Open Data, Order-Flow Privacy, and Market Quality

Lin William Cong[1] Siguang Li[2] Mingzhe Zheng[2]

[1]Nanyang Technological University, ABFER, CEPR, and NBER
[2]Society Hub, HKUST (Guangzhou)

ABFER, Singapore
May 2026

Outline

1. Introduction
2. Model Setup
3. Trading Equilibrium
4. Endogenous Data Sales
5. Optimal Data Design
6. Empirical Evidence
7. Conclusion

Data antitrust is a global regulatory trend

Regulators increasingly intervene to dismantle information monopolies:

- **US.** SEC *Market Data Infrastructure Rule*: competing consolidators.
- **EU.** 2020 in-depth review of the LSEG / Refinitiv acquisition.
- **China.** 2023 antitrust action ending an exclusive bond-data agreement.

The common belief behind "Open Data" initiatives:

turning data monopolies into competitive markets improves market quality.

This paper identifies a critical oversight in this view.

Open Data involves two different types of data

Opening data, in practice, involves two different kinds of data:

Fundamental data

Signals about asset value.

e.g. Open banking, credit histories.

Reduces information asymmetry.

Non-fundamental data

Signals about trading.

e.g. Real-time order, PFOF-style data.

Exposes trading intentions.

This paper

Does competition in data markets always improve market quality, once data includes non-fundamental order-flow information?

We frame the answer as the open data paradox in capital markets:

1. Extend **Kyle (1985)**: traders may acquire data on the size of noise trading.
2. Introduce an **upstream data market**; compare six data market structures.
3. Characterize the optimal data design: selective openness.
4. Test predictions using **China's 2023 bond-data antitrust** intervention.

Outline

1 Introduction

2 Model Setup

3 Trading Equilibrium

4 Endogenous Data Sales

5 Optimal Data Design

6 Empirical Evidence

7 Conclusion

Kyle (1985) with two types of data

A single risky asset with terminal value $\theta \sim N(0, \tau_\theta^{-1})$, prior mean normalized to 0.

- N **potential rational traders.** Risk-neutral, submit market orders x_i.
- **Noise traders.** Submit an aggregate random order $u \sim N(0, \tau_u^{-1})$, independent of θ.
- **Competitive market maker.** Observes only aggregate order flow $y = \sum_i x_i + u$, sets $p(y) = \mathbb{E}[\theta \mid y]$ (zero expected profit).

Upstream data sellers can produce two signals:

Fundamental data

$$s_f = \theta + \epsilon_f, \quad \epsilon_f \sim N(0, \tau_f^{-1})$$

helps form a belief about θ.

Non-fundamental data

$$s_n = u + \epsilon_n, \quad \epsilon_n \sim N(0, \tau_n^{-1})$$

Not useful on its own: $u \perp \theta$.

Information structure: four trader types

The data market partitions the N rational traders into four types:

Type	Information set	Mass
FN-type	$\{s_f, s_n\}$, the full package	N_{fn}
F-type	$\{s_f\}$, fundamental only	N_f
N-type	$\{s_n\}$, order-flow only	N_n
U-type	$\varnothing$, uninformed	N_u

Two sufficient statistics summarize the information structure:

$$M \equiv N_f + N_{fn} \quad \text{(fundamental access)}, \qquad K \equiv N_n + N_{fn} \quad \text{(order-flow access)}.$$

Six data-sales regimes

"Open Data" means non-discriminatory *access*, not free data. Sellers monetize access.

	Integrated (one seller, both)	**Segmented** (one data type each)
Monopoly	I. Integrated monopoly	II. Segmented monopoly
Competition	IV. Integrated competition	III. Segmented competition
Asymmetric	V. & VI. Generalist vs. specialist (overlap in s_f or s_n)	

- **Integration.** Whether one seller controls both data types (internalization).
- **Competition.** Whether ≥ 2 sellers operate within a data market.
- **Asymmetric oligopoly.** A generalist vs. a specialist.

Timing and equilibrium

The game has three stages:

- $t = 0$. **Data-market competition.** Each data provider chooses the quantity of data to supply, (N_{fn}, N_f, N_n), to maximize its ex-ante profit.
- $t = 1$. **Trading.** Traders realize their types, observe signals, and submit orders simultaneously.
- $t = 2$. **Pricing.** The market maker observes y, sets $p(y)$; payoffs realize.

Outline

1 Introduction

2 Model Setup

3 Trading Equilibrium

4 Endogenous Data Sales

5 Optimal Data Design

6 Empirical Evidence

7 Conclusion

The unique linear equilibrium

Take the information structure (N_{fn}, N_f, N_n, N_u) as given. **Linear strategies:**
F-type: $x^f = \beta_f \hat{\theta}_f$ with $\hat{\theta}_f \equiv \mathbb{E}[\theta \mid s_f]$;
N-type: $x^n = \gamma_n \hat{u}_n$ with $\hat{u}_n \equiv \mathbb{E}[u \mid s_n]$;
FN-type: $x^{fn} = \beta_{fn} \hat{\theta}_f + \gamma_{fn} \hat{u}_n$;
U-type: $x^u = 0$.

Proposition 1

Let $M \equiv N_f + N_{fn}$ and $K \equiv N_n + N_{fn}$. A unique symmetric linear equilibrium exists:

$$\beta = \frac{1}{\lambda(M+1)}, \qquad \gamma = -\frac{1}{K+1}, \qquad \lambda = \frac{\sqrt{M}}{M+1}\sqrt{\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}}\,\frac{1}{\sqrt{h(K)}},$$

where $h(K)$ is the residual noise in order flow, strictly decreasing in K.

Proposition 1: four key findings

1. **Linear additivity.** $\beta_f = \beta_{fn}$ and $\gamma_n = \gamma_{fn}$, and $x^{fn} = x^f + x^n$: holding one signal does not change how a trader uses the other.

2. **Trading against expected noise.** $\gamma = -\frac{1}{K+1} < 0$: traders with s_n trade *against* the noise they predict.

3. **Profitable without fundamental information.** A non-fundamental trader earns rents, even without any knowledge about fundamentals, both ex-ante and ex-post.

4. **Fundamental data always matters.** If $M = 0$, order flow carries no information, $\lambda = 0$, and all order-flow rents vanish. The value of s_n comes from the price impact.

Non-fundamental data is independently profitable

The market maker cannot separate informed orders from noise, so she prices *all* order flow the same way, $p = \lambda y$. A trader who observes u knows part of that price move is "wrong" and bets against it.

Example: $M = K = 1$, perfect signals, $\tau_\theta = \tau_u = 1 \Rightarrow \lambda = 1,\ \beta = 0.5,\ \gamma = -0.5$.

	x^f	x^n	y	p	Non-fund. profit $(\theta - p)\, x^n$
State 1: $\theta = 2,\ u = +1$ (same sign)	1	-0.5	1.5	1.5	-0.25
State 2: $\theta = 2,\ u = -1$ (opposite sign)	1	$+0.5$	0.5	0.5	$+0.75$

When u aligns with θ, the trade is on the wrong side but the mispricing is small. When u opposes θ, the mispricing is large and the trade is on the right side.

Gains exceed losses $\Rightarrow$ strictly positive expected profit, with zero fundamental knowledge.

Data access and market quality

How do fundamental access (M) and order-flow access (K) move market quality?

Market-quality measure	$M \uparrow$	$K \uparrow$
Market depth $MD = 1/\lambda$	$\uparrow$	$\downarrow$
Informational efficiency IE	$\uparrow$	$\rightarrow$ no effect
Order-flow volatility $\mathrm{Var}(y)$	$\uparrow$	$\downarrow$
Noise-trader welfare W_{noise}	$\uparrow$	$\uparrow$

- Order-flow data hurts depth yet does nothing for price discovery.
- Low volatility need not mean a liquid market: $K \uparrow$ lowers both volatility and depth.

The value of data

Ex-ante expected profit of a j-type trader: $\pi_j \equiv \mathbb{E}\big[\,\mathbb{E}[x^j(\theta - p) \mid \mathcal{I}_j]\,\big]$.

The value of data is the difference in π_j with and without access: $v_f \equiv \pi_{fn} - \pi_n$, $v_n \equiv \pi_{fn} - \pi_f$.

$$v_f(M,K) = \frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}\,\frac{1}{\lambda(M+1)^2}, \qquad v_n(M,K) = \frac{\tau_n}{\tau_u(\tau_u+\tau_n)}\,\frac{\lambda}{(K+1)^2}.$$

- **Additivity:** $\pi_{fn} = \pi_f + \pi_n$.

Data sales generate externalities among buyers

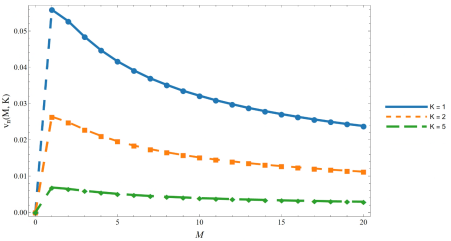
Strategic substitutability

$$\frac{\partial v_f}{\partial M} < 0, \quad \frac{\partial v_f}{\partial K} < 0, \quad \frac{\partial v_n}{\partial K} < 0.$$

Each extra informed trader imposes a negative externality on the others.

Non-monotone externality in M

$v_n = 0$ at $M = 0$; jumps to its global max at $M = 1$; then strictly falls for $M > 1$.



Value of non-fundamental data $v_n(M, K)$.

A monopolist has an incentive to internalize this externality and restrict sales.

Outline

1 Introduction

2 Model Setup

3 Trading Equilibrium

4 Endogenous Data Sales

5 Optimal Data Design

6 Empirical Evidence

7 Conclusion

Why study data-market structure?

Regulating data products is hard: vendors repackage equivalent information. A more feasible regulatory method is to intervene structure of the data market.

We vary the data market along two dimensions:

- **Monopoly vs. competition.** How many sellers operate within a data market? With ≥ 2 sellers, each internalizes only a fraction of any externality.
- **Integrated vs. segmented operation.** Does one seller control both data types?

An seller may internalize two externalities:

- **Own-market dilution.** Selling more of a data type lowers its own marginal value: informed traders' rents are spread thinner.
- **Cross-market externality.** The two data types are strategic substitutes.

Equilibrium data sales across the six regimes

Regime	(M, K)	Economic content
I. Integrated monopoly	$(1, 0)$	internalizes externality; suppresses s_n
II. Segmented monopoly	$(1, \geq 1)$	s_n seller ignores the externality
III. Segmented competition	(N, N)	full saturation once ≥ 2 sellers
IV. Integrated competition	$(\leq N, N)$	partial internalization
V. Asym. oligopoly (s_f)	$(\geq 1, \leq K^{II})$	generalist may self-segment
VI. Asym. oligopoly (s_n)	$(1, N)$	s_f stays monopolistic

- A fundamental-data monopolist always sells to one trader ($M = 1$).
- With ≥ 2 non-cross-selling sellers, they sell as much data as possible: non-linear Cournot.

No single regime is optimal

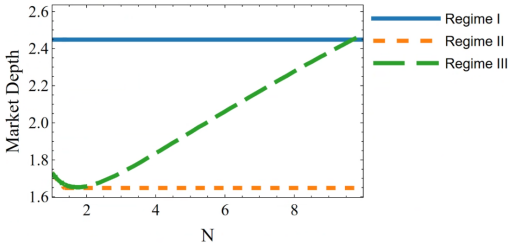
The optimality of data market structures depend on measures of financial market quality:

- **Informational efficiency:** $IE^I = IE^{II} \leq IE^{III}$. Competition always wins.
- **Market depth:** $MD^{III} \leq MD^I$ iff $N \leq \widehat{N}$. Monopoly wins in small markets.
- **Order-flow volatility:** minimized by the *segmented monopoly*.
- **Noise-trader welfare:** $W^I_{\text{noise}} < W^{II}_{\text{noise}} \leq W^{III}_{\text{noise}}$. Competition wins.

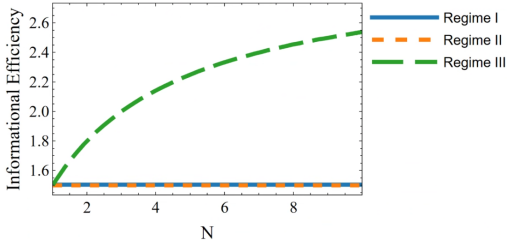
where $\widehat{N} = 1 + 2\rho_n + \sqrt{4\rho_n^2 + 3\rho_n}$, with $\rho_n \equiv \tau_n/\tau_u$, increasing in ρ_n.

The optimal regime depends on the regulator's objective and on market size N.

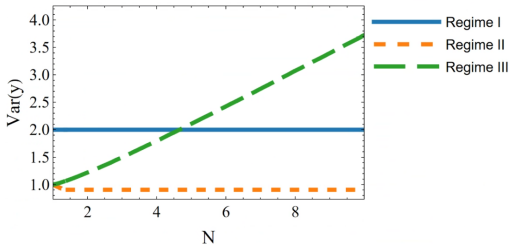
Market size N and market quality



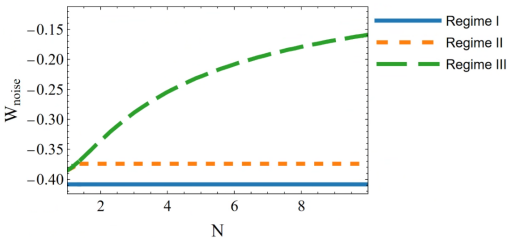
(a) Market depth



(b) Informational efficiency



(c) Order-flow volatility



(d) Noise-trader welfare

Outline

1 Introduction

2 Model Setup

3 Trading Equilibrium

4 Endogenous Data Sales

5 Optimal Data Design

6 Empirical Evidence

7 Conclusion

Is the open data paradox unavoidable?

The six regimes generate only a limited collection of allocations (M, K).

Let a regulator choose (M, K) directly, subject to the equilibrium pricing of Proposition 1. The objective is an *institutional mandate*:

$$W(M,K) = \alpha\, IE + \beta\, MD + \gamma\, W_{\text{noise}} - \zeta\, \text{Var}(y), \qquad (\alpha,\beta,\gamma,\zeta) \in \mathbb{R}^4_+.$$

- α, β: price discovery and liquidity.
- γ: retail-investor welfare. ζ: order-flow-volatility penalty.

Selective openness

Proposition: selective openness

Under an efficiency–liquidity mandate ($\alpha, \beta > 0$, $\gamma = \zeta = 0$), the regulator's problem has a unique maximizer

$$(M^*, K^*) = (N, 0) :$$

full fundamental access, full order-flow suppression. It *weakly dominates every one of the six regimes* in both *IE* and *MD*.

- The open data paradox reflects the *bundling* of fundamental access with order-flow disclosure.
- Implementing $(N, 0)$ requires competition in fundamental data, while restricting order-flow data sales.
- **Robustness region $\mathcal{R}$.** For $\gamma > 0, \zeta > 0$, there exists $\mathcal{R} \neq \emptyset$, the set of weights $(\alpha, \beta, \gamma, \zeta)$ for which $(N, 0)$ remains the unique optimum. It is an open neighborhood of the pure efficiency–liquidity slice $\{(\alpha, \beta, 0, 0)\}$.

Outline

1 Introduction

2 Model Setup

3 Trading Equilibrium

4 Endogenous Data Sales

5 Optimal Data Design

6 Empirical Evidence

7 Conclusion

A natural experiment: China's interbank bond market

Setting. About 80% of bond trading; few participants and broker-reliant, hence a small-N market.

The monopoly. A dominant data vendor (*Firm S*) held a decade-long exclusive agreement with one of six major brokers (*Broker X*), becoming the only platform aggregating all six brokers' real-time quote and transaction data.

The shock, March 2023. China's first data antitrust action banned exclusivity clauses; within days, rival platforms integrated Broker X's data.

A clean, exogenous shift: data monopoly $\rightarrow$ data competition.

Empirical design

Difference-in-differences, bond-month panel, 2021–2025; 700 bonds, 24,025 obs.

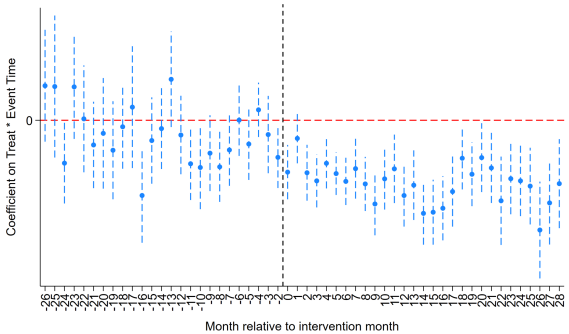
$$Y_{it} = \alpha + \beta\,(Treat_i \times Post_t) + \gamma X_{it} + \mu_i + \lambda_t + \varepsilon_{it}$$

- **Treat:** bonds ever matched by Broker X before the intervention.
- **Post:** March 2023 and later.
- **Informational efficiency:** price deviation $= \left|\frac{\text{close price}-\text{valuation}}{\text{valuation}}\right|$.
- **Liquidity:** turnover = volume / outstanding market value.
- Controls: rating, maturity, duration, convexity, type; bond & month FE; SE clustered by bond.

Result 1: informational efficiency improved

	(1)	(4)
$Treat \times Post$	-0.011^{***}	-0.004^{***}
	(0.001)	(0.001)
Controls / FE	No	Yes
Observations	24,025	24,025

Outcome: price deviation.



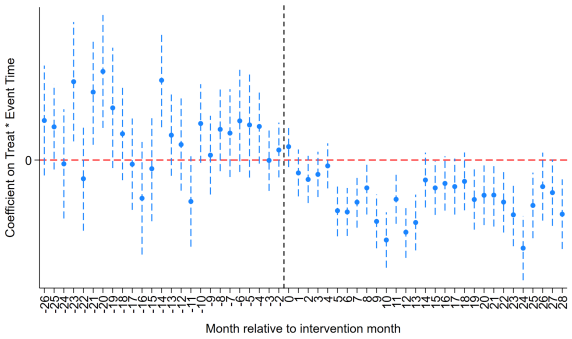
Event study: price deviation.

Price deviations fell 30 to 40 bps, immediate and persistent.

Result 2: market liquidity declined

	(1)	(4)
Treat × *Post*	−0.158***	−0.042***
	(0.010)	(0.014)
Controls / FE	No	Yes
Observations	24,025	24,025

Outcome: turnover.



Event study: turnover.

Turnover declined steadily: the open data paradox

Outline

1 Introduction

2 Model Setup

3 Trading Equilibrium

4 Endogenous Data Sales

5 Optimal Data Design

6 Empirical Evidence

7 Conclusion

Conclusion

- Data market structure governs the type of data released.
- Fundamental and non-fundamental data have distinct, interacting effects: competition aids price discovery but can harm liquidity, the open data paradox.
- A regulator can escape the paradox through selective openness: open fundamental data fully, withhold or coarsen order-flow data.
- China's 2023 bond-data antitrust confirms the trade-off: efficiency $\uparrow$, turnover $\downarrow$.
- Open Data is not unambiguously welfare-improving. Optimal regulation must account for data type, market size and structure.

Thank you!

Comments and questions welcome.