The Value of Data

Simone Galperti
UC San Diego

Aleksandr Levkun
UC San Diego

Jacopo Perego
Columbia University

December 2020

Motivation

Data is essential input in modern economies

Often data collected "for free" absent formal markets

Towards a market for data:

> "*A first and necessary step is getting a quantitative grip on* the value of data. *Things that are not measured are not priced.*" (Posner, Weyl '18)

This paper:

- A theory to assess the value of *datapoints* in a *database*

DATAPOINTS

A **datapoint** is a measurement of the agents' type. Examples:

- In a Buyer-Seller trade: Buyer's valuation
- In an Auction: Bidders' valuations
- Firm and Worker matching: Worker's productivty

Each datapoint characterizes a single economic interaction

A **database** is the set of datapoints

In **Decision Problems**, the value of data is well-understood

- E.g. Seller has data about buyer's valuations and maximizes its profits

value of data	=	direct payoff from best action given the datapoint

Decision vs Mediation Problems

In **Decision Problems**, the value of data is well-understood

- E.g. Seller has data about buyer's valuations and maximizes its profits

value of data = direct payoff from best action given the datapoint

In **Mediation Problems**, the value of data is less understood

- E.g. Platform has data about buyers' valuations. It communicates with seller to maximize some objective

value of data = *direct* payoff from best action given the datapoint + *indirect* payoff from information externalities

Decision vs Mediation Problems

In **Decision Problems**, the value of data is well-understood

- E.g. Seller has data about buyer's valuations and maximizes its profits

value of data = direct payoff from best action given the datapoint

In **Mediation Problems**, the value of data is less understood

- E.g. Platform has data about buyers' valuations. It communicates with seller to maximize some objective

value of data = *direct* payoff from best action given the datapoint + *indirect* payoff from information externalities

We build on a **simple** insight:

- **Data-Value** Problem intimately relates to **Data-Use** Problem
- When "carefully formulated," the two are in a special mathematical relationship: **Duality**

Main Ideas

We build on a **simple** insight:

- **Data-Value** Problem intimately relates to **Data-Use** Problem
- When "carefully formulated," the two are in a special mathematical relationship: **Duality**

Data-Use Problem: Designer chooses mechanism/information using datapoints as inputs

Data-Value Problem: Designer assigns individual value to each datapoint

We build on a **simple** insight:

- **Data-Value** Problem intimately relates to **Data-Use** Problem
- When "carefully formulated," the two are in a special mathematical relationship: **Duality**

Data-Use Problem: Designer chooses mechanism/information using datapoints as inputs

Data-Value Problem: Designer assigns individual value to each datapoint

Plan for today:

1. Formalize the Data-Value Problem and interpret it
2. Characterize these information externalities
3. Use framework to study effects of privacy on value of data

Related Literature

	Our Paper
Mechanism Design. Myerson ('82, '83) ...	— Formulation of data-use problem
Information Design. Kamenica & Gentzkow ('11), Bergemann & Morris ('16,'19) ...	— Subclass of data-use problem
Duality & Correlated Equilibrium. Nau & McCardle ('90), Nau ('92), Hart & Schmeidler ('89), Myerson ('97)	— Duality to characterize CE — Feasible mechanisms for principal
Duality & Bayesian Persuasion. Kolotilin ('18), Dworczak & Martini ('19), Dizdar & Kovac ('19), Dworczak & Kolotilin ('19)	— Dual not as a solution method, but as focus of analysis — Independent question from ID — Games and mechanisms
Markets for Information. Bergemann & Bonatti ('15), Bergmann, Bonatti, Smolin ('18), Posner & Weyl ('18), Bergemann & Bonatti ('19)	— Focus on value of data*points*
Information Privacy. Acquisti, Taylor, Wagman ('16), Ali, Lewis, Vasserman ('20), Bergemann, Bonatti, Gan ('20), Acemoglu, Makhdoumi, Malekian, Ozdaglar, ('20)	— A method for assessing effects of privacy on value of data

an example

An Online Marketplace

Example builds on Bergemann, Brooks, Morris (2015)

Three parties:

- An online **platform** / information designer
- A monopolistic **seller** (mc=0)
- A finite set of potential **buyers** with unit demand

An Online Marketplace

Example builds on Bergemann, Brooks, Morris (2015)

Three parties:

- An online **platform** / information designer
- A monopolistic **seller** (mc=0)
- A finite set of potential **buyers** with unit demand

The platform **owns** a database: a list of ω's, one for each buyer

Datapoint is measurement of buyers' **valuation** for seller's product

Platform sends information to seller about ω, who then charges a price to buyer

The Database and Platform's Problem

Database

Buyer ID	Datapoint
⋮	⋮
123	ω_{123}
124	ω_{124}
⋮	⋮

$\leftarrow$ summary

The Database and Platform's Problem

example

Database

Buyer ID	Datapoint
$\vdots$	$\vdots$
123	ω_{123}
124	ω_{124}
$\vdots$	$\vdots$

← summary

$$\text{Three types of datapoints: } \omega = \begin{cases} 2 & \text{for 60\% of buyers} \\ 1 & \text{for 30\% of buyers} \\ \varnothing & \text{for 10\% of buyers} \end{cases}$$

If $\omega = \varnothing$, buyer has valuation 2 with probability $h \geq \frac{1}{2}$ and 1 otherwise.

Question: What is the value of each datapoint for the platform?

Case 1: Maximize Profits

Suppose platform maximizes **seller's profits**

Case 1: Maximize Profits

Suppose platform maximizes **seller's profits**

It is optimal to fully reveal ω $\rightarrow$ Perfect price discrimination

Case 1: Maximize Profits

Suppose platform maximizes **seller's profits**

It is optimal to fully reveal ω $\rightarrow$ Perfect price discrimination

	s'	s''	s'''
$\omega = 1$	30	0	0
$\omega = \varnothing$	0	10	0
$\omega = 2$	0	0	60

Case 1: Maximize Profits

Suppose platform maximizes **seller's profits**

It is optimal to fully reveal ω $\rightarrow$ Perfect price discrimination

	s'	s''	s'''
$\omega = 1$	30	0	0
$\omega = \varnothing$	0	10	0
$\omega = 2$	0	0	60
a^*	1	2	2

Case 1: Maximize Profits

Suppose platform maximizes **seller's profits**

It is optimal to fully reveal ω $\rightarrow$ Perfect price discrimination

	s'	s''	s'''	$u^*(\omega)$
$\omega = 1$	30	0	0	1
$\omega = \varnothing$	0	10	0	$2h$
$\omega = 2$	0	0	60	2
a^*	1	2	2	

Case 1: Maximize Profits

Suppose platform maximizes **seller's profits**

It is optimal to fully reveal ω $\rightarrow$ Perfect price discrimination

	s'	s''	s'''	$u^*(\omega)$
$\omega = 1$	30	0	0	1
$\omega = \varnothing$	0	10	0	$2h$
$\omega = 2$	0	0	60	2
a^*	1	2	2	

For each datapoint ω, the platform's **direct payoff** is $u^*(\omega)$

Case 1: Maximize Profits

Suppose platform maximizes **seller's profits**

It is optimal to fully reveal ω $\rightarrow$ Perfect price discrimination

	s'	s''	s'''	$u^*(\omega)$
$\omega = 1$	30	0	0	1
$\omega = \varnothing$	0	10	0	$2h$
$\omega = 2$	0	0	60	2
a^*	1	2	2	

For each datapoint ω, the platform's **direct payoff** is $u^*(\omega)$

Question:

- What datapoint is more valuable for platform: $\omega = 1, \omega = \varnothing$, or $\omega = 2$?

Suppose platform maximizes **seller's profits**

It is optimal to fully reveal ω $\rightarrow$ Perfect price discrimination

	s'	s''	s'''	$u^*(\omega)$
$\omega = 1$	30	0	0	1
$\omega = \varnothing$	0	10	0	$2h$
$\omega = 2$	0	0	60	2
a^*	1	2	2	

For each datapoint ω, the platform's **direct payoff** is $u^*(\omega)$

Question:

- What datapoint is more valuable for platform: $\omega = 1$, $\omega = \varnothing$, or $\omega = 2$?

Obs. Decision problem; $v^*(\omega) = u^*(\omega)$; Independent of (Ω, μ)

Case 2: Maximize Surplus

Suppose instead platform maximizes **buyers' surplus**

Suppose instead platform maximizes **buyers' surplus**

An optimal information structure is:

	s'	s''
$\omega = 1$	30	0
$\omega = \varnothing$	10	0
$\omega = 2$	$20(2-h)$	$20(1+h)$

Case 2: Maximize Surplus

Suppose instead platform maximizes **buyers' surplus**

An optimal information structure is:

	s'	s''
$\omega = 1$	30	0
$\omega = \varnothing$	10	0
$\omega = 2$	$20(2-h)$	$20(1+h)$

Case 2: Maximize Surplus

Suppose instead platform maximizes **buyers' surplus**

An optimal information structure is:

	s'	s''
$\omega = 1$	30	0
$\omega = \varnothing$	10	0
$\omega = 2$	$20(2-h)$	$20(1+h)$
a^*	1	2

Case 2: Maximize Surplus

Suppose instead platform maximizes **buyers' surplus**

An optimal information structure is:

	s'	s''	$u^*(\omega)$
$\omega = 1$	30	0	0
$\omega = \varnothing$	10	0	h
$\omega = 2$	$20(2-h)$	$20(1+h)$	$\frac{1}{3}(2-h)$
a^*	1	2	

Case 2: Maximize Surplus

Suppose instead platform maximizes **buyers' surplus**

An optimal information structure is:

	s'	s''	$u^*(\omega)$
$\omega = 1$	30	0	0
$\omega = \varnothing$	10	0	h
$\omega = 2$	$20(2-h)$	$20(1+h)$	$\frac{1}{3}(2-h)$
a^*	1	2	

For each datapoint ω, the platform's **direct payoff** is $u^*(\omega)$

Case 2: Maximize Surplus

Suppose instead platform maximizes **buyers' surplus**

An optimal information structure is:

	s'	s''	$u^*(\omega)$
$\omega = 1$	30	0	0
$\omega = \varnothing$	10	0	h
$\omega = 2$	$20(2-h)$	$20(1+h)$	$\frac{1}{3}(2-h)$
a^*	1	2	

For each datapoint ω, the platform's **direct payoff** is $u^*(\omega)$

Question:

- What datapoint is more valuable for the platform: $\omega = 1$, $\omega = \varnothing$, or $\omega = 2$?

Case 2: Maximize Surplus

Suppose instead platform maximizes **buyers' surplus**

An optimal information structure is:

	s'	s''	$u^*(\omega)$
$\omega = 1$	30	0	0
$\omega = \varnothing$	10	0	h
$\omega = 2$	$20(2-h)$	$20(1+h)$	$\frac{1}{3}(2-h)$
a^*	1	2	

For each datapoint ω, the platform's **direct payoff** is $u^*(\omega)$

Question:

- What datapoint is more valuable for the platform: $\omega = 1$, $\omega = \varnothing$, or $\omega = 2$?

Case 2: Maximize Surplus

The **most valuable** datapoint is the one yielding the **lowest** direct payoff

	direct payoff $u^*(\omega)$	value of data $v^*(\omega)$
$\omega = 1$	0	1
$\omega = \varnothing$	h	$1 - h$
$\omega = 2$	$\frac{1}{3}(2-h)$	0

Indeed, new $\omega = 1 \rightarrow$ Move old $\omega = 2$ from s'' to $s' \rightarrow$ Earn surplus of 1

Case 2: Maximize Surplus

The **most valuable** datapoint is the one yielding the **lowest** direct payoff

	direct payoff $u^*(\omega)$	value of data $v^*(\omega)$
$\omega = 1$	0	1
$\omega = \varnothing$	h	$1 - h$
$\omega = 2$	$\frac{1}{3}(2 - h)$	0

Indeed, new $\omega = 1 \to$ Move old $\omega = 2$ from s'' to $s' \to$ Earn surplus of 1

Unlike in Decision Problems,

- Direct payoff u^* is misleading measure of value
- **Conflict of interest** leads to pooling $\rightsquigarrow$ Information Externalities
- What is v^*? How to compute it? What are its properties?

model

The Model

Parties: Designer $i = 0$, Agents $i \in I = \{1, \ldots, n\}$

Let $\Omega = \{\omega, \omega', \ldots, \omega''\}$ be a finite set

Party i privately controls action $a_i \in A_i$: $\quad A = A_0 \times A_1 \times \ldots \times A_n$

Payoff function of party i: $u_i : A \times \Omega \to \mathbb{R}$

Database is (Ω, μ), where

– $\mu(\omega)$ is stock of ω-datapoints in database ($\sim$ as a share of total)

Parties: Designer $i = 0$, Agents $i \in I = \{1, \ldots, n\}$

Let $\Omega = \{\omega, \omega', \ldots, \omega''\}$ be a finite set

Party i privately controls action $a_i \in A_i$: $\quad A = A_0 \times A_1 \times \ldots \times A_n$

Payoff function of party i: $u_i : A \times \Omega \to \mathbb{R}$

Database is (Ω, μ), where

- $\mu(\omega)$ is stock of ω-datapoints in database ($\sim$ as a share of total)

Discussion:

(1) A "frequentist" interpretation of (Ω, μ)
(2) More primitive states

We start with plain-vanilla **Information Design**

1. Designer privately observes each datapoint ω (omniscience)
2. $|A_0| = 1$ (no mech design)
3. She chooses information structure $\pi : \Omega \to \Delta(S)$ (commitment)
4. Agents observe signals and play BNE (implementation)

Later today, we drop 1 and 2

value of data

Data-Use Problem

As usual, wlog to focus on "recommendation mechanisms" $x : \Omega \to \Delta(A)$ that satisfy

- **Obedience**: it is optimal for each agent to follow recommended a_i

The **Data-Use** problem involves:

- Inputs = Datapoints ω from database (Ω, μ)
- Production Technologies = Obedient Mechanisms x
- Objective = $u_0(a, \omega)$

Data-Use Problem

Problem $\mathcal{U}$

$$U^* = \max_x \quad \sum_{\omega, a} u_0(a, \omega) x(a|\omega) \mu(\omega)$$

s.t. for all i, a_i, and a_i'

$$\sum_{\omega, a_{-i}} \Big(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \Big) x(a_i, a_{-i}|\omega) \mu(\omega) \geq 0$$

Data-Use Problem

Problem $\mathcal{U}$

$$U^* = \max_x \sum_{\omega, a} u_0(a, \omega) x(a|\omega) \mu(\omega)$$

s.t. for all i, a_i, and a_i'

$$\sum_{\omega, a_{-i}} \Big(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \Big) x(a_i, a_{-i}|\omega) \mu(\omega) \geq 0$$

Definition. The **direct payoff** of datapoint ω is

$$u^*(\omega) = \sum_a u_0(a, \omega) x^*(a|\omega)$$

Problem $\mathcal{U}$

$$U^* = \max_x \sum_{\omega, a} u_0(a, \omega) x(a|\omega) \mu(\omega)$$

s.t. for all i, a_i, and a_i'

$$\sum_{\omega, a_{-i}} \Big(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \Big) x(a_i, a_{-i}|\omega) \mu(\omega) \geq 0$$

Definition. The **direct payoff** of datapoint ω is

$$u^*(\omega) = \sum_a u_0(a, \omega) x^*(a|\omega)$$

Data-Value Problem consists of finding

$$v : \Omega \to \mathbb{R}$$

such that $v(\omega)$ is the value each ω-datapoint generates for designer

Data-Value Problem

Designer chooses for each agent i and a_i

$$\ell_i(\cdot|a_i) \in \Delta(A_i) \qquad \text{and} \qquad q_i(a_i) \in \mathbb{R}_{++}$$

Problem $\mathcal{V}$

$$V^* = \min_{\ell,q} \quad \sum_{\omega} v(\omega)\mu(\omega)$$

s.t. for all ω,

$$\boxed{v(\omega) = \max_{a \in A} \left\{ u_0(a,\omega) + \sum_i T_{\ell_i,q_i}(a,\omega) \right\}}$$

$$T_{\ell_i,q_i}(a,\omega) = q_i(a_i) \sum_{a_i' \in A_i} \Big(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \Big) \ell_i(a_i'|a_i)$$

Data-Value Problem

Designer chooses for each agent i and a_i

$$\ell_i(\cdot|a_i) \in \Delta(A_i) \qquad \text{and} \qquad q_i(a_i) \in \mathbb{R}_{++}$$

Problem $\mathcal{V}$

$$V^* = \min_{\ell,q} \quad \sum_{\omega} v(\omega)\mu(\omega)$$

s.t. for all ω,

$$\boxed{v(\omega) = \max_{a \in A} \left\{ u_0(a,\omega) + \sum_i T_{\ell_i,q_i}(a,\omega) \right\}}$$

$$T_{\ell_i,q_i}(a,\omega) = q_i(a_i) \sum_{a_i' \in A_i} \Big(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \Big) \ell_i(a_i'|a_i)$$

Why is $\mathcal{V}$ the "right" Data-Value problem?

Lemma

Problem $\mathcal{V}$ is equivalent to the **dual** of Problem $\mathcal{U}$. Also,

$$\sum_{\omega} \underbrace{v^*(\omega)}_{\substack{\text{value of} \\ \text{datapoint}}} \mu(\omega) = \underbrace{U^*}_{\substack{\text{value of} \\ \text{database}}}$$

Dual Relationship

Why is $\mathcal{V}$ the "right" Data-Value problem?

Lemma

Problem $\mathcal{V}$ is equivalent to the **dual** of Problem $\mathcal{U}$. Also,

$$\sum_{\omega} \underbrace{v^*(\omega)}_{\text{value of datapoint}} \mu(\omega) = \underbrace{U^*}_{\text{value of database}}$$

- $v(\omega)$ variables corresponds to $\mathcal{U}$-constraints

$$\sum_{a} x(a|\omega) \qquad = 1 \qquad \forall \omega$$

Dual Relationship

Why is $\mathcal{V}$ the "right" Data-Value problem?

Lemma

Problem $\mathcal{V}$ is equivalent to the **dual** of Problem $\mathcal{U}$. Also,

$$\sum_{\omega} \underbrace{v^*(\omega)}_{\text{value of datapoint}} \mu(\omega) = \underbrace{U^*}_{\text{value of database}}$$

- $v(\omega)$ variables corresponds to $\mathcal{U}$-constraints

$$\sum_{a} x(a|\omega)\mu(\omega) = 1\mu(\omega) \qquad \forall \omega$$

Why is $\mathcal{V}$ the "right" Data-Value problem?

Lemma

Problem $\mathcal{V}$ is equivalent to the **dual** of Problem $\mathcal{U}$. Also,

$$\sum_\omega \underbrace{v^*(\omega)}_{\substack{\text{value of}\\ \text{datapoint}}} \mu(\omega) = \underbrace{U^*}_{\substack{\text{value of}\\ \text{database}}}$$

- $v(\omega)$ variables corresponds to $\mathcal{U}$-constraints

$$\sum_a \chi(a, \omega) = \mu(\omega) \qquad \forall \omega$$

Why is $\mathcal{V}$ the “right” Data-Value problem?

Lemma

Problem $\mathcal{V}$ is equivalent to the **dual** of Problem $\mathcal{U}$. Also,

$$\sum_{\omega} \underbrace{v^*(\omega)}_{\text{value of datapoint}} \mu(\omega) = \underbrace{U^*}_{\text{value of database}}$$

- $v(\omega)$ variables corresponds to $\mathcal{U}$-constraints

$$\sum_{a} \chi(a, \omega) = \mu(\omega) \qquad \forall \omega$$

- $v(\omega)$ captures shadow **value** of a datapoint ω to $\mu(\mu)$

Dual Relationship

Why is $\mathcal{V}$ the "right" Data-Value problem?

Lemma

Problem $\mathcal{V}$ is equivalent to the **dual** of Problem $\mathcal{U}$. Also,

$$\sum_{\omega} \underbrace{v^*(\omega)}_{\text{value of datapoint}} \mu(\omega) = \underbrace{U^*}_{\text{value of database}}$$

- $v(\omega)$ variables corresponds to $\mathcal{U}$-constraints

$$\sum_{a} \chi(a, \omega) = \mu(\omega) \qquad \forall \omega$$

- $v(\omega)$ captures shadow **value** of a datapoint ω to $\mu(\mu)$
- Values v^* is generically unique with respect to μ

Two interpretations for $v^*(\omega)$:

- $v(\omega)$ reflects designer's WTP for **marginal** datapoint ω given (Ω, μ)
- $v(\omega)$ assess "fair" compensation for individual data providers

Why focus on single datapoints vs database?

- Guide allocation of scarce resources: e.g. user retention or acquisition
- $v(\omega)$ as the **demand curve** for data
- Efficiency benchmark for markets for data

information externalities

Externalities Between Datapoints

In $\mathcal{U}$, designer pools datapoints to produce information

Direct payoff $u^*(\omega)$ depends on other ω' that are pooled with ω

Those each ω' generates **externalities** for other ω's

We can characterize quantifies these externalities combining $\mathcal{V}$ and $\mathcal{U}$

Definition. The **indirect payoff** of datapoint ω is

$$T^*(\omega) = \sum_{i,a} T_{\ell_i^*, q_i^*}(\omega, a) x^*(a|\omega)$$

Definition. The **indirect payoff** of datapoint ω is

$$T^*(\omega) = \sum_{i,a} T_{\ell_i^*, q_i^*}(\omega, a) x^*(a|\omega)$$

Proposition

Let x^* and (ℓ^*, q^*) be optimal for $\mathcal{U}$ and $\mathcal{V}$. Then

$$\underbrace{v^*(\omega)}_{\text{value}} = \underbrace{u^*(\omega)}_{\text{direct payoff}} + \underbrace{T^*(\omega)}_{\text{indirect payoff}}$$

Moreover,

$$T^*(\omega) > 0 \text{ for some } \omega \quad \iff \quad T^*(\omega') < 0 \text{ for some } \omega'$$

Why transfer value from ω-datapoints to ω'-datapoints?

Proposition

If $T^*(\omega) < 0$, then there is $a \in A$ such that

- $x^*(a|\omega) > 0$
- $u_0(a, \omega) > \max_{y \in CE(G_\omega)} \sum_a u_0(a, \omega) y(a)$

Intuition: ω pooled with some other ω' to induce outcomes that are otherwise unachievable if ω was common knowledge

Sufficient condition for no externalities

Proposition

If $x^*(\cdot|\omega) \in CE(G_\omega)$ for all ω, then $T^*(\omega) = 0$.

- No conflicts of interest leads to no pooling, hence no externalities
- $T^* = 0 \quad \Rightarrow \quad v^* = u^*$

When there are conflicts of interest between designer and agents:

- Partial information, externalities $T^* \neq 0$, missed by u^*

back to example

Seller's Payoff:

$u_1(a,\omega)$	$a=1$	$a=2$
$\omega=1$	1	0
$\omega=\varnothing$	1	$2h$
$\omega=2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a,\omega)$	$a=1$	$a=2$
$\omega=1$	0	0
$\omega=\varnothing$	h	0
$\omega=2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell,q} \quad \sum_{\omega} v(\omega)\mu(\omega)$$

s.t. for all ω,

$$v(\omega) = \max_{a\in A} \left\{ u_0(a,\omega) + T_{\ell,q}(a,\omega) \right\}$$

Computing The Value of Data

Seller's Payoff:

$u_1(a,\omega)$	$a=1$	$a=2$
$\omega=1$	1	0
$\omega=\varnothing$	1	$2h$
$\omega=2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a,\omega)$	$a=1$	$a=2$
$\omega=1$	0	0
$\omega=\varnothing$	h	0
$\omega=2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell,q} \quad \sum_{\omega} v(\omega)\mu(\omega)$$

s.t. for all ω,

$$v(\omega) = \max_{a\in A}\left\{u_0(a,\omega) + T_{\ell,q}(a,\omega)\right\}$$

Computing The Value of Data

Seller's Payoff:

$u_1(a, \omega)$	$a = 1$	$a = 2$
$\omega = 1$	1	0
$\omega = \varnothing$	1	$2h$
$\omega = 2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a, \omega)$	$a = 1$	$a = 2$
$\omega = 1$	0	0
$\omega = \varnothing$	h	0
$\omega = 2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell, q} \quad v(1)\mu(1) + v(\varnothing)\mu(\varnothing) + v(2)\mu(2)$$

s.t. for all ω,

$$v(\omega) = \max_{a \in A} \left\{ u_0(a, \omega) + T_{\ell,q}(a, \omega) \right\}$$

Computing The Value of Data

Seller's Payoff:

$u_1(a,\omega)$	$a=1$	$a=2$
$\omega=1$	1	0
$\omega=\varnothing$	1	$2h$
$\omega=2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a,\omega)$	$a=1$	$a=2$
$\omega=1$	0	0
$\omega=\varnothing$	h	0
$\omega=2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell,q} \quad v(1)\mu(1)+v(\varnothing)\mu(\varnothing)+v(2)\mu(2)$$

s.t.

$$v(1)=\max\left\{u_0(1,1)+T_{\ell,q}(1,1),\,u_0(2,1)+T_{\ell,q}(2,1)\right\}$$

$$v(\varnothing)=\max\left\{u_0(1,\varnothing)+T_{\ell,q}(1,\varnothing),\,u_0(2,\varnothing)+T_{\ell,q}(2,\varnothing)\right\}$$

$$v(2)=\max\left\{u_0(1,2)+T_{\ell,q}(1,2),\,u_0(2,2)+T_{\ell,q}(2,2)\right\}$$

Seller's Payoff:

$u_1(a,\omega)$	$a=1$	$a=2$
$\omega=1$	1	0
$\omega=\varnothing$	1	$2h$
$\omega=2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a,\omega)$	$a=1$	$a=2$
$\omega=1$	0	0
$\omega=\varnothing$	h	0
$\omega=2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell,q} \quad v(1)\mu(1)+v(\varnothing)\mu(\varnothing)+v(2)\mu(2)$$

$$\text{s.t.} \quad v(1)=\max\left\{q(1)\ell(2|1),-q(2)\ell(2|1)\right\}$$

$$v(\varnothing)=\max\left\{h+(1-2h)q(1)\ell(2|1),(2h-1)q(2)\ell(1|2)\right\}$$

$$v(2)=\max\left\{1-q(1)\ell(2|1),q(2)\ell(1|2)\right\}$$

Seller's Payoff:

$u_1(a,\omega)$	$a=1$	$a=2$
$\omega=1$	1	0
$\omega=\varnothing$	1	$2h$
$\omega=2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a,\omega)$	$a=1$	$a=2$
$\omega=1$	0	0
$\omega=\varnothing$	h	0
$\omega=2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell,q} \quad v(1)\mu(1)+v(\varnothing)\mu(\varnothing)+v(2)\mu(2)$$

$$\text{s.t.} \quad v(1)=\max\left\{q(1)\ell(2|1),-q(2)\ell(2|1)\right\}=q(1)\ell(2|1)$$

$$v(\varnothing)=\max\left\{h+(1-2h)q(1)\ell(2|1),(2h-1)q(2)\ell(1|2)\right\}$$

$$v(2)=\max\left\{1-q(1)\ell(2|1),q(2)\ell(1|2)\right\}$$

Computing The Value of Data

Seller's Payoff:

$u_1(a,\omega)$	$a=1$	$a=2$
$\omega=1$	1	0
$\omega=\varnothing$	1	$2h$
$\omega=2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a,\omega)$	$a=1$	$a=2$
$\omega=1$	0	0
$\omega=\varnothing$	h	0
$\omega=2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell,q} \quad v(1)\mu(1)+v(\varnothing)\mu(\varnothing)+v(2)\mu(2)$$

$$\text{s.t.} \quad v(1)=q(1)\ell(2|1)$$

$$v(\varnothing)=\max\left\{h+(1-2h)q(1)\ell(2|1),(2h-1)q(2)\ell(1|2)\right\}$$

$$v(2)=\max\left\{1-q(1)\ell(2|1),q(2)\ell(1|2)\right\}$$

Seller's Payoff:

$u_1(a,\omega)$	$a=1$	$a=2$
$\omega=1$	1	0
$\omega=\varnothing$	1	$2h$
$\omega=2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a,\omega)$	$a=1$	$a=2$
$\omega=1$	0	0
$\omega=\varnothing$	h	0
$\omega=2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell,q} \quad v(1)\mu(1)+v(\varnothing)\mu(\varnothing)+v(2)\mu(2)$$

$$\text{s.t.} \quad v(1)=q(1)\ell(2|1)$$

$$v(\varnothing)=\max\left\{h+(1-2h)q(1)\ell(2|1),\,(2h-1)q(2)\ell(1|2)\right\}$$

$$v(2)=\max\left\{1-q(1)\ell(2|1),\,q(2)\ell(1|2)\right\}$$

Computing The Value of Data

Seller's Payoff:

$u_1(a,\omega)$	$a=1$	$a=2$
$\omega=1$	1	0
$\omega=\varnothing$	1	$2h$
$\omega=2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a,\omega)$	$a=1$	$a=2$
$\omega=1$	0	0
$\omega=\varnothing$	h	0
$\omega=2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell,q} \quad v(1)\mu(1)+v(\varnothing)\mu(\varnothing)+v(2)\mu(2)$$

$$\text{s.t.} \quad v(1)=q(1)\ell(2|1)$$

$$v(\varnothing)=\max\left\{h+(1-2h)q(1)\ell(2|1),0\right\}$$

$$v(2)=\max\left\{1-q(1)\ell(2|1),0\right\}$$

Computing The Value of Data

Seller's Payoff:

$u_1(a,\omega)$	$a=1$	$a=2$
$\omega=1$	1	0
$\omega=\varnothing$	1	$2h$
$\omega=2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a,\omega)$	$a=1$	$a=2$
$\omega=1$	0	0
$\omega=\varnothing$	h	0
$\omega=2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell,q} \quad v(1)\mu(1)+v(\varnothing)\mu(\varnothing)+v(2)\mu(2)$$

$$\text{s.t.} \quad v(1)=q(1)\ell(2|1)$$

$$v(\varnothing)=\max\left\{h+(1-2h)q(1)\ell(2|1),0\right\}$$

$$v(2)=\max\left\{1-q(1)\ell(2|1),0\right\}$$

Since $\mu(2)>\frac{1}{2}$, optimal to set $q^*(1)\ell^*(2|1)=1$

Seller's Payoff:

$u_1(a,\omega)$	$a=1$	$a=2$
$\omega=1$	1	0
$\omega=\varnothing$	1	$2h$
$\omega=2$	1	2

Designer's Payoff = Buyer's surplus:

$u_0(a,\omega)$	$a=1$	$a=2$
$\omega=1$	0	0
$\omega=\varnothing$	h	0
$\omega=2$	1	0

Data-value problem (seller is the only agent)

$$\min_{\ell,q} \quad v(1)\mu(1)+v(\varnothing)\mu(\varnothing)+v(2)\mu(2)$$

s.t. $v(1)=1$

$$v(\varnothing)=1-h$$

$$v(2)=0$$

Since $\mu(2)>\frac{1}{2}$, optimal to set $q^*(1)\ell^*(2|1)=1$

The values $v^*(\omega)$ as a guide for the acquisition of **new data**

1. Data about Existing Buyers

Suppose existing buyer with $\omega = \varnothing$ wants to **sell her data** to platform

Platform's WTP is: $$(1-h)v^*(1) + hv^*(2) - v(\varnothing)$$

The values $v^*(\omega)$ as a guide for the acquisition of **new data**

1. Data about Existing Buyers

Suppose existing buyer with $\omega = \varnothing$ wants to **sell her data** to platform

Platform's WTP is: $(1-h)v^*(1) + hv^*(2) - v(\varnothing)$

For all $h \in [0,1]$, we find that:

- Platform is **unwilling to pay** to disclose $\varnothing$
- Even if platform acts on the "realization" of $\varnothing$ (i.e. $x^\star$ changes)
- This counters our intuition from Decision Problems

The values $v^*(\omega)$ as a guide for the acquisition of **new data**

The values $v^*(\omega)$ as a guide for the acquisition of **new data**

2. Data about New Buyers

Suppose a prospective buyer has valuation 2 wp $h' \in [0,1]$

Platform's WTP is: $$(1-h')v^*(1) + h'v^*(2)$$

v^* is useful to "price" buyers whose datapoints do not exist in database

Discussion. The stability of v^* in μ

what drives v^*

Towards an independent interpretation of $\mathcal{V}$ to understand what drives v^*

Towards an independent interpretation of $\mathcal{V}$ to understand what drives v^*

Towards an independent interpretation of $\mathcal{V}$ to understand what drives v^*

Data-Value Problem:

$$\begin{aligned} \min_{\ell, q} \quad & \sum_{\omega} v(\omega)\mu(\omega) \\ \text{s.t.} \quad & v(\omega) = \max_{a \in A} \left\{ u_0(a,\omega) + \sum_i T_{\ell_i, q_i}(a,\omega) \right\} \quad \forall \omega \end{aligned}$$

But what are ℓ and q?

Fix (a, ω) and recall $q_i(a_i) \in \mathbb{R}_{++}$, $\ell_i(\cdot|a_i) \in \Delta(A_i)$, and

$$T_{\ell_i, q_i}(a, \omega) = q_i(a_i) \sum_{a_i' \in A_i} \Big(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \Big) \ell_i(a_i'|a_i)$$

Dual Side of Information Design

Fix (a, ω) and recall $q_i(a_i) \in \mathbb{R}_{++}$, $\ell_i(\cdot|a_i) \in \Delta(A_i)$, and

$$T_{\ell_i,q_i}(a,\omega) = q_i(a_i) \sum_{a_i' \in A_i} \Big(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \Big) \ell_i(a_i'|a_i)$$

Principal designs **gambles** against agents **contingent** on (a, ω)

- (ℓ_i, q_i) family of gambles (lottery & stake) contingent on a_i
- given (a, ω), $\ell_i(?|a_i)$ yields **prize** $u_i(a_i, a_{-i}, \omega) - u_i(?, a_{-i}, \omega)$

Dual Side of Information Design

Fix (a,ω) and recall $q_i(a_i) \in \mathbb{R}_{++}$, $\ell_i(\cdot|a_i) \in \Delta(A_i)$, and

$$T_{\ell_i,q_i}(a,\omega) = q_i(a_i) \sum_{a_i' \in A_i} \Big(u_i(a_i,a_{-i},\omega) - u_i(a_i',a_{-i},\omega)\Big)\ell_i(a_i'|a_i)$$

Principal designs **gambles** against agents **contingent** on (a,ω)

- (ℓ_i, q_i) family of gambles (lottery & stake) contingent on a_i
- given (a,ω), $\ell_i(?|a_i)$ yields **prize** $u_i(a_i,a_{-i},\omega) - u_i(?,a_{-i},\omega)$
- designer **wins** iff $u_i(a_i,a_{-i},\omega) < u_i(a_i',a_{-i},\omega)$
  $\leftrightarrow$ Had i known (a_{-i},ω), he would have preferred $a_i' \neq a_i$ (**regret**)

Fix (a, ω) and recall $q_i(a_i) \in \mathbb{R}_{++}$, $\ell_i(\cdot|a_i) \in \Delta(A_i)$, and

$$T_{\ell_i,q_i}(a,\omega) = q_i(a_i) \sum_{a_i' \in A_i} \Big(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \Big) \ell_i(a_i'|a_i)$$

Principal designs **gambles** against agents **contingent** on (a, ω)

- (ℓ_i, q_i) family of gambles (lottery & stake) contingent on a_i
- given (a, ω), $\ell_i(?|a_i)$ yields **prize** $u_i(a_i, a_{-i}, \omega) - u_i(?, a_{-i}, \omega)$
- designer **wins** iff $u_i(a_i, a_{-i}, \omega) < u_i(a_i', a_{-i}, \omega)$
  $\leftrightarrow$ Had i known (a_{-i}, ω), he would have preferred $a_i' \neq a_i$ (**regret**)

$v^*(\omega)$ is lower when agents are tricked into choosing actions they regret ex post

$\min_{\ell,q} \sum p(\omega)\mu(\omega)$ $\rightsquigarrow$ designer wants to win gambles as much as possible

$\min_{\ell,q} \sum p(\omega)\mu(\omega) \rightsquigarrow$ designer wants to win gambles as much as possible

Constraint 1: Limited Flexibility

Gambles against i can be tailored only to a_i, but not (a_{-i}, ω)

$\rightsquigarrow$ using (ℓ_i, q_i) to lower $p(\omega)$ may cause $p(\omega')$ to go up

Structure of Gambles

$\min_{\ell,q} \sum p(\omega)\mu(\omega)$ $\rightsquigarrow$ designer wants to win gambles as much as possible

Constraint 1: Limited Flexibility

Gambles against i can be tailored only to a_i, but not (a_{-i}, ω)

$\rightsquigarrow$ using (ℓ_i, q_i) to lower $p(\omega)$ may cause $p(\omega')$ to go up

Constraint 2: Agents' Joint Rationality (Nau '92)

Proposition

For every* (ℓ, q), if $\sum_i T_{\ell_i,q_i}(a,\omega) < 0$ for (a,ω), there must exist (a',ω') such that $\sum_i T_{\ell_i,q_i}(a',\omega') > 0$

Winning less important for relatively scarce datapoints (low μ) $\rightsquigarrow$ higher value, downward-sloping "demand"

$\min_{\ell,q} \sum p(\omega)\mu(\omega) \rightsquigarrow$ designer wants to win gambles as much as possible

Constraint 1: Limited Flexibility

Gambles against i can be tailored only to a_i, but not (a_{-i}, ω)

$\rightsquigarrow$ using (ℓ_i, q_i) to lower $p(\omega)$ may cause $p(\omega')$ to go up

Constraint 2: Agents' Joint Rationality (Nau '92)

Proposition

For every* (ℓ, q), if $\sum_i T_{\ell_i,q_i}(a,\omega) < 0$ for (a,ω), there must exist (a',ω') such that $\sum_i T_{\ell_i,q_i}(a',\omega') > 0$

Winning less important for relatively scarce datapoints (low μ) $\rightsquigarrow$ higher value, downward-sloping "demand"

general analysis

General Model

Our analysis extends to larger class of data-usage problems

Who has the data? So far, principal was omniscient

- More realistically, each party has **private** payoff-relevant data
- Principal has to **elicit** these data to use them

General Model

Our analysis extends to larger class of data-usage problems

Who has the data? So far, principal was omniscient

- More realistically, each party has **private** payoff-relevant data
- Principal has to **elicit** these data to use them

What can principal do with data? So far, only information

- More generally, principal could also take **own actions**

General Model

Our analysis extends to larger class of data-usage problems

Who has the data? So far, principal was omniscient

- More realistically, each party has **private** payoff-relevant data
- Principal has to **elicit** these data to use them

What can principal do with data? So far, only information

- More generally, principal could also take **own actions**

Key: formulate data usage as "Bayes incentive" problem (Myerson '83, '84)
$\rightsquigarrow$ dual = data-value problem with similar structure

Why?

Examples:

- **Online Marketplace**: Both *Platform* and competing *Firms* have private data about demand
- **Auctions**: *Bidders* have data about own valuation of item
- **Navigation System**: *App* has data about traffic, *Drivers* have data about traffic and destinations

Constraint of incentive-compatible elicitation seems useful tool to study how value of data is affected by privacy protection

Agents voluntarily provide private data depending on how designer commits to using them

Privacy and the Value of Data

Constraint of incentive-compatible elicitation seems useful tool to study how value of data is affected by **privacy protection**

Agents **voluntarily** provide private data depending on how designer commits to using them

Immediate: privacy protection **decreases** overall value of any database, U^*

However, some datapoints can become **more valuable** under privacy (information externalities)

Classic Mech Design: Principal is revenue-maximizing auctioneer

Each auction:

- one homogeneous item
- two agents/bidders, independent valuations, $\omega_i \sim U[0,1]$

Question: how much value does each (ω_1, ω_2)-auction generate?

Classic Example: IPV Auctions

Classic Mech Design: Principal is revenue-maximizing auctioneer

Each auction:

- one homogeneous item
- two agents/bidders, independent valuations, $\omega_i \sim U[0,1]$

Question: how much value does each (ω_1, ω_2)-auction generate?

Solving **data-value** problem, we find

$$v^*(\omega_1, \omega_2) = \max\left\{0, \omega_1 - (1-\omega_1), \omega_2 - (1-\omega_2)\right\}$$

where $\omega_i - (1-\omega_i)$ is bidder i's **virtual valuation**

Classic Example: IPV Auctions

Classic Mech Design: Principal is revenue-maximizing auctioneer

Each auction:

- one homogeneous item
- two agents/bidders, independent valuations, $\omega_i \sim U[0,1]$

Question: how much value does each (ω_1, ω_2)-auction generate?

Solving **data-value** problem, we find

$$v^*(\omega_1, \omega_2) = \max\Big\{0, \omega_1 - (1-\omega_1), \omega_2 - (1-\omega_2)\Big\}$$

where $\omega_i - (1-\omega_i)$ is bidder i's **virtual valuation**

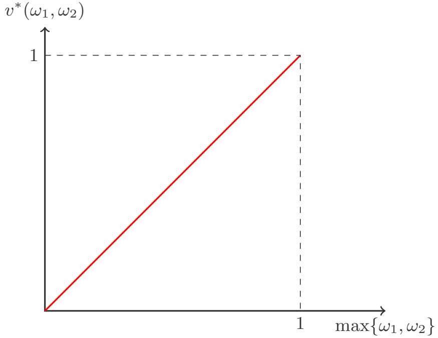
A sanity check: marginal revenues for monopolistic seller

These values incorporate the difficulty of eliciting private data

The Effect of Privacy

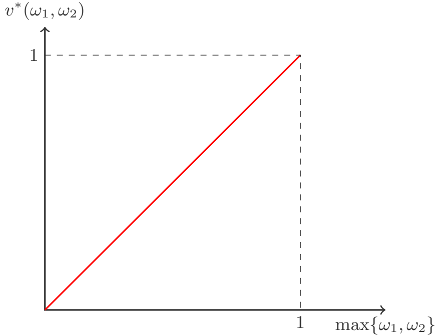
These values incorporate the difficulty of eliciting private data

Red: Scenario where auctioneer knows bidders valuations ω



These values incorporate the difficulty of eliciting private data

Red: Scenario where auctioneer knows bidders valuations ω

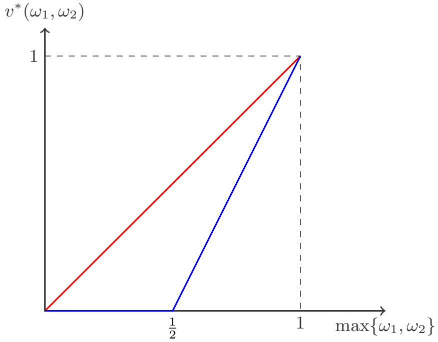


Gap reflects information rents

These values incorporate the difficulty of eliciting private data

Red: Scenario where auctioneer knows bidders valuations ω

Blue: The real auction



Gap reflects information rents

summary

Summary

A theory of how to assess **value of data** in mediation problems

Central Insight: Exploit duality between

- Data-Usage problem = mechanism+information design problem
- Data-Value problem = contingent gambling against the agents

Direct payoff is a misleading measure of value for mediation problems

We characterized *information externalities* across datapoints

A method to assess effects of *privacy* protection on value of data

Related Literature

	Our Paper
Mechanism Design. Myerson ('82, '83) ...	— Formulation of data-use problem
Information Design. Kamenica & Gentzkow ('11), Bergemann & Morris ('16,'19) ...	— Subclass of data-use problem
Duality & Correlated Equilibrium. Nau & McCardle ('90), Nau ('92), Hart & Schmeidler ('89), Myerson ('97)	— Duality to characterize CE — Feasible mechanisms for principal
Duality & Bayesian Persuasion. Kolotilin ('18), Dworczak & Martini ('19), Dizdar & Kovac ('19), Dworczak & Kolotilin ('19)	— Dual not as a solution method, but as focus of analysis — Independent question from ID — Games and mechanisms
Markets for Information. Bergemann & Bonatti ('15), Bergmann, Bonatti, Smolin ('18), Posner & Weyl ('18), Bergemann & Bonatti ('19)	— Focus on value of data*points*
Information Privacy. Acquisti, Taylor, Wagman ('16), Ali, Lewis, Vasserman ('20), Bergemann, Bonatti, Gan ('20), Acemoglu, Makhdoumi, Malekian, Ozdaglar, ('20)	— A method for assessing effects of privacy on value of data