

# Selling Training Data (Preliminary)

Jingmin Huang, Wei Zhao and Renjie Zhong\*

Renmin University of China

Berkeley Theory Lunch

# Selling Training Data

- Consider data buyers purchase supplementary data from a monopolistic data broker to train its predictive model.
- Data buyers may have baseline datasets either collected by themselves or from other external sources.
- Private datasets impact evaluation of supplementary datasets by:
  - 1 altering buyers' outside option
  - 2 affecting the way how supplementary datasets are merged (with private datasets) in statistical decision making
- Our question: what is the **optimal data selling mechanism** for a **screening problem** with **private datasets** as buyer's type.

# Timeline

- 1 The data broker posts a mechanism  $\mathcal{M} = \{\mathcal{E}, t\}$ 
  - 1 a collection of datasets (information structures)  $\mathcal{E}$
  - 2 associated tariff  $t : \mathcal{E} \rightarrow \mathbb{R}_+$
- 2 The data buyer purchases the supplementary dataset  $E$  and pays the price  $t(E)$ . Then he trains his predictive model by merging his private dataset  $E^P$  and  $E$ .
- 3 The true state  $\omega$  is realized.
- 4 The buyer employs his predictive model to make predictions and chooses action to maximize his expected payoff.

# Statistical Decision Making

a simplified cute model for this talk (hypothesis testing):

- two states :  $\omega_1$  (null hypothesis),  $\omega_2$  (alternative hypothesis), prior:  $\mu_0 = (\frac{1}{2}, \frac{1}{2})$
- hypothesis test: binary action  $\{a_1, a_2\}$  and payoff  $u(a_i, \omega_j) = 1_{i=j}$  (correct identification)

two states for tractability: the likelihood ratio between states  $\mu(\omega_1)/\mu(\omega_2)$  can be ranked  
others can be generalized (finite actions, finite signals, arbitrary payoff matrix, any prior now)

# Private Data

- private experiment: two signals  $s'_1$  (acceptance),  $s'_2$  (rejection)
- private type:  $(\alpha, \beta)$ , Type I error  $\alpha = \Pr(s'_2|\omega_1)\mu(\omega_1)$ , Type II error  $\beta = \Pr(s'_1|\omega_2)\mu(\omega_2)$
- buyer with high-quality private dataset is low type (data quality preference  $\alpha + \beta$ )

		$E'$	$s'_1$	$s'_2$	Statistical Errors	
Null hypothesis	→	$\omega_1$	$\pi'_1$	$1 - \pi'_1$	→	Type I Error $\alpha = \Pr(s'_2 \omega_1)\mu(\omega_1)$
Alternative hypothesis	→	$\omega_2$	$1 - \pi'_2$	$\pi'_2$	→	Type II Error $\beta = \Pr(s'_1 \omega_2)\mu(\omega_2)$

Private Experiment ( $\pi'_1 + \pi'_2 \geq 1$ )

$(\pi'_1, \pi'_2)$  is an equivalent way to describe buyer's private type; but the reduced form  $(\alpha, \beta)$  directly reflects the preference to the supplementary data.

## Supplementary Data

$E$  is obedient for type  $(\alpha, \beta)$  if every signal  $s_k = (a_{k_1}, a_{k_2})$  is obeyed for  $(\alpha, \beta)$ , i.e.

$$a_{k_j} \in \arg \max_{a_{j'} \in A} E[u_{ij'} | s_k, s'_{j'}] \text{ for all } s_k \text{ and } j = 1, 2.$$

### Lemma

*The outcome of every menu  $\mathcal{M}$  can be attained by a direct and straight mechanism  $\mathcal{M} = \{\mathcal{E}_\Theta, t\}$ , where each type  $\theta = (\alpha, \beta)$  buys obedient  $E_\theta$  from  $\mathcal{E}_\Theta$ , and pays  $t : \mathcal{E}_\Theta \rightarrow \mathbb{R}_+$ .*

$E$	$(a_1, a_1)$	$(a_1, a_2)$	$(a_2, a_1)$	$(a_2, a_2)$
$\omega_1$	$\pi_{11}$	$\pi_{12}$	$\pi_{13}$	$\pi_{14}$
$\omega_2$	$\pi_{21}$	$\pi_{22}$	$\pi_{23}$	$\pi_{24}$

Table: Straight Experiment

Data broker recommends action profiles for different private signal implementations.

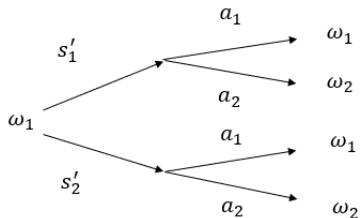


Figure: Data Merging

- Type-wise reduction data structure  $(\pi_1, \pi_2)$ : reduce Type i error by a ratio  $\pi_i$

$\pi_i$ : probability inducing **Type i** error from identifying  $\omega_{-i}$  in  $\omega_i$

## Lemma

The revenues can always be weakly improved by replacing a direct and straight mechanism  $\mathcal{M} = \{\mathcal{E}_\Theta, t\}$  with an alternative direct and straight mechanism  $\mathcal{M} = \{\mathcal{E}'_\Theta, t'\}$ , where  $E'_\theta \in \mathcal{E}'_\Theta$  is Type-wise reduction for all  $\theta$ .

- obedience constraint:  $\pi_1\alpha + \pi_2\beta \leq \min\{\frac{1}{2}\pi_1, \frac{1}{2}\pi_2\}$

$E$	$(a_1, a_1)$	$(a_1, a_2)$	$(a_2, a_1)$	$(a_2, a_2)$
$\omega_1$	$1 - \pi_1$	$\pi_1$	0	0
$\omega_2$	0	$\pi_2$	0	$1 - \pi_2$

Table: Type-wise Reduction Experiment

In the reduced-form, designer allocates the reduction ratio of Type I and II error

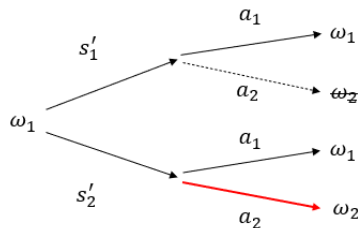


Figure: Reducing Type I error

revelation principle for  $\theta = (\alpha, \beta)$ :

1 direct mechanism  $\mathcal{M} = \{E_\theta, t_\theta\}_{\theta \in \Theta}$ : IC + IR

2 straight information design with Type-wise reduction  $E_\theta (\pi_1(\theta), \pi_2(\theta))$ : Ob-edience  
value of experiment  $(\pi_1, \pi_2)$  for  $(\alpha, \beta)$ : incremental probability of correct identification

$$V(E, \theta) = \underbrace{\alpha + \beta}_{\text{initial overall error}} - \underbrace{\min \left\{ \alpha\pi_1 + \beta\pi_2, \frac{1}{2}\pi_1, \frac{1}{2}\pi_2 \right\}}_{\text{new overall error}}$$

Designer's Problem

$$\max_{\mathcal{M}} \int_{\Theta} t_{\theta} dF(\theta)$$

$$\alpha\pi_1(\theta) + \beta\pi_2(\theta) \leq \min\{\frac{1}{2}\pi_1(\theta), \frac{1}{2}\pi_2(\theta)\}$$

$$\alpha + \beta - \alpha\pi_1(\theta) - \beta\pi_2(\theta) - t_{\theta} \geq 0, \quad \forall \theta \in \Theta$$

$$\alpha + \beta - \alpha\pi_1(\theta) - \beta\pi_2(\theta) - t_{\theta} \geq \underbrace{\alpha + \beta - \min\{\alpha\pi_1(\theta') + \beta\pi_2(\theta'), \frac{1}{2}\pi_1(\theta'), \frac{1}{2}\pi_2(\theta')\}}_{\text{two-step deviation}} - t_{\theta'}, \quad \forall \theta, \theta' \in \Theta$$

Remark: double deviation always exists in our model and traditional FOC method cannot work



# Key Attributes of Data Goods

- data goods: sell (reduced) statistical error (specific multi-dimensional goods)
- interdependence between different Types of error imposes **rigidity** on the menu structure:
  - 1 obedience,  $\alpha\pi_1 + \beta\pi_2 \leq \min\{\frac{1}{2}\pi_1, \frac{1}{2}\pi_2\}$ , constrains the allocation of statistical error
  - 2 double-deviation,  $\min\{\alpha\pi_1 + \beta\pi_2, \frac{1}{2}\pi_1, \frac{1}{2}\pi_2\}$ , weakens the differentiation
- inclusion, exclusion and extraction principles + allocation rigidity shape the bundling policy
- trade-off: extraction of low type surplus v.s. reduce information rent
- the seller can **exploit the horizontal difference** to **neutralize the vertical difference**, through subtly designing the lower-tiered dataset to **nullify the impact of private dataset**

# Data Goods and Other Goods

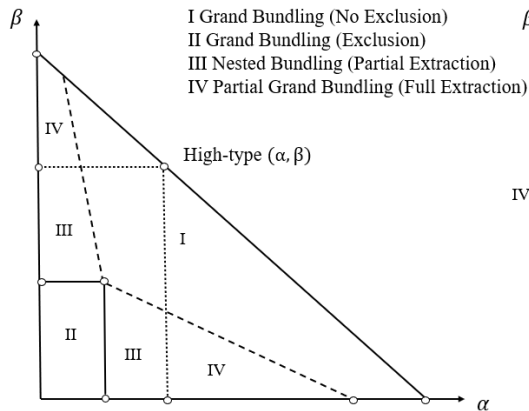
flexibility: physical goods < information goods < data goods < bundling of physical goods

- 1 physical goods: posted-price, no-haggling
  - 2 information goods: position and informativeness (separately “multi-dimensional” goods)
    - 1 position: the Type of error (either Type I or II) -  $(\alpha, 0)$  or  $(0, \beta)$
    - 2 informativeness: the probability of corresponding Type error -  $\alpha$  or  $\beta$
    - 3 allocation: reducing corresponding Type error-  $(\pi_1, \pi_2) = (\pi_1, 1)$  when  $(\alpha, 0)$ , or  $(1, \pi_2)$  when  $(0, \beta)$
- the design and price of information  $\iff$  one-dimensional allocation (differentiated informativeness) + one-dimensional preference with incongruent order
- 3 data goods: allocate different Types of error (specific multi-dimensional goods)
  - 4 multi-dimensional goods: optimal bundling policy tends to be complex and infinite

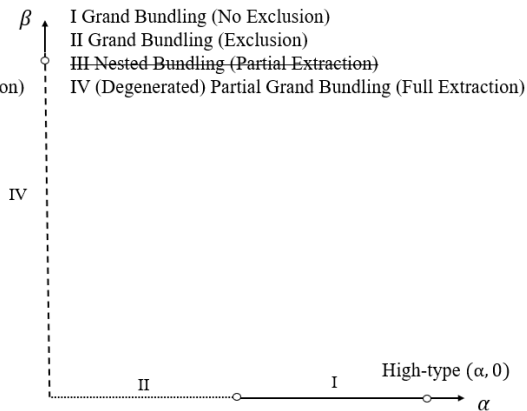
# Data v.s. Information (Bergemann et al.2018): Two Type Case

high type  $(\alpha, \beta)$ , low type  $(\alpha', \beta')$ ,  $\alpha + \beta \geq \alpha' + \beta'$  ▶ Binary Situation

data v.s. information  $\iff$  private experiment  $(\alpha, \beta)$  v.s. private signal  $(\alpha, 0)$  or  $(0, \beta)$



The Low-type Buyer  $(\alpha', \beta')$  located inside  $\alpha' + \beta' \leq \alpha + \beta$



The Low-type Buyer  $(\alpha', 0)$  or  $(0, \beta')$  located inside  $\alpha' \leq \alpha$  or  $\beta' \leq \alpha$

# Continuous Type Space

## Assumption

The statistical error of buyer's private data is perfectly correlated: for the private type  $(\alpha, \beta)$ , it holds that  $k\alpha + \beta = m$ , with  $m \in [0, \frac{1}{2})$  and  $k \in [0, 2m]$ .

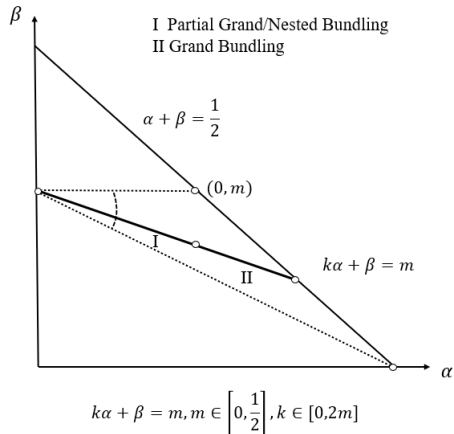
one-dimensional preference:  $(\alpha, m - k\alpha)$ , where  $\alpha \in \mathcal{A} = [\underline{\alpha}, \bar{\alpha}] = [0, \frac{\frac{1}{2}-m}{1-k}]$  draws from distribution  $F$  with a continuous, strictly positive density  $f$

three characteristics:

- 1 the coexistence of both horizontal and vertical differences
- 2 the obedience constraints
- 3 the possibilities of double deviation

The optimal selling mechanism is two-tiered pricing:

- 1  $(E_\alpha, t_\alpha) = (\bar{E}, \bar{t})$  for  $\alpha \in [\alpha^*, \bar{\alpha}]$
- 2  $(E_\alpha, t_\alpha) = (E^*, t^*)$  for  $\alpha \in [\underline{\alpha}, \alpha^*)$ , where  $\pi_1^* = 1 - k(1 - \frac{\alpha^*}{\bar{\alpha}})$ ,  $\pi_2^* = \frac{\alpha^*}{\bar{\alpha}}$
- 3  $\alpha^* \in \arg \max_{\alpha} \alpha ((1 - k)\bar{\alpha} - \frac{1}{2}F(\alpha))$



Double deviation reduce the dimension of multi-goods allocation (rigidity).

## Lemma (Dimension Reduction)

*In the optimal mechanism,  $\frac{1}{2}\pi_2(\alpha) + t_\alpha = \frac{1}{2}\pi_2(\alpha') + t_{\alpha'}$  for all  $\alpha, \alpha' \in [\underline{\alpha}, \bar{\alpha}]$ ,*

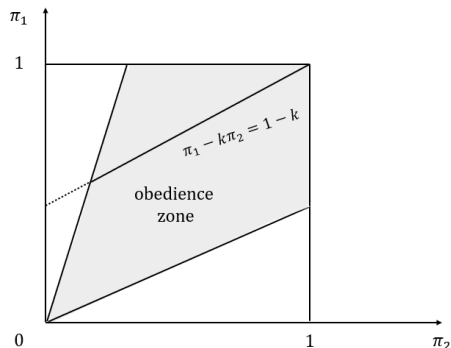
- double deviation  $\{\frac{1}{2}\pi_1, \frac{1}{2}\pi_2\}$ : buyer only commits one Type of error in hypothesis testing.  
 $\Rightarrow$  “drop” his own dataset in statistical decision making
- to exclude double deviations, between two supplementary datasets,  
price gap = informativeness/value gap in statistical decision making  
which is the differences in reduction ratio of a specific type error
- such linkage reduces the dimensions  $\Rightarrow$  one-dimensional screening  
the possibility of two- step deviation limits the flexibility of menu structure brought by multi-dimension nature of data allocation.

Now we restrict to two-tiered structure, the data seller can:

- exploit horizontal differences to neutralize vertical difference  $\Rightarrow$  inclusion of the low type
- improve  $E$  along the neutralization line  $\frac{1-\pi_1}{1-\pi_2} = k$  and extract all the additional value

$$V(E^*, \alpha) = \alpha + (m - k\alpha) - \alpha\pi_1^* - (m - k\alpha)\pi_2^* = m(1 - \pi_2^*) + \alpha[(1 - k) - (\pi_1^* - k\pi_2^*)] = m(1 - \pi_2^*)$$

Such operation can be continued until it hits the obedient boundary (rigidity)



# Proof Sketch

Denote  $V(E, \alpha) = \max\{V_r(E, \alpha), V_n(E, \alpha)\}$ , where

$$V_r(E, \alpha) = \alpha(1 - \pi_1) + (m - k\alpha)(1 - \pi_2), \quad V_n(E, \alpha) = \alpha + (m - k\alpha) - \min\{\frac{1}{2}\pi_1, \frac{1}{2}\pi_2\}$$

two properties of the value functions:

**Property 1. (“same difference”)**

$$V_n(E', \alpha') - V_n(E, \alpha') = V_n(E', \alpha) - V_n(E, \alpha), \forall E, E', \alpha, \alpha'.$$

**Property 2. (“increasing difference”)**

$V_r(E', \alpha') - V_r(E, \alpha') \geq V_r(E', \alpha) - V_r(E, \alpha), \forall \alpha' > \alpha$  if and only if  $\pi'_1 - k\pi'_2 \leq \pi_1 - k\pi_2$ ,  
where inequality binds if and only if  $\pi'_1 - k\pi'_2 = \pi_1 - k\pi_2$ .



Denote  $\lambda(\alpha) : \mathcal{A} \rightarrow \mathcal{A}$  the type of buyers who are exactly indifferent to following seller's recommendation or not, when merging his own private dataset.

$$(i) \lambda(\alpha) = \frac{(\frac{1}{2}-m)\pi_2(\alpha)}{\pi_1(\alpha)-k\pi_2(\alpha)} \text{ if } \pi_1(\alpha) \neq 0; (ii) \lambda(\alpha) = \bar{\alpha} \text{ otherwise.}$$

## Lemma (Characterization of Obedience Zone)

*Optimal menu  $(E_\alpha, t_\alpha)$  satisfies*

- 1  $\pi_2(\alpha)/\pi_1(\alpha) \leq 1$
- 2 *There exists a threshold  $\alpha^*$  such that*
  - 1 *for any  $\alpha < \alpha^*$ ,  $\alpha < \lambda(\alpha)$  and there exists some  $\alpha' > \lambda(\alpha)$  such that  $IC[\alpha' \rightarrow \alpha]$  binds;*
  - 2  $E_\alpha = \bar{E}$  *if and only if*  $\alpha \geq \alpha^*$ .

a class of perturbations  $\{(-k\Delta\pi, -\Delta\pi : \Delta\pi \geq 0)\}$  on supplementary datasets, which does not change the difference in evaluating the dataset between  $V_r$ , but enlarge it between  $V_r$  and  $V_n$

exploit such perturbation of informativeness improvement to the maximal degree

$\Rightarrow$  either double-deviation IC, or Ob is binding

Define  $\gamma(\alpha)$  some type who is indifferent between choosing  $E_{\gamma(\alpha)}$  and conducting double deviation by choosing  $E_\alpha$ .

$$\gamma(\alpha) = \begin{cases} \alpha & \text{if } \alpha = \lambda(\alpha) \\ \tilde{\alpha} \in \{\alpha' > \lambda(\alpha) : \text{IC}[\alpha' \rightarrow \alpha] \text{ is binding}\} & \text{if } \alpha < \lambda(\alpha) \end{cases}$$

### Lemma (Properties of $\lambda$ and $\gamma$ )

*In optimal menu,*

- 1  $\lambda(\alpha) \leq \lambda(\hat{\alpha}) \leq \gamma(\alpha)$  for  $\hat{\alpha} \in [\alpha, \lambda(\alpha)]$ .
- 2  $\pi(\alpha) := \pi_1(\alpha) - k\pi_2(\alpha)$  is non-increasing for  $\alpha \in [0, \bar{\alpha}]$ ;

Two key observations:

- 1 The supplementary dataset amplifies the quality gap of baseline/private datasets.
- 2 The private dataset narrows the quality gap of supplementary datasets.

## Lemma (Equivalent Transformation of Constraints)

*In the optimal mechanism, the IC, IR and Ob conditions are equivalent to*

- 1  $\frac{1}{2}\pi_2(\alpha) + t_\alpha = t^*$  for all  $\alpha \in [\underline{\alpha}, \bar{\alpha}]$ , where  $t^*$  is the associated tariff for  $\bar{E}$ ;
- 2  $V(E_\alpha, \alpha) = \int_{\underline{\alpha}}^{\alpha} (1 - k - \pi(t))dt + V(E_{\underline{\alpha}}, \underline{\alpha})$
- 3  $\pi(\alpha) : [\underline{\alpha}, \bar{\alpha}] \rightarrow [0, 1 - k]$  is non-increasing;
- 4 IR $[\hat{\theta}]$  holds for some  $\hat{\alpha} = \inf\{\alpha | \pi(\alpha) \leq 1 - k\}$ .

Condition 1: price gap should exactly measure difference in Type II error reduction.

Seller's optimization problem can be transformed as a classic optimization problem

$$\begin{aligned} \max_{\pi} \int_{\underline{\alpha}}^{\bar{\alpha}} \frac{-1}{1 - 2m} \left[ \int_{\alpha}^{\bar{\alpha}} (1 - F(t) - tf(t))dt + 2m\alpha \right] d\pi(\alpha) \\ \text{s.t. } \begin{cases} \pi : [\underline{\alpha}, \bar{\alpha}] \rightarrow [0, 1 - k] \text{ is non-increasing} \\ \pi(\bar{\alpha}) = 0 \end{cases} \end{aligned}$$


# Future Work

- robust optimal mechanism
- distributional assumptions sufficient for a simple optimal menu (?)
- many states
- concrete applications

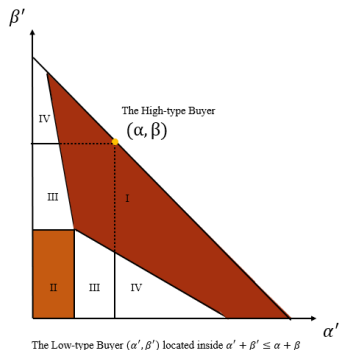
- 1 Information Design as Screening Tools: Admati and Pfleiderer (1986), Admati and Pfleiderer (1990), Babaioff et al. (2012), Bergemann et al. (2018), Yang (2022), Segura-Rodriguez (2022), Bonatti et al. (2023), Bonatti et al. (2024), Rodriguez Olivera (2024)
- 2 Multi-dimensional Screening: Adams and Yellen (1976), McAfee et al. (1989), Armstrong and Rochet (1999), Manelli and Vincent (2007), Hart and Reny (2015), Daskalakis et al. (2017), Carroll (2017), Haghpanah and Hartline (2021); Yang (2022), Deb and Roesler (2023)

**Thank You!**

# Binary Situation

binary type (low  $(\alpha', \beta')$  and high  $(\alpha, \beta)$ ,  $\alpha' + \beta' \leq \alpha + \beta$ , uniform distribution): four polices 

Lemma: sell fully informative  $\bar{E}$  to type-H & type-L experiment  $E$  should be obedient for type-H



Region	Data Menu	Selling Policy
I	$(\bar{E}, \bar{E})$	Inclusive Grand Bundling
II	$(\bar{E}, \phi)$	Exclusive Grand Bundling

grand bundling: sell  $\bar{E}$  with  $(\pi_1, \pi_2) = (0, 0)$

I: low rent, high type-L surplus - including type-L

II: high rent, low type-L surplus - excluding type-L

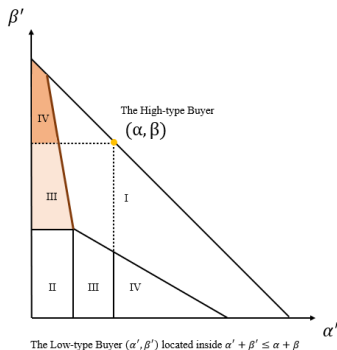
boundary line: inclusion/exclusion of type-L (rent extraction v.s. surplus extraction)

Region	Data Menu	Selling Policy	Binding Constraints
III	$(\bar{E}, E^*)$	Nested Bundling	(IR-L),(IC-H),(Ob-H)

$$(\alpha, \beta) > (\alpha', \beta') \Rightarrow \pi_1 \alpha + \pi_2 \beta > \pi_1 \alpha' + \pi_2 \beta', \text{ given } (\pi_1, \pi_2)$$

$\Rightarrow$  information rent  $> 0$  & higher tendency to make another Type error ((Ob-H) is binding)

$E^*$ : **only** reduce some Type error by a **constant ratio** (e.g. when  $\frac{\beta}{2} < \beta' < \beta$  and  $\alpha' < \frac{\alpha}{2}$ )



exploitation of data structure:

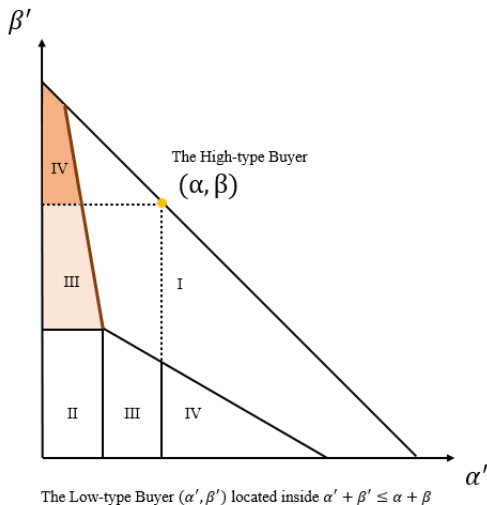
$$1 \cdot \alpha + \pi_2^* \beta = \frac{1}{2} \pi_2^*$$

	$(s_1, s_1)$	$(s_1, s_2)$	$(s_2, s_2)$
$\omega_1$	0	1	0
$\omega_2$	0	$\pi_2^*$	$1 - \pi_2^*$

two benefits for including  $\beta'$ : relatively low rent and high surplus



Region	Data Menu	Selling Policy	Binding Constraints
IV	$(\bar{E}, E_{(\alpha, \beta)})$	Partial Grand Bundling	(IR-L), (IC-H), (Ob), (IR-H)



$E_{(\alpha, \beta)}$ : reduce both Types of error

(i) exploitation of data structure:

$$\alpha\pi_1^* + \beta\pi_2^* = \frac{1}{2}\pi_i^*$$

(ii) no information rent:

$$\alpha\pi_1^* + \beta\pi_2^* = \alpha'\pi_1^* + \beta'\pi_2^*$$

	$(s_1, s_1)$	$(s_1, s_2)$	$(s_2, s_2)$
$\omega_1$	$1 - \pi_1^*(\alpha', \beta')$	$\pi_1^*(\alpha', \beta')$	0
$\omega_2$	0	$\pi_2^*(\alpha', \beta')$	$1 - \pi_2^*(\alpha', \beta')$