Selling Data as a Digital Good with Scaling Valuations

Ningyuan Li[*1], Yanru Guan[*1], Xiaotie Deng[1], Zihe Wang[2], Jie Zhang[3]

[1] Peking University, [2] Renmin University of China, [3] University of Bath

[*] Equal Contribution

PDF

Setup & Contributions

Data markets differ from classical auctions:

1. **Data is non-rival**: the same dataset can be allocated to many buyers
2. **Supply is endogenous**: seller chooses production level $D \geq 0$ at constant marginal cost $c > 0$
3. **Value follows scaling law**: buyer i receiving allocation x_i gets value $v_i \cdot x_i^{\alpha}$ $(0 < \alpha < 1)$

Design Goal: optimize profit over mechanisms that are **truthful** & individual rational

= *allocation + payment*

Our contributions:

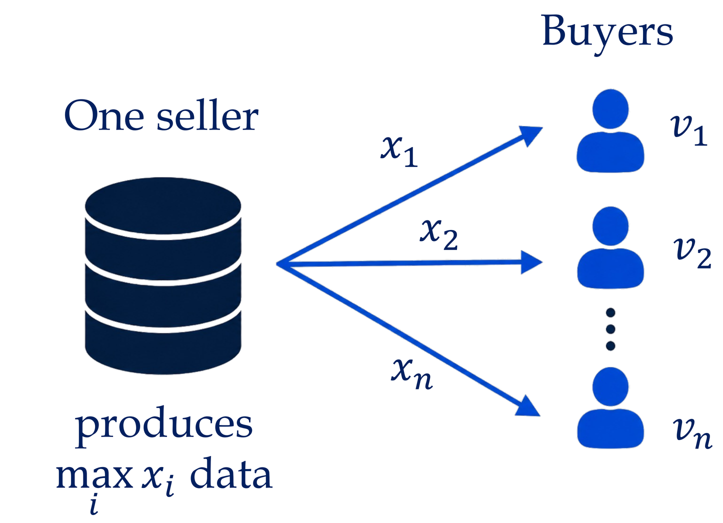
1. **Offline:** optimal truthful mechanisms pool demand into one shared dataset.
2. **Online:** a two-stage truthful mechanism gives a constant approximation.
3. **Bilateral trade:** even with two-sided private information, simple truthful proposal rules recover a constant fraction of First-best Gain-From-Trade.

1. Offline setting: optimal mechanism pools demand

Offline, all buyer types are reported simultaneously.

n buyers, private value $v_i \sim F_i, \forall 1 \leq i \leq n$

Objective: maximize seller's profit $\mathbb{E}_{v\sim F}\left[\sum_i p_i - c \cdot \max_i x_i\right]$



Virtual values (Myerson, 1981)

- $\phi_i(v_i) = v_i - \frac{1-F_i(v_i)}{f_i(v_i)}$
- For irregular distributions, use ironed virtual values ϕ_i^{ir}
- Define $r_i := \left(\phi_i^{\mathrm{ir}}(v_i)\right)_+$

Theorem 1. All buyers with positive virtual values receive same shared data amount

$$x_i^* = 1[r_i > 0] \cdot D^*, \text{ where } D^* = \left(\frac{\alpha}{c}\sum_i r_i\right)^{\frac{1}{1-\alpha}}$$

Given a monotone allocation, payments are set by the standard envelope formula so truthful reporting is optimal.

2. Online setting: uncertainty makes production hard

Buyers arrive sequentially; production decisions are irreversible.

Objective: approximate optimal offline (with hindsight) profit OPT

Simple baselines fail

- *Myopic greedy*: optimizes per-buyer profit, ignoring future demand; worst case expected profit $\leq O\left(n^{-\frac{\alpha}{1-\alpha}}\right) \cdot \mathrm{ONL}$
- *Non-adaptive*: fixed D ex-ante, worst case expected profit $\leq \epsilon \cdot \mathrm{ONL}, \forall \epsilon > 0$

ONL := optimal online profit, characterized by DP

A two-stage mechanism

1. Initial production $D_0 = \left(\frac{\alpha}{c}(1-\delta)\sum_i \mathbb{E}[r_i]\right)^{\frac{1}{1-\alpha}}$
   - δ only depends on α
2. Upon arrival of buyer i, compute r_i and myopic target $x_i^{\mathrm{MG}} = \left(\frac{\alpha r_i}{c}\right)^{\frac{1}{1-\alpha}}$
3. Expand data: $D_i = \max\left\{D_{i-1}, x_i^{\mathrm{MG}}\right\}$

Theorem 2. The two-stage mechanism achieves a constant-factor approximation

Expected profit $\geq \kappa_\alpha \cdot \mathrm{OPT}$, where κ_α $(*)$ only depends on α

$(*)$ $\kappa_\alpha = \left(\left(1+\frac{1}{1-\alpha}\right)C_{1/(1-\alpha)}\right)^{-1}$, where C_p is the constant from Rosenthal's inequality.

3. Bilateral trade: simple truthful mechanisms work

Now both the buyer's value and the seller's cost are private.

One buyer, private value $v \sim F_B$; One seller, private cost $c \sim F_S$.

A mechanism asks both sides to report private information. Truthful reporting should be optimal.

Objective: maximize Gain-From-Trade $\mathbb{E}_{v\sim F_B, c\sim F_S}[vx(v,c)^{\alpha} - cx(v,c)]$

Seller-proposing (SP)

Seller decides how much data to trade

$$x^{\mathrm{SP}} = \left(\frac{\alpha(\phi_B(v))_+}{c}\right)^{\frac{1}{1-\alpha}}$$

Approximation guarantee:
$\alpha^{\frac{1}{1-\alpha}}$ of First-best GFT

Buyer-proposing (BP)

Buyer decides how much data to trade

$$x^{\mathrm{BP}} = \left(\frac{\alpha v}{(\phi_S(c))_+}\right)^{\frac{1}{1-\alpha}}$$

Approximation guarantee:
$\alpha^{\frac{\alpha}{1-\alpha}}$ of First-best GFT
in particular at least $1/e$

- SP and BP give proposal power to one side of the market. Both are truthful and individually rational.
- First-best GFT: directly solves $\max_{x\geq 0}[vx^{\alpha} - cx]$; ignores incentive and budget-balance constraints.

$\phi_B(v) = v - \frac{1-F_B(v)}{f_B(v)}, \phi_S(c) = c + \frac{F_S(c)}{f_S(c)}$