this to characterize properties of the **demand function**

How does $v_q^*(\omega)$ depend on $q(\omega)$?

How does $v_q^*(\omega)$ depend on $q(\omega)$? It is downward-sloping

How does $v_q^*(\omega)$ depend on $q(\omega)$? It is downward-sloping

In fact, a much more general result holds:

Notation: $\mu_q(\omega) \triangleq \frac{q(\omega)}{\sum_{\omega'} q(\omega')}$ is the frequency of type-ω records

Proposition (Scarcity Principle)

Fix q and q'. If $\mu_q(\omega) < \mu_{q'}(\omega)$, then $v_q^*(\omega) \geq v_{q'}^*(\omega)$.

Demand for Records

How does $v_q^*(\omega)$ depend on $q(\omega)$? It is **downward-sloping**

In fact, a much more general result holds:

Notation: $\mu_q(\omega) \triangleq \dfrac{q(\omega)}{\sum_{\omega'} q(\omega')}$ is the frequency of type-ω records

Proposition (Scarcity Principle)

Fix q and q'. If $\mu_q(\omega) < \mu_{q'}(\omega)$, then $v_q^*(\omega) \geq v_{q'}^*(\omega)$.

Moreover, when $\mu_q(\omega)$ grows, $v_q^*(\omega) \searrow \bar{u}(\omega) \triangleq$ payoff under full-disclosure

How does $v_q^*(\omega)$ depend on $q(\omega)$? It is downward-sloping

In fact, a much more general result holds:

Notation: $\mu_q(\omega) \triangleq \frac{q(\omega)}{\sum_{\omega'} q(\omega')}$ is the frequency of type-ω records

Proposition (Scarcity Principle)

Fix q and q'. If $\mu_q(\omega) < \mu_{q'}(\omega)$, then $v_q^*(\omega) \geq v_{q'}^*(\omega)$.

Moreover, when $\mu_q(\omega)$ grows, $v_q^*(\omega) \searrow \bar{u}(\omega) \triangleq$ payoff under full-disclosure

How does $v_q^*(\omega)$ depend on $q(\omega)$? It is **downward-sloping**

In fact, a much more general result holds:

Notation: $\mu_q(\omega) \triangleq \dfrac{q(\omega)}{\sum_{\omega'} q(\omega')}$ is the frequency of type-ω records

Proposition (Scarcity Principle)

Fix q and q'. If $\mu_q(\omega) < \mu_{q'}(\omega)$, then $v_q^*(\omega) \geq v_{q'}^*(\omega)$.

Moreover, when $\mu_q(\omega)$ grows, $v_q^*(\omega) \searrow \bar{u}(\omega) \triangleq$ payoff under full-disclosure

Moreover, values v^* are stable for local changes of q:

How does $v_q^*(\omega)$ depend on $q(\omega)$? It is **downward-sloping**

In fact, a much more general result holds:

Notation: $\mu_q(\omega) \triangleq \dfrac{q(\omega)}{\sum_{\omega'} q(\omega')}$ is the frequency of type-ω records

Proposition (Scarcity Principle)

Fix q and q'. If $\mu_q(\omega) < \mu_{q'}(\omega)$, then $v_q^*(\omega) \geq v_{q'}^*(\omega)$.

Moreover, when $\mu_q(\omega)$ grows, $v_q^*(\omega) \searrow \bar{u}(\omega) \triangleq$ payoff under full-disclosure

Moreover, values v^* are stable for local changes of q:

Proposition (Locally Constant)

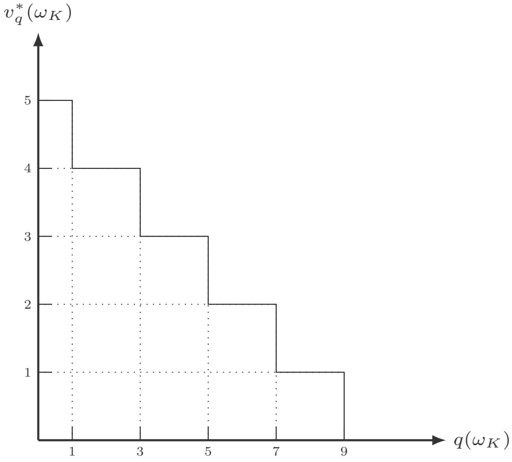
There is finite collection $\{Q_1, \ldots, Q_N\} \subseteq \mathbb{R}_+^\Omega$ of open, convex, disjoint sets s.t. $\bigcup Q_n$ has full measure and v_q^* is **constant** in $q \in Q_n$ for each n

Example (continued): Demand Curve

An example of a demand curve for records of type ω_K

Example (continued): Demand Curve

An example of a **demand curve** for records of type ω_K



Setting $\beta = 0$, $K = 10$, $q(\omega_k) = 1 \; \forall k < K$

Are different types of data records complements or substitutes?

Result. Data records exhibit complementaries iff platform withholds some information

Let's first see this through an example

Example (continued): Indifference Curves

Recall that: $u_0(a, \omega) = \beta\Big(\text{seller's profit}\Big) + (1-\beta)\Big(\text{buyer's surplus}\Big)$

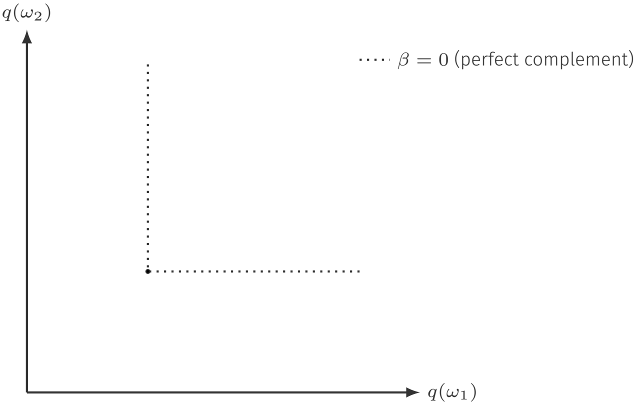
An example of the platform's **indifference curves**

Set $K = 2$

Example (continued): Indifference Curves

Recall that: $u_0(a, \omega) = \beta\Big(\text{seller's profit}\Big) + (1-\beta)\Big(\text{buyer's surplus}\Big)$

An example of the platform's **indifference curves**

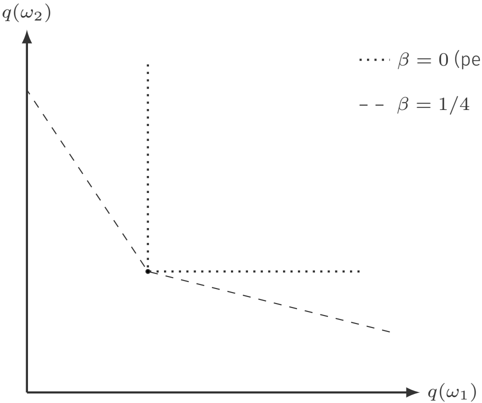


Set $K = 2$

Example (continued): Indifference Curves

Recall that: $u_0(a, \omega) = \beta\Big(\text{seller's profit}\Big) + (1-\beta)\Big(\text{buyer's surplus}\Big)$

An example of the platform's **indifference curves**

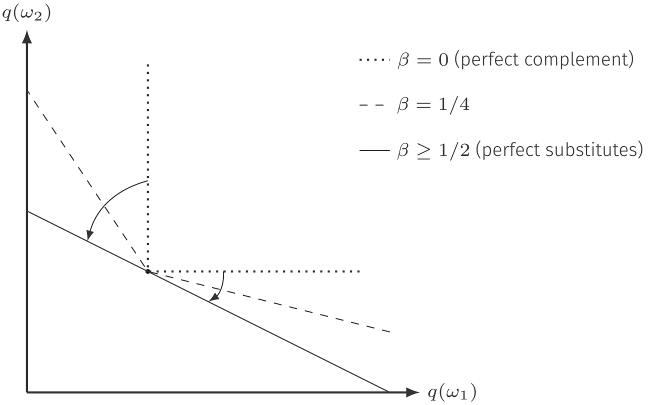


Set $K = 2$

Example (continued): Indifference Curves

Recall that: $u_0(a, \omega) = \beta\Big(\text{seller's profit}\Big) + (1-\beta)\Big(\text{buyer's surplus}\Big)$

An example of the platform's **indifference curves**



Set $K = 2$

Are Records Complement or Substitutes?

General result:

Proposition

Records are perfect substitutes iff for some q it is optimal to fully disclose every record. In this case, full disclosure is optimal for every q.

General result:

Proposition

Records are **perfect substitutes** iff for **some** q it is optimal to **fully disclose** every record. In this case, full disclosure is optimal for every q.

Implications:

1. Optimal database: Well-behaved problem, interior solution $\rightsquigarrow$ Standard demand analysis

Are Records Complement or Substitutes?

General result:

Proposition

Records are **perfect substitutes** iff for **some** q it is optimal to **fully disclose** every record. In this case, full disclosure is optimal for every q.

Implications:

1. Optimal database: Well-behaved problem, interior solution $\rightsquigarrow$ Standard demand analysis
2. When value of merging two datasets is higher than their sum

Are Records Complement or Substitutes?

more records

General result:

Proposition

Records are **perfect substitutes** iff for **some** q it is optimal to **fully disclose** every record. In this case, full disclosure is optimal for every q.

Implications:

1. Optimal database: Well-behaved problem, interior solution $\rightsquigarrow$ Standard demand analysis
2. When value of merging two datasets is higher than their sum
3. How does platform use its data?
   – If detect imperfect substitutability at q, then platform is withholding information from agents

Are Records Complement or Substitutes?

more records

General result:

Proposition

Records are **perfect substitutes** iff for **some** q it is optimal to **fully disclose** every record. In this case, full disclosure is optimal for every q.

Implications:

1. Optimal database: Well-behaved problem, interior solution $\rightsquigarrow$ Standard demand analysis
2. When value of merging two datasets is higher than their sum
3. How does platform use its data?
   - If detect imperfect substitutability at q, then platform is withholding information from agents

Having More Data

What is the platform's WTP for more data?

The colloquial "*having more data*" can indicate two different things:

1. The platform obtains *more* records
2. The platform obtains *better* records

We can study both problems by exploring how $v_q^*(\omega)$ depends on q

Refining Existing Records

Records are often only partially informative about θ and platform can **learn** more about them $\rightsquigarrow$ we call this "**refining**" a record

Questions:

- How do refinements change the value derived from *each* record?
- Do refinements benefit platform *overall* $\rightsquigarrow$ positive WTP?

A classic question from a new perspective

Definition. (link to formalism)

A refinement refines:

- A share $\alpha \in [0,1]$ of the existing records of type ω
- Does so according to rule $\sigma_\omega \in \Delta(\Omega)$
- Does so **independently**

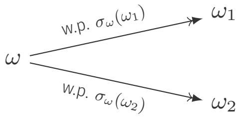
Definition.

(link to formalism)

A refinement refines:

- A share $\alpha \in [0,1]$ of the existing records of type ω
- Does so according to rule $\sigma_\omega \in \Delta(\Omega)$
- Does so **independently**

Example:



(age, $\varnothing$) (age, education)

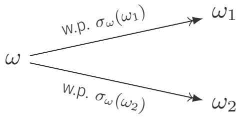
Formalizing Refinements

Definition. (link to formalism)

A refinement refines:

- A share $\alpha \in [0, 1]$ of the existing records of type ω
- Does so according to rule $\sigma_\omega \in \Delta(\Omega)$
- Does so **independently**

Example:



(age, $\varnothing$) (age, education)

Thus, it transforms the original database $q \rightsquigarrow q_\alpha$ such that:

$$q_\alpha(\omega) < q(\omega) \quad \text{and} \quad q_\alpha(\omega') > q(\omega') \quad \forall \omega' \in \operatorname{supp} \sigma_\omega$$

How do refinements change the value derived from *each* record?

How do refinements change the value derived from *each* record?

Corollary:

Consider refining α-share of type-ω records:

Direct Effects. The value of each *refined* record increases:

$$\sum_{\omega' \in \Omega} v^*_{q_\alpha}(\omega')\sigma_\omega(\omega') \geq v^*_q(\omega)$$

Indirect Effects. The value of *unrefined* records affected too:

$$v^*_{q_\alpha}(\omega) \geq v^*_q(\omega) \qquad \text{and} \qquad v^*_{q_\alpha}(\omega') \leq v^*_q(\omega') \qquad \forall \omega' \in \operatorname{supp} \sigma_\omega$$

How do refinements change the value derived from *each* record?

Corollary:

Consider refining α-share of type-ω records:

Direct Effects. The value of each *refined* record increases:

$$\sum_{\omega' \in \Omega} v^*_{q_\alpha}(\omega')\sigma_\omega(\omega') \geq v^*_q(\omega)$$

Indirect Effects. The value of *unrefined* records affected too:

$$v^*_{q_\alpha}(\omega) \geq v^*_q(\omega) \qquad \text{and} \qquad v^*_{q_\alpha}(\omega') \leq v^*_q(\omega') \qquad \forall \omega' \in \operatorname{supp} \sigma_\omega$$

This characterizes some of the possible externalities when disclosing personal data

Given these mixed effects, do refinements benefit platform overall?

Given these mixed effects, do refinements benefit platform overall?

Proposition

The platform's benefit from the refinement is:

- Weakly **positive**, $U^*(q_\alpha) \geq U^*(q)$

Yes. WTP for a refinement is positive

Given these mixed effects, do refinements benefit platform overall?

Proposition

The platform's benefit from the refinement is:

- Weakly **positive**, $U^*(q_\alpha) \geq U^*(q)$

Yes. WTP for a refinement is positive

However, WTP can be 0 even if platform acts on new information (x_q^* changes)

- In sharp contrast with decision problems

Given these mixed effects, do refinements benefit platform overall?

Proposition

The platform's benefit from the refinement is:

- Weakly **positive**, $U^*(q_\alpha) \geq U^*(q)$
- **Zero** for all α iff there is $a \in \operatorname{supp} x_q^*(\cdot|\omega'')$ for $\omega'' = \omega$ & $\omega'' \in \operatorname{supp} \sigma_\omega$

Yes. WTP for a refinement is positive

However, WTP can be 0 even if platform acts on new information (x_q^* changes)

- In sharp contrast with decision problems

WTP for Better Records

Given these mixed effects, do refinements benefit platform overall?

Proposition

The platform's benefit from the refinement is:

- Weakly **positive**, $U^*(q_\alpha) \geq U^*(q)$
- **Zero** for all α iff there is $a \in \operatorname{supp} x_q^*(\cdot|\omega'')$ for $\omega'' = \omega$ & $\omega'' \in \operatorname{supp} \sigma_\omega$
- Marginally **decreasing** in α

Yes. WTP for a refinement is positive

However, WTP can be 0 even if platform acts on new information (x_q^* changes)

- In sharp contrast with decision problems

conclusions

Summary

We show how to compute the **unit value** of a buyer's specific data record

- Uncover novel data **externalities**, specific to **intermediation** problems

  Due to pooling records to withhold information

- Direct payoff gives a biased account of the value of a record

Use our theory to characterize basic properties of the demand for data:

- *"More"* records: demand for records, complements vs substitutes
- *"Better"* records: mixed effects on unit values, overall WTP

Overall, an investigation of the **demand side** of data markets

Next Steps: Privacy

Work in progress: how protecting privacy affects the value of data

Next Steps: Privacy

Work in progress: how protecting privacy affects the value of data

In a richer model:

- Each buyer as an agent and ω as her **private** data
- Buyer can agree to disclose ω to platform if she wants
- Thus, platform has to elicit such data in order to use it

Use + Elicitation = LP problem $\rightsquigarrow$ Same approach as in this paper

Preliminary findings:

- Privacy decreases total value of the database (of course!)
- But it can **increase** the value of some records (redistributive effects)

appendix

How does v_q^* depend on q?

Stability

How does v_q^* depend on q?

v_q^* goes beyond a *marginal* interpretation $\rightsquigarrow$ WTP for discrete changes in q

Proposition (Stability)

There exists finite collection $\{Q_1, \ldots, Q_K\}$ of open sets in $\mathbb{R}_+^\Omega$ s.t.:

- $\bigcup Q_k$ has full measure
- (v_q^*, λ_q^*) is **unique** and **constant** in $q \in Q_k$ for each k

How does v_q^* depend on q?

v_q^* goes beyond a *marginal* interpretation $\rightsquigarrow$ WTP for discrete changes in q

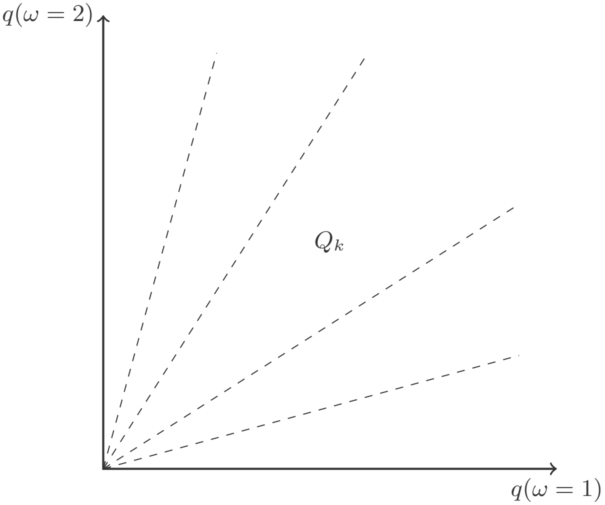
Proposition (Stability)

There exists finite collection $\{Q_1, \ldots, Q_K\}$ of open sets in $\mathbb{R}_+^\Omega$ s.t.:

- $\bigcup Q_k$ has full measure
- (v_q^*, λ_q^*) is **unique** and **constant** in $q \in Q_k$ for each k

Note : v_q^* constant in Q_k even though x_q^* changes

Proof idea : algebraic representation of extreme points & optimality


$q(\omega = 2)$
Q_k
$q(\omega = 1)$

Proposition

For $\beta \leq \frac{1}{2}$,

$$v_q^*(\omega) = \begin{cases} (1-\beta)\omega & \text{if } \omega < a_q \\ \beta a_q + (1-\beta)(\omega - a_q) & \text{if } \omega \geq a_q; \end{cases}$$

Moreover, $t_q^*(\omega) > 0$ for $\omega < a_q$ and $t_q^*(\omega) \leq 0$ for $\omega \geq a_q$

For $\beta \geq \frac{1}{2}$ we have $v_q^*(\omega) = u_q^*(\omega) = \beta\omega$ for all ω

Formalizing Refinements

Let $p_\omega \in \Delta(\Theta)$ be belief about buyer's θ if her record is of type ω

A **refinement** is $\sigma_\omega \in \Delta(\Omega)$ s.t. $\sum_{\omega' \in \Omega} \sigma_\omega(\omega') p_{\omega'} = p_\omega$

Let $p_\omega \in \Delta(\Theta)$ be belief about buyer's θ if her record is of type ω

A **refinement** is $\sigma_\omega \in \Delta(\Omega)$ s.t. $\sum_{\omega' \in \Omega} \sigma_\omega(\omega') p_{\omega'} = p_\omega$

We consider refining multiple records (extensive margin):

- let $\alpha \in [0, 1]$ be **share** of $q(\omega)$
- refine each record **independently** according to σ_ω

Formalizing Refinements

Let $p_\omega \in \Delta(\Theta)$ be belief about buyer's θ if her record is of type ω

A **refinement** is $\sigma_\omega \in \Delta(\Omega)$ s.t. $\sum_{\omega' \in \Omega} \sigma_\omega(\omega') p_{\omega'} = p_\omega$

We consider refining multiple records (extensive margin):

- let $\alpha \in [0, 1]$ be **share** of $q(\omega)$
- refine each record **independently** according to σ_ω

It transforms the original database $q \rightsquigarrow q_\alpha$ such that:

$$q_\alpha(\omega) < q(\omega) \qquad \text{and} \qquad q_\alpha(\omega') > q(\omega') \qquad \forall \omega' \in \operatorname{supp} \sigma_\omega$$