Open Data and Financial Market Quality*

Lin William Cong† Siguang Li‡ Mingzhe Zheng§

First draft: June 2025; this draft: January 2026.

Abstract

Global financial regulators increasingly promote open-data mandates to dismantle information monopolies and improve market transparency. This paper identifies a fundamental equilibrium tension—the open data paradox—arising from the endogenous composition of disseminated information. We develop a theoretical model in which competition among data vendors expands access to fundamental signals, improving price efficiency, but also incentivizes the dissemination of non-fundamental order-flow information that indirectly reduces liquidity provision. While fundamental data reduces informational frictions, non-fundamental data increases adverse selection by facilitating strategic sniping of noise-trader flow, thereby impairing market liquidity. We test the model's predictions using a quasi-natural experiment generated by China's 2023 antitrust intervention in the bond-market data industry. Consistent with the theory, dismantling the data monopoly enhances pricing efficiency but reduces market turnover. These results highlight that open-data policies operate through opposing informational channels, implying that optimal data regulation must balance the benefits of information access against the liquidity risks of exposing order-flow privacy, particularly in thin or regulated markets.

JEL Classification:. G14. G18. D82. L41.

Keywords: Open Data Paradox, Non-fundamental Information, Data Antitrust, Interbank Bond Market, Financial Market Quality, Adverse Selection.

*The authors are especially grateful to Jaden Chen, Liyan Yang, and Mao Ye for helpful discussions and suggestions. Siguang Li acknowledges the generous research support from the National Science Foundation of China (Grant No. 72573146). All errors are our own.

†Cornell SC Johnson College of Business (Johnson), ABFER, and NBER. Email: will.cong@cornell.edu

‡Society Hub, HKUST (Guangzhou). Email: siguangli@hkust-gz.edu.cn

§Society Hub, HKUST (Guangzhou). Email: mzheng842@connect.hkust-gz.edu.cn

1 Introduction

Data antitrust has become a prominent trend in modern financial regulation. From the SEC's infrastructure reforms[1] to aggressive antitrust enforcement in Europe[2] and Asia[3], regulators are increasingly intervening to dismantle information monopolies. The primary goal is to lower costs and broaden access to proprietary data feeds. This policy stance rests on a common belief driving global "Open Data" initiatives: that transforming data monopolies into competitive markets can enhance market quality. This paper identifies a critical oversight in this view. We argue that data market structure governs not only the price of data, but also the type of information released. Our central finding is that while competition maximizes transparency of fundamentals, it incentivizes the over-selling of predatory order-flow signals, thereby damaging market liquidity.

We frame this trade-off as the open data paradox in capital markets. To understand the consequences of data liberalization, we must distinguish between two types of transparency. In open banking, regulators mandate the sharing of fundamental data, such as consumers' credit histories, to reduce asymmetry and foster efficiency. In contrast, in trading markets, proposals to share real-time transaction-level data, analogous to the retail order flows in payment-for-order-flow (PFOF) schemes, involve sharing non-fundamental data. While fundamental data helps verify asset quality, non-fundamental data exposes trading intentions. We theorize that data monopolies often function as mechanisms for order privacy, protecting market depth by limiting the circulation of these order-flow signals. When antitrust actions force these data open, this results in a broad release of non-fundamental information that enables targeted opponent strategies.

To formalize the mechanism, we extend the framework of Kyle (1985) to allow traders to acquire non-fundamental data, which informs them about the size of noise trading. This non-fundamental signal has no intrinsic value for asset valuation; its value is purely strategic. Our model also introduces an upstream data market to investigate the joint sale of fundamental and non-fundamental data. It is important to clarify that, in this context, "Open Data" does not imply free data, but rather non-discriminatory access. Our frame-

[1]See SEC, "Market Data Infrastructure," www.govinfo.gov/content/pkg/FR-2021-04-09/pdf/2020-28370.pdf.

[2]In 2020, the European Commission opened an in-depth investigation to assess the proposed acquisition of Refinitiv by the London Stock Exchange Group. Their main concern was that the integration might limit market participants' access to financial data products. See https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1140.

[3]In 2023, Chinese regulators initiated an antitrust investigation into bond market data providers, which resulted in the termination of a decade-long exclusive data-sharing agreement between the largest money broker in the China interbank market and a dominant bond data platform. See https://www.samr.gov.cn/fldes/tzgg/xzcf/art/2024/art_a98612333b5f48ca98e3ce1c5710dde4.html.

work reflects the standard authorized redistribution model seen in financial markets. Just as exchanges or open banking APIs authorize intermediary aggregators to package and sell data to end users, the market we study involves data sellers monetizing access. We rigorously investigate a spectrum of data market structures, ranging from exclusive redistribution, where a single integrated vendor controls the sale of both data types, to competitive redistribution, where multiple vendors engage in quantity competition within a data market. We compare the consequences of each regime, providing key insights into the policy trade-offs for the governance of financial market quality.

Our model yields several novel findings. First, fundamental and non-fundamental data affect market quality in opposite ways. Increasing the number of fundamentally informed traders deepens the market and improves informational efficiency. As more fundamental traders compete in a batch order game, each must trade more aggressively on their signals, accelerating the incorporation of information into prices. The resulting competition paradoxically lowers adverse selection risk for market makers, reducing price impact and enhancing market depth. By contrast, non-fundamental data exerts the opposite effect on liquidity: when more traders act on signals about noise order flow, they dampen the randomness in total order flow, making prices appear more informative. Market makers then perceive higher informational risk and raise their price impact, reducing depth. Interestingly, while non-fundamental traders influence liquidity, they contribute nothing to informational efficiency. The information in prices remains solely determined by fundamental signals. The reason is subtle: although non-fundamental traders remove some noise from aggregate orders, fundamental traders adjust their aggressiveness in response, leaving overall price informativeness unchanged. Hence, non-fundamental data redistributes rents without improving information transmission.

From a volatility perspective, the composition of informed traders also affects order flow volatility. More fundamental traders amplify aggregate trading activity, increasing order flow volatility. In contrast, more non-fundamental traders smooth total orders by offsetting noise trading, effectively stabilizing the flow of liquidity. This finding challenges the conventional intuition that low volatility signals a more liquid market: in our setting, volatility and liquidity can move in opposite directions.

Another interesting result is that trading on non-fundamental data can be profitable even without any information on fundamentals, both ex ante and ex post. In contrast to prior studies, where non-fundamental signals improve traders' inference about fundamentals by filtering noise from total order flow or observable prices (Ganguli and Yang, 2009; Manzano and Vives, 2011; Madrigal, 1996; Yang and Zhu, 2020), our model departs sharply from

this logic. This result stems from a simple but powerful friction: the market maker cannot distinguish informed from noise-driven orders and must price all order flow symmetrically. Traders who can anticipate the noise component exploit this mispricing channel, yielding a non-fundamental information rent.

To illustrate this mechanism, consider a simple case with one fundamental trader and one non-fundamental trader, each perfectly informed about their respective signals. When the fundamental shock and the noise shock move in the same direction, the non-fundamental trader's orders partially offset the fundamental trader's orders, leading to small losses. When the shocks move in opposite directions, however, their trades amplify price deviations and generate larger profits. Because gains are larger than losses in absolute value, the non-fundamental trader earns a positive expected profit despite having no knowledge of fundamentals. This asymmetry, arising purely from the market maker's inability to disentangle order sources, demonstrates how information on non-fundamental order flow alone can sustain profitable trading.

This clear result is due to the linear additivity of trading strategies in our model. For a trader who has access to both types of data, the trading intensity associated with either data source is the same as that of a trader with access to only one type. We show that there does not exist an equilibrium in which traders with access to both data types employ a non-linear strategy with an interaction term. This implies that there is no non-linear interaction effect between trading strategies based on fundamental data and non-fundamental data. This result follows from the absence of risk aversion and from the fact that prices (or order flows) are not publicly observable. It helps to demonstrate a direct channel explaining the independent profitability of non-fundamental trading. In addition, we emphasize that the profitability of non-fundamental trading relies on the presence of fundamental traders in the market. When there is no fundamental information, market makers correctly infer that order flow is uninformative and price solely based on prior beliefs.

Because trading strategies are linearly additive, the market can be decomposed into two orthogonal layers: a fundamental market in which information drives price discovery, and a non-fundamental market in which liquidity provision to noise traders itself becomes a source of rent. This structure provides a microfoundation for the profitability of activities such as PFOF, which involve the sale of retail order flow data to liquidity providers without requiring portfolio skill or informational advantage. It also highlights a new policy insight: selling data on noise trading can generate welfare and liquidity effects similar to those of off-exchange retail execution, even when no trades are actually internalized. This equivalence suggests that the informational dimension of market design—who observes which data—may matter

as much as institutional trading rules themselves, and that data concentration can shape market efficiency and depth in non-trivial ways.

This motivates us to investigate how data market structure can be used to govern the quality of financial markets. To capture different market structures, we compare six regimes: (i) (horizontally) integrated data monopoly, in which a monopolist controls the sale of both fundamental and non-fundamental data; (ii) segmented data monopoly, in which two separate monopolists sell fundamental and non-fundamental data, respectively; (iii) segmented competition, in which at least two sellers in a market simultaneously choose quantities and share a common price; (iv) integrated competition, in which at least two sellers are able to sell both types of data; (v) asymmetric oligopoly with overlap for fundamental data, in which a specialist sells only fundamental data and a generalist sells both types; and (vi) asymmetric oligopoly with overlap for non-fundamental data, in which a specialist sells only non-fundamental data and a generalist sells both types.

For each regime, we solve for the equilibrium quantities of data sold and analyze the resulting implications for market quality, including liquidity, informational efficiency, volatility, and welfare. The results reveal that data market structure exerts a first-order impact on financial market quality. An integrated data monopolist fully internalizes the negative cross-market externality: selling non-fundamental data reduces the value of fundamental data. To protect its primary profit source, the monopolist optimally restricts all non-fundamental sales and sells the fundamental signal to only one trader. This outcome maximizes market depth by preventing liquidity-reducing noise trading, but at the cost of minimal information aggregation and highly concentrated informational power.

When the market is segmented, the monopolist controlling non-fundamental data fails to internalize its adverse impact on the fundamental market. As a result, non-fundamental sales emerge endogenously, generating additional trading activity but reducing overall liquidity. The fundamental data monopolist still restricts sales to one trader, but the coexistence of non-fundamental data leads to a thinner and noisier market. In this sense, segmentation introduces competition across data types but also generates inefficiencies by ignoring interdependencies in information sales.

Under segmented competition, however, the outcome diverges sharply from that of standard Cournot competition. In a conventional setting with linear demand, the equilibrium quantity lies between the monopoly and perfect competition outcomes. In contrast, our model exhibits a corner solution: once there are at least two data sellers, the market becomes fully saturated. Each seller bears only a fraction of the negative externality from expanding sales, and because data are nonrival, the incentive to expand output persists until

all traders are served. As a result, equilibrium features maximal data dissemination, with every trader acquiring both types of signals. Thus, the presence of even minimal competition in data markets drives the outcome directly to the competitive limit.

These results suggest that no single data market structure is universally optimal. The integrated monopoly maximizes liquidity but minimizes informativeness; competition maximizes informativeness but may destabilize liquidity; and segmentation lies in between, mitigating volatility but not fully correcting inefficiencies. The optimal regulatory choice therefore depends on the policymaker's objective and the size of the financial market. In small markets, restricting data competition can improve liquidity, whereas in large markets, promoting data competition enhances overall efficiency. These findings provide a theoretical foundation for current regulatory debates on data market concentration, data access, and related practices such as PFOF and proprietary data distribution.

The theoretical results highlight a key policy trade-off. Competition maximizes efficiency but can harm liquidity, especially in small markets. To examine these predictions empirically, we exploit a quasi-natural experiment: China's first antitrust intervention against a data monopoly in the interbank bond market. Prior to March 2023, a dominant data seller maintained an exclusive agreement with one of the six major bond brokers, making it the only platform aggregating real-time quote and transaction data from all brokers. This exclusivity effectively monopolized access to broker data, a crucial source of market information. In March 2023, regulators prohibited such exclusivity clauses, forcing the data seller to open access and allowing competitors to integrate the broker's data within days. This regulatory shock abruptly transformed the data market from monopoly to competition, providing a clean empirical setting to study the consequences of this change in data market structure.

Using a difference-in-differences design, we compare bonds historically covered by the broker of interest with those never covered, before and after the intervention. We measure two key dimensions of market quality: (i) price deviation, capturing informational efficiency and defined as the relative deviation between the market close price and model-implied valuation; and (ii) turnover, capturing market liquidity and measured as trading volume relative to outstanding amount. Consistent with our theoretical framework, we find that breaking the data monopoly significantly improves informational efficiency but reduces liquidity. The timing of the effects, confirmed by event-study plots, shows that pricing efficiency improves immediately and persistently, while turnover declines steadily after the intervention. These results indicate that enhanced data access improves information aggregation but simultaneously intensifies adverse selection and noise trading, reducing liquidity in a market with few participants.

Taken together, the empirical findings mirror the core prediction of our model: the transition from monopoly to competition in data markets leads to a trade-off between informativeness and liquidity. In small markets, the liquidity cost of data liberalization can dominate, suggesting that moderate concentration or coordination among data providers may sometimes enhance market stability. Our evidence provides new insights for debates on data access, competition, and the design of financial data markets.

The rest of the paper is organized as follows. The remainder of Section 1 positions our contribution in the literature. Section 2 outlines the model. Section 3 establishes the financial market equilibrium with exogenous data allocation. Section 4 endogenizes data sales and analyzes the link between data market structure and financial market quality. Section 5 presents the empirical results. Section 6 concludes. All proofs are collected in the Appendix.

Literature Review This article relates to three strands of literature: the implications of big data in financial markets, information sales in financial market, and the data economy.

A number of recent papers have analyzed the implications of big data on financial markets (Goldstein et al., 2021, 2025). Data technologies in financial markets can be categorized into two types. One focuses on processing data faster, while the other aims to extract information from a broader range of data. The former has primarily evolved into high-frequency trading and algorithmic trading, such as Biais et al. (2015) analyzed different-sized institutions exhibit distinct trading behaviors in high-frequency trading markets due to information asymmetry. As for the latter, the richer data sources and multi-dimensions play a crucial role in this aspect. Dugast and Foucault (2018) studied the behavior of investors and the impact on market price informativeness when there are multiple data sources in the market. An important insight they propose is that low-quality data with low cost can crowd out high-quality data, thereby diminishing the price informativeness. Huang et al. (2022) abstract the concept that alternative data requires high learning costs and endogenizes the skills investors need to process the data. They find a nonlinear relationship between the cost of acquiring these skills and the financial market efficiency. Dugast and Foucault (2023) distinguish the data abundance and data mining cost. These two distinct dimensions of data technologies have different implications for financial markets. Complementing the existing literature on the economic consequences of data technologies on financial markets, this paper investigates how data market structures shape data access and, in turn, affect financial market quality. Our model predicts that data monopoly has the incentive to block the entry of alternative data into financial markets. Furthermore, we provide empirical evidence showing that data monopolies negatively affect market efficiency. It also joins a growing body of empirical work

examining the impact of technology and networks on market microstructure (O'Hara et al., 2018).

The literature on information sales in financial markets discusses the optimal selling strategies of information sellers and their impacts on market efficiency. The theoretical developments revolve around investigating different market microstructures, including perfectly competitive rational expectations equilibrium (Admati and Pfleiderer, 1986, 1990; Veldkamp, 2006; Cespa, 2008) and strategic trading (Admati and Pfleiderer, 1988; Garcia and Sangiorgi, 2011; Chen and Wilhelm Jr, 2012). Some studies analyze the sale of price information as a specific type of information product (Cespa and Foucault, 2014; Easley et al., 2016). Our study complements this literature in two ways. First, we examine the joint sale of fundamental and non-fundamental information. This introduces a new externality, the strategic substitutability between the two types of data, which alters the elasticity considerations in the information seller's pricing problem. Second, we shift the focus from financial market structures, which is the emphasis of most existing work, to data market structures. While the literature typically studies monopolistic data sales, we provide a more detailed analysis of the consequences of introducing competition and allowing horizontal integration.

Our research is also related to the literature on the data economy and the market structures of data sales. There is a growing literature documenting the data externalities that arise from consumers sharing data (Acemoglu et al., 2022; Bergemann and Bonatti, 2023; Choi et al., 2019; Ichihashi, 2021b). In our model, the data market creates externalities for the financial market. Data investment impacts both data acquisition and market quality within the financial market. We examine how data sales can generate externalities, leading to complex impacts on the financial market quality. A recent body of work has expressed concerns about the market power arising from data, both in product markets (Begenau et al., 2018; Prüfer and Schottmüller, 2021; Eeckhout and Veldkamp, 2022), and platforms (Bergemann and Bonatti, 2023). These papers describe two channels through which data enhances market power. The first is the scale effect channel, where firms accumulate data to produce at lower marginal costs. The second is the information advantage channel, where firms use consumer information to design personalized products or engage in price discrimination. Complementing these papers, we focus on the competition and market power within data markets, which arises from the data vendor's ability to manipulate the allocation of data. It implies novel implications for data market regulation. Our focus on competition within the data market is similar to Ichihashi (2021a) and Liu et al. (2024). Ichihashi (2021a) shows that due to the nonrivalry of data, introducing competition among data intermediaries pricing user data may lead to excessively low data prices. Liu et al. (2024) models the strategic

substitutability of data and the lack of commitment power by data vendors. In a dynamic model, data vendors compete with their future selves. As a result, they state that despite the data market being a monopoly, the market power of data vendors remains limited. As a complement, we construct a two-dimensional data structure with strategic substitutability. Our findings suggest that data vendors can manipulate information allocation in downstream markets by controlling the circulation of both data. This implies that the market power of monopolistic data vendors remains a significant concern.

2 Model Setup

We consider an economy where multiple upstream data providers decide the production and supply of information for an imperfectly competitive financial market. This framework allows us to investigate the strategic implications of data market structure on financial market quality—a central concern for regulators overseeing "Open Data" and "Open Finance" mandates. Unlike the conventional view that informational monopolies are the sole source of market friction, our setup treats information as a nonrival infrastructure, where the *composition* of data types (fundamental versus non-fundamental) is endogenously determined by the degree of competition among vendors.

The Financial Market. The financial market consists of a single risky asset with a terminal value $\tilde{\theta}$. We normalize the common prior mean of $\tilde{\theta}$ to zero. The prior belief is $\tilde{\theta} \sim N(0, \tau_\theta^{-1})$, where τ_θ represents the public precision of the fundamental value. The market is populated by three classes of risk-neutral agents: a pool of N potential rational traders, noise traders, and a perfectly competitive market maker. Specifically, noise traders submit an aggregate, random order $\tilde{u}$ to satisfy exogenous liquidity needs. This shock is independent of $\tilde{\theta}$ and distributed as $\tilde{u} \sim N(0, \tau_u^{-1})$, where τ_u is the precision of the noise trading shock. The market maker is competitive and risk-neutral, observing only the aggregate order flow y and setting the price $p(y)$ to earn zero expected profit. Trading occurs via a batch-order mechanism, common in dealer markets, where orders are submitted simultaneously and cannot be conditioned on the contemporaneous clearing price. See Remark 1 below for further discussion on batch auction.

The Data Market and Open Data Policy. The data market is defined by a set of $J \geq 1$ data providers. In line with the authorized redistribution models seen in global market data reforms, these providers monetize access to proprietary datasets. We assume two categories

of data can be generated at a fixed cost, reflecting the high-upfront/low-marginal-cost nature of digital assets:

- **Fundamental Data (s_f):** Provides a noisy signal regarding the asset's payoff:

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

  where $\tau_f > 0$. The fixed cost of production is c_f. In policy terms, this represents "Fundamental Transparency"—the sharing of credit histories or corporate disclosures to reduce information asymmetry.

- **Non-Fundamental Data (s_n):** Provides a noisy signal regarding the noise trading volume:

$$s_n = u + \epsilon_n, \quad \epsilon_n \sim N(0, \tau_n^{-1}) \tag{2}$$

  The fixed cost of production is c_n. This represents "Order-Flow Transparency," analogous to the real-time transaction data utilized in Payment for Order Flow (PFOF) schemes.

The two signals exhibit a hierarchical informational structure. The fundamental signal s_f has standalone value, as it enables a trader to form a posterior belief about the asset's payoff θ. By contrast, the non-fundamental signal s_n has no independent value for a risk-neutral trader. Because u is orthogonal to θ, s_n conveys no information about fundamentals. Its value is purely strategic: conditional on observing s_f, s_n allows a trader to better anticipate aggregate order flow y and optimally trade against the market maker's pricing rule.

Remark 1 (Price Observability vs. Batch Auction)**.** *This distinction is model-dependent. In rational expectations equilibrium models with price observability, non-fundamental signals may help traders infer θ from prices (Ganguli and Yang, 2009), but such environments typically admit multiple equilibria in information acquisition, complicating the analysis of data provision. To maintain tractability and isolate the strategic role of order-flow information, we abstract from price observability and adopt a batch-order framework in which traders submit orders simultaneously and cannot condition on the market-clearing price. This institutional feature is standard in dealer markets.*

Information Structure and Data Sales Regimes. The data market structure endogenously partitions rational traders into three groups: N_{fn} traders who acquire the full information package $\{s_f, s_n\}$ (*FN-type*); N_f traders who acquire only the fundamental package

$\{s_f\}$ (*F-type*); N_n traders who acquire only the fundamental package $\{s_n\}$ (*N-type*);and the remaining N_u agents who remain uninformed. The resulting triple (N_{fn}, N_f, N_n, N_u) defines the financial market's information structure.

We analyze the consequences of open data policy by evaluating six regimes of data sales, ranging from integrated monopolies to competitive multi-provider landscapes:

- **Regime I (Integrated Monopoly):** A single vendor controls the entire sales of both s_f and s_n.
- **Regime II (Segmented Monopoly):** Two separate monopolists control sales of s_f and s_n independently.
- **Regime III (Segmented Competition):** Multiple independent sellers ($J \geq 2$) compete within the s_f and s_n markets separately.
- **Regime IV (Integrated Competition):** At least two providers ($J \geq 2$) possess the capability to sell both types of data.
- **Regime V (Asymmetric Oligopolies):** A "generalist" vendor (e.g., a global aggregator) competes with a "specialist" vendor restricted to fundamental data.
- **Regime VI (Asymmetric Oligopolies):** A "generalist" vendor (e.g., a global aggregator) competes with a "specialist" vendor restricted to non-fundamental data.

Timing and Equilibrium. The game unfolds over three sequential stages.

Stage 0: Data Provision. At $t = 0$, the data seller(s) receive an initial endowment of information and determine their optimal selling strategies. The sellers choose the number of contracts for each trader type, (N_{fn}, N_f, N_n), to maximize expected profits.

Stage 1: Trading and Order Submission. At $t = 1$, potential traders acquire data, realize their specific types, and observe their respective signals. Based on this private information, N_{fn} FN-type traders, N_f F-type traders, N_n N-type traders, and N_u U-type traders, along with liquidity (noise) traders, submit their orders $(x_j^{FN}, x_i^F, x_i^N, u)$ simultaneously.

Stage 2: Pricing and Settlement. At $t = 2$, the market maker observes only the aggregate order flow, y, defined as:

$$y = \sum_{i=1}^{N_{fn}} x_i^{FN} + \sum_{j=1}^{N_f} x_j^F + \sum_{k=1}^{N_n} x_i^N + \sum_{l=1}^{N_u} x_l^N + u \tag{3}$$

The market maker sets the price $p(y)$ as the expectation of the asset value conditional on the observed flow. All payoffs are then realized.

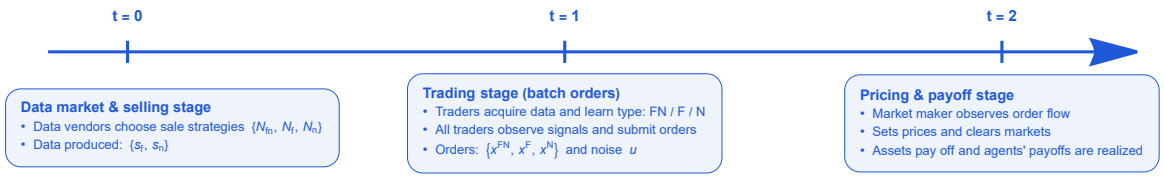


Figure 1: Timeline.

We characterize the perfect Bayesian equilibrium of the game by backward induction and restrict attention to a symmetric linear equilibrium. In this equilibrium, traders of the same information type employ identical linear trading strategies, and the market maker sets prices as a linear function of aggregate order flow, $p(y) = \lambda y$.

Definition 1. *An equilibrium consists of (i) a linear pricing rule $p^*(y) = \lambda y$ for the risky asset, (ii) equilibrium orders $(x_i^*)_i$ submitted by investors, and (iii) optimal data sales $(N_j^*)_{j\in\{f,n,fn\}}$, such that:*

1. *the data seller chooses (N_{fn}^*, N_f^*, N_n^*) to maximize profit,*

$$\Pi(N_{fn}^*, N_f^*, N_n^*, N_u^*) \geq \Pi(N_{fn}, N_f, N_n, N_u), \quad \text{for all } (N_{fn}, N_f, N_n, N_u) \in \mathbb{N}^4;$$

2. *given their information sets and the pricing rule, all investors optimally choose their orders to maximize expected trading profits;*

3. *the market maker sets prices competitively to satisfy the zero-profit condition,*

$$p^*(y) = \mathbb{E}[\theta \mid y].$$

3 Strategic Trading with Non-fundamental Data

We solve the game by backward induction. The analysis begins with the trading subgame, taking the information allocation (N_{fn}, N_f, N_n, N_u) as given. We then characterize the value of information and, finally, examine how the data allocation affects market quality. This allows us to connect equilibrium trading behavior to policy-relevant features of data access.

3.1 Financial Market Equilibrium

In the trading subgame, the market's information structure (N_{fn}, N_f, N_n, N_u) is common knowledge. FN-type traders observe $\mathcal{I}_{fn} = \{s_f, s_n\}$, F-type traders observe $\mathcal{I}_f = \{s_f\}$, N-type traders observe $\mathcal{I}_n = \{s_n\}$, and U-type traders are uninformed (i.e., $\mathcal{I}_u = \text{Var}nothing$). All agents are risk-neutral. We focus on a symmetric linear equilibrium in which traders' strategies are linear in their signals and the market maker sets prices as a linear function of aggregate order flow y. We begin with uninformed traders, who possess no private information.

Lemma 1 (No trading by uninformed traders)**.** *An uninformed risk-neutral trader optimally submits a zero order (i.e., $x_u = 0$).*

Given Lemma 1, aggregate order flow is

$$y = N_{fn}x^{fn} + N_f x^f + N_n x^n + u.$$

We conjecture a linear pricing rule $p(y) = p_0 + \lambda y$. The market maker's break-even condition implies $p(y) = \mathbb{E}[\theta \mid y]$. Since $\mathbb{E}[\theta] = 0$, equilibrium trading strategies are linear in zero-mean signals, and noise trading u has zero mean, we have $\mathbb{E}[y] = 0$ and hence $p_0 = 0$. The pricing rule therefore simplifies to $p(y) = \lambda y$.

We conjecture linear strategies for informed traders. An F-type trader i submits

$$x_i^f = \beta_f \hat{\theta}_f, \qquad \hat{\theta}_f \equiv \mathbb{E}[\theta \mid s_f] = \frac{\tau_f}{\tau_\theta + \tau_f} s_f.$$

An N-type trader submits

$$x_i^n = \gamma_n \hat{u}_n, \qquad \hat{u}_n \equiv \mathbb{E}[u \mid s_n] = \frac{\tau_n}{\tau_u + \tau_n} s_n.$$

An FN-type trader conditions trading on both signals. By orthogonality, $\mathbb{E}[\theta \mid s_f, s_n] = \hat{\theta}_f$ and $\mathbb{E}[u \mid s_f, s_n] = \hat{u}_n$, so

$$x_j^{fn} = \beta_{fn}\hat{\theta}_f + \gamma_{fn}\hat{u}_n.$$

Each trader i acts as a Cournot competitor, choosing x_i to maximize

$$\mathbb{E}[x_i(\theta - \lambda y) \mid \mathcal{I}_i],$$

taking others' strategies as given. Let $y_{-i} \equiv \sum_{j \neq i} x_j + u$, so total order flow is $y = x_i + y_{-i}$.

The first-order condition is

$$\mathbb{E}[\theta - \lambda y \mid \mathcal{I}_i] - \lambda x_i = 0,$$

or equivalently,

$$2\lambda x_i = \mathbb{E}[\theta \mid \mathcal{I}_i] - \lambda\, \mathbb{E}[y_{-i} \mid \mathcal{I}_i].$$

Together with the pricing rule $\lambda = \mathrm{Cov}(\theta, y)/\mathrm{Var}(y)$, these conditions yield a unique linear equilibrium.

Proposition 1 (Unique Symmetric Linear Equilibrium)**.** *Let $M \equiv N_f + N_{fn}$ denote the number of traders with fundamental data and $K \equiv N_n + N_{fn}$ the number with non-fundamental data. A unique symmetric linear equilibrium exists and is characterized by:*

1. ***Fundamental trading intensity.*** *All traders with fundamental data trade with the same intensity,*

$$\beta \equiv \beta_f = \beta_{fn} = \begin{cases} \frac{1}{\lambda(M+1)}, & M > 0, \\ \textit{undefined}, & M = 0; \end{cases}$$

2. ***Non-fundamental trading intensity.*** *All traders with non-fundamental data trade with the same intensity,*

$$\gamma \equiv \gamma_n = \gamma_{fn} = \begin{cases} -\frac{1}{K+1}, & M > 0,\ K > 0, \\ 0, & M = 0,\ K > 0, \\ \textit{undefined}, & K = 0; \end{cases}$$

3. ***Pricing rule.*** *The market maker's price impact is*

$$\lambda = \begin{cases} \frac{\sqrt{M}}{M+1} \sqrt{\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}}\, \frac{1}{\sqrt{h(K)}}, & M > 0, \\ 0, & M = 0, \end{cases}$$

where Define

$$h(K) \equiv \mathrm{Var} K\gamma\, \frac{\tau_n}{\tau_n + \tau_u}\, s_n + u = \frac{1}{\tau_u} - \frac{\tau_n}{\tau_u(\tau_n + \tau_u)} \frac{K(K+2)}{(K+1)^2}.$$

Proposition 1 yields four central insights. First, trading on each signal is separable. We have $\beta_f = \beta_{fn}$ and $\gamma_n = \gamma_{fn}$, so possessing non-fundamental data does not alter how a trader

uses fundamental data, and vice versa. Strategies are additive:

$$x^{fn} = \beta\hat{\theta}_f + \gamma\hat{u}_n = x^f + x^n.$$

Economically, FN-type traders behave as if they simultaneously participate in two independent Cournot games—one over fundamentals and one over non-fundamental data. This additivity follows from orthogonality and the linear-normal environment.[4]

Second, non-fundamental data enables profitable trading through strategic anticipation of noise. The equilibrium intensity $\gamma = -1/(K+1) < 0$ implies that traders with s_n optimally trade against expected noise orders. This behavior reduces price impact when noise demand is high and increases it when noise demand is low, allowing informed traders to camouflage their trades.

Third, the equilibrium supports profitable traders with only non-fundamental data. Unlike rational-expectations models in which non-fundamental signals improve inference about fundamentals, here profits arise purely from exploiting the market maker's inability to disentangle information from noise. Non-fundamental traders extract rents from order-flow uncertainty without contributing to price discovery. This result highlights that markets for purely non-fundamental data can exist even when such data has zero informational content about fundamentals.[5]

Fourth, non-fundamental trading is profitable only in the presence of fundamental trading. When there are no fundamental traders, aggregate order flow is uninformative about θ, and the market maker optimally sets $\lambda = 0$, eliminating all rents from non-fundamental data. Non-fundamental information is therefore a second-order good whose value derives entirely from the price impact generated by fundamental traders.[6]

[4]AWe conjecture that linearity is not merely a simplifying assumption, but a robust feature of the equilibrium. Allowing FN-type traders to adopt nonlinear strategies with an interaction term $c\,\hat{\theta}_f\hat{u}_n$ yields $c = 0$ in equilibrium, collapsing back to the linear strategy. The intuition is that linear–normal order flow induces a linear pricing rule, $p(y) = \lambda y$; facing linear price impact, traders optimally respond with linear strategies.

[5]This viability can arise in rational-expectations equilibria when non-fundamental information improves the precision of price signals (Ganguli and Yang, 2009; Manzano and Vives, 2011). In the Kyle (1985) framework, however, traders do not observe prices or informative order flow and therefore do not engage in price discovery about θ. Instead, profits arise from exploiting the market maker's informational disadvantage. Because the market maker cannot distinguish non-fundamental trades (from the N-type traders) from liquidity trades u, she must impose a positive price impact ($\lambda > 0$) on aggregate order flow to break even. An N-type trader, observing s_n, anticipates the noise component u and recognizes that it will partially and incorrectly move prices in a direction orthogonal to the fundamental value θ. By submitting orders against expected noise (as indicated by $\gamma < 0$), the N-type trader systematically profits from predictable, non-fundamental price movements. This constitutes pure rent extraction from order-flow uncertainty, with the information rent ultimately paid for by the market maker's inability to separate noise from information.

[6]When $M = 0$, aggregate order flow depends only on non-fundamental shocks (u, ϵ_n) and is orthogonal to the fundamental value θ. The market maker's competitive pricing rule, $p(y) = \mathbb{E}[\theta \mid y]$, therefore collapses to the prior $\mathbb{E}[\theta] = 0$ for all y. Anticipating that order flow is non-informational, the market maker ceases to

3.2 The Value of Data

Following Blackwell (1953), we define the value of data as the difference in ex ante expected profits with and without access to that data. Let π_f, π_n, and π_{fn} denote the expected profits of F-, N-, and FN-type traders, respectively.

Proposition 2. *Given (M, K), the value of data is characterized as follows:*

1. ***Fundamental data.*** *The value of fundamental data, defined as the difference between π_{fn} and π_n, is*

$$v_f(M,K) = \begin{cases} \dfrac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}\dfrac{1}{\lambda(M+1)^2}, & M>0, \\ \textit{undefined}, & M=0. \end{cases}$$

2. ***Non-fundamental data.*** *The value of non-fundamental data, defined as the difference between π_{fn} and π_f, is*

$$v_n(M,K) = \begin{cases} \dfrac{\tau_n}{\tau_u(\tau_u+\tau_n)}\dfrac{\lambda}{(K+1)^2}, & M>0,\ K>0, \\ 0, & M=0,\ K>0, \\ \textit{undefined}, & K=0. \end{cases}$$

3. ***Additivity.*** *Expected profits are additive across data types: $\pi_{fn} = \pi_f + \pi_n = v_f + v_n$.*

This proposition quantifies the private value of fundamental and non-fundamental data. The term v_f denotes the ex ante profit from trading on s_f, shared among all traders with access to fundamental information, while v_n denotes the ex ante profit from trading on s_n, shared among traders who possess non-fundamental data. The proposition further shows that the value of the full data package is additively separable into the values of fundamental and non-fundamental information.

This additive separability follows from the orthogonality of the underlying signals and implies that the two data types are neither complements nor substitutes in value creation. Access to s_n does not increase the marginal value of s_f, and vice versa; rather, each signal generates an independent profit stream. As a result, the data provider faces a simple problem of selling two distinct goods, rather than a bundled product with interactive value.

On the other hand, data sales generate externalities among data buyers. In particular, the acquisition of fundamental and non-fundamental data can exhibit *strategic substitutability*

impose an upward-sloping price schedule. In such a market, an N-type trader—despite observing s_n—cannot earn profits, confirming that non-fundamental data has no private value in isolation and derives its strategic value solely from exploiting the price impact ($\lambda > 0$) induced by fundamental traders.

or *complementarity*. These forces are summarized in Lemma 2.

Lemma 2. *The values of fundamental and non-fundamental data, v_f and v_n, exhibit externalities in data sales:*

1. *Strategic substitutability. The value of fundamental data v_f is strictly decreasing in $M > 0$ and $K \geq 0$, while the value of non-fundamental data v_n is strictly decreasing in $K > 0$:*
$$\frac{\partial v_f}{\partial M} < 0, \qquad \frac{\partial v_f}{\partial K} < 0, \qquad \frac{\partial v_n}{\partial K} < 0.$$

2. *Non-monotonic externality. The value of non-fundamental data $v_n(M, K)$ is non-monotonic in M. Specifically, $v_n = 0$ at $M = 0$, jumps to its global maximum at $M = 1$, and is strictly decreasing for all $M > 1$.*

Lemma 2 shows that data purchase decisions are predominantly strategic substitutes. When an additional trader acquires data – of either type – they impose a negative financial externality on other informed traders. While this conclusion applies broadly, the underlying economic mechanisms differ sharply between fundamental and non-fundamental data.

For fundamental data, $v_f(M, K)$ reflects two opposing forces as the number of fundamental traders M increases. First, a direct competition effect reduces profits by spreading trading rents across more informed traders, as captured by the $1/(M+1)^2$ term. Second, a price impact effect operates through market liquidity: higher M intensifies competition and lowers the price impact λ, mitigating adverse selection and benefiting informed traders. Our analysis establishes that the direct competition effect always dominates, implying $\partial v_f/\partial M < 0$.

The effect of K on v_f is unambiguous. While non-fundamental traders do not directly compete with fundamental traders, an increase in K worsens market liquidity by making aggregate order flow more informative about fundamentals. Anticipating greater adverse selection, the market maker raises the price impact λ, which strictly reduces the profitability of fundamental trading. Hence, $\partial v_f/\partial K < 0$.

The value of non-fundamental data, $v_n(M, K)$, is governed by similar but distinct forces. An increase in K again generates a direct competition effect that erodes rents, alongside a price impact effect that operates in the opposite direction. Since v_n reflects rents extracted from predictable mispricing, a higher λ amplifies price deviations from fundamentals, increasing the exploitable signal for non-fundamental traders. Nevertheless, the negative competition effect dominates, ensuring that $\partial v_n/\partial K < 0$.

Finally, Lemma 2 uncovers a non-monotonic externality with respect to M. As shown in Proposition 2, $v_n = 0$ when $M = 0$, since prices are uninformative. Introducing the

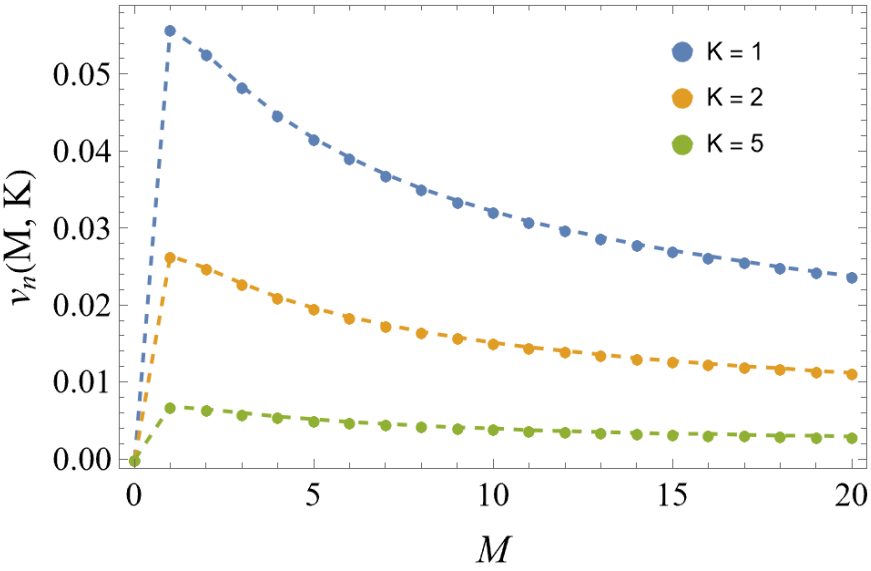


Figure 2: The value of non-fundamental data.

Notes: The figure illustrates how v_n varies with M for different fixed values of $K \in \{1, 2, 5\}$. Parameters are set as: $\tau_n = \tau_u = \tau_f = \tau_\theta = 1$.

first fundamental trader ($M = 1$) generates a positive price impact and creates maximal mispricing rents, causing v_n to jump to its global maximum. For all $M > 1$, further entry of fundamental traders improves liquidity and compresses mispricing, leading v_n to decline monotonically. This mechanism is illustrated in Figure 2.

This non-monotonicity identifies a precise externality that a data monopolist has incentives to internalize: restricting both M and K preserves illiquidity and sustains rents from non-fundamental data. Consequently, optimal data sales by a monopolist are more limited than in standard monopoly problems, reflecting the endogenous interaction between information structure and market liquidity.

3.3 Data Sales and Market Quality

Before solving for the endogenous equilibrium information structure (M, K), we first examine how the exogenous number of traders of each type affects financial market quality. This exercise isolates the economic consequences of data dissemination decisions and provides guidance for open data policies that expand access to different types of information. We focus on four dimensions of market quality: market depth, informational efficiency, order flow volatility, and noise trader welfare. Throughout this subsection, unless otherwise stated, we restrict attention to the case $M > 0$, so that at least one trader is informed about fundamentals.

Market depth. Market depth (MD) measures the ability to trade large quantities quickly and at low cost with a small price impact, and thus is defined as the inverse of price impact (that is, $MD \equiv 1/\lambda$). We show in Appendix A that

$$\frac{\partial MD}{\partial M} > 0, \qquad \frac{\partial MD}{\partial K} < 0.$$

An increase in M intensifies competition among fundamental traders. To capture a larger share of trading profits, each trader responds more aggressively to her signal s_f, which accelerates information aggregation in prices. This competition effect dominates the associated increase in informed trading, lowering the market maker's adverse selection risk for a given order and thus reducing λ. As a result, market depth increases.

In contrast, an increase in K unambiguously reduces market depth. Non-fundamental traders trade against noise orders, thereby reducing the noise component in aggregate order flow y. This makes y a cleaner signal of fundamentals. Anticipating greater adverse selection, the market maker raises λ, leading to a decline in market depth. From a policy perspective, broad access to non-fundamental order-flow-related data may therefore reduce liquidity, even though such data do not convey fundamental information.

Informational efficiency. Informational efficiency (IE) captures how well prices reflect the fundamental value of the asset. We measure it by the precision of the posterior belief about θ conditional on the price, $IE \equiv 1/\mathrm{Var}(\theta \mid p)$. In our model,

$$IE = \left(\frac{1}{\tau_\theta} - \frac{M}{M+1} \frac{\tau_f}{\tau_\theta(\tau_f + \tau_\theta)} \right)^{-1},$$

which implies

$$\frac{\partial IE}{\partial M} > 0, \qquad \frac{\partial IE}{\partial K} = 0.$$

As M increases, more fundamental information is impounded into prices through trading, improving price informativeness. By contrast, non-fundamental data have no effect on informational efficiency. Although non-fundamental trading reduces noise in order flow and induces the market maker to adjust the pricing rule, fundamental traders optimally scale back their trading intensity in response. These effects exactly offset, leaving price informativeness unchanged. This result highlights a key distinction for open data policy: expanding access to fundamental data improves price discovery, whereas expanding access to non-fundamental data does not.

Order flow volatility. Order flow volatility provides a liquidity-based measure of market activity. In our model,

$$\mathrm{Var}(y) = (M+1)h(K), \tag{4}$$

so that

$$\frac{\partial \mathrm{Var}(y)}{\partial M} > 0, \qquad \frac{\partial \mathrm{Var}(y)}{\partial K} < 0.$$

Holding K fixed, additional fundamental traders increase informed trading volume, raising the volatility of aggregate order flow. In contrast, a larger K allows non-fundamental traders to more effectively offset noise shocks by trading against u, thereby smoothing order flow by absorbing u and reducing volatility. This pattern cautions against interpreting low volatility as a sufficient indicator of market liquidity: greater dissemination of non-fundamental data can simultaneously lower volatility and reduce market depth.

Noise trader welfare. Noise traders trade for non-speculative reasons and demand liquidity. Their ex ante expected welfare is $W_{noise} \equiv \mathbb{E}[(\theta - p)u]$ and can be computed as:

$$W_{\text{noise}} = -\lambda \left(\frac{1}{\tau_u} - \frac{K}{K+1} \frac{\tau_n}{\tau_u(\tau_n + \tau_u)} \right).$$

An increase in M benefits noise traders:

$$\frac{\partial W_{\text{noise}}}{\partial M} \propto -\frac{\partial \lambda}{\partial M} > 0.$$

Thus, greater competition among fundamental traders lowers price impact, allowing noise traders to transact at prices closer to fundamentals.

The effect of K combines two opposing forces:

$$\frac{\partial W_{\text{noise}}}{\partial K} = \underbrace{-\frac{\partial \lambda}{\partial K} \left(\frac{1}{\tau_u} - \frac{K}{K+1} \frac{\tau_n}{\tau_u(\tau_n + \tau_u)} \right)}_{\text{price impact effect } <0} + \underbrace{\frac{\lambda \tau_n}{\tau_u(\tau_n + \tau_u)} \frac{1}{(K+1)^2}}_{\text{noise absorption effect } >0}.$$

The price impact effect reflects higher λ, which worsens execution prices for a given noise order. The noise absorption effect arises because non-fundamental traders optimally trade against noise trading, partially absorbing these orders before they reach the market maker. The latter effect always dominates, implying that noise trader welfare strictly increases in K. Thus, while non-fundamental data may harm liquidity, it can nevertheless benefit liquidity demanders.

Summary. The effects of (M, K) on market quality are summarized below.

Lemma 3 (Data Sales and Financial Market Quality)**.**

1. *Market depth strictly increases in M and strictly decreases in K.*
2. *Informational efficiency strictly increases in M and does not depend on K.*
3. *Order flow volatility strictly increases in M and strictly decreases in K.*
4. *Noise trader welfare strictly increases in both M and K.*

Taken together, these results underscore a central tension for open data policy. Expanding access to fundamental data robustly improves price discovery and liquidity, whereas expanding access to non-fundamental data redistributes rents and benefits liquidity demanders but can impair market depth without improving informational efficiency.

3.4 Discussion on Non-fundamental Data

Our analysis of the financial trading subgame and its associated market quality metrics reveals some novel and often counter-intuitive results, all driven by the introduction of non-fundamental data.

The first important finding is non-fundamental data can help traders profit without any fundamental information both ex-ante and ex-post. Several studies also point out the profitability of observing signals of noise trading, but in their setting, non-fundamental information is used to indirectly infer fundamental information by improving the precision of price signals under the REE framework (Ganguli and Yang, 2009; Manzano and Vives, 2011) or removing noise orders from observable total orders under Kyle (1985)'s multi-period framework (Madrigal, 1996; Yang and Zhu, 2020), thereby indirectly inferring informed orders. Traders with non-fundamental information, in these models, eventually make profits from fundamental signals. The novelty of our model lies in the fact that non-fundamental traders do not have any signals about the fundamental payoff.

Our model achieves linear additivity of trading strategies. There is no interaction between trading strategies based on fundamental data and non fundamental data. This is because we have not introduced risk aversion or observable prices (order flow). Our model demonstrates a more direct channel to explain the profitability of non-fundamental trading. The market maker, unable to distinguish informed trades from noise, must set a price $p = \lambda y$ that responds to all order flow. Non-fundamental traders, by observing s_n, can anticipate the

noise component u and understands that this component will mistakenly push the price in a direction uncorrelated with θ. This creates a profitable, non-fundamental information rent.

Noise trading u is independent of the fundamental payoff θ. Fundamental trading x^f is, on average, aligned with θ, whereas non-fundamental trading x^n is aligned against the noise component u. Consequently, when realizations of u and θ have the same sign, non-fundamental trades incur losses; when their signs differ, such trades generate profits. Under symmetry and independence, one might expect these gains and losses to cancel, yielding zero expected profits. However, because the market maker cannot distinguish the source of order flow, profits and losses are asymmetric in magnitude. For a given realization of θ, favorable trades generate larger gains than the losses incurred in unfavorable trades. As a result, non-fundamental trading earns strictly positive expected returns. We illustrate this asymmetry with a simple numerical example.

Example 1 (Profitability of Non-fundamental Trading)**.** *Consider a baseline case with a single fundamental trader ($M = 1$) and a single non-fundamental trader ($K = 1$). Assume both agents receive perfectly informative signals, such that $\tau_f \to \infty$ and $\tau_n \to \infty$. Thus, the fundamental trader observes the asset's payoff θ, and the non-fundamental trader observes the noise trading volume u. Let the prior precisions be normalized to $\tau_\theta = \tau_u = 1$. Under these parameters, the equilibrium coefficients are $\lambda = 1$, $\beta = 0.5$, and $\gamma = -0.5$.*

To illustrate the profit asymmetry, consider two symmetric realizations of u for a fixed fundamental value $\theta = 2$:

1. ***State 1** ($u = 1$)**:** The fundamental trader submits an order $x^f = 1$, while the non-fundamental trader submits $x^n = -0.5$. The resulting aggregate order flow is $y = 1.5$, leading to an equilibrium price $p = 1.5$. The non-fundamental trader incurs a loss of $(\theta - p)x^n = (2 - 1.5)(-0.5) = -0.25$.*

2. ***State 2** ($u = -1$)**:** The fundamental trader remains at $x^f = 1$, but the non-fundamental trader submits $x^n = 0.5$. The aggregate order flow is $y = 0.5$, yielding a price $p = 0.5$. In this state, the non-fundamental trader earns a profit of $(\theta - p)x^n = (2 - 0.5)(0.5) = 0.75$.*

The profits of non-fundamental traders exhibit a fundamental asymmetry across states, characterized by higher gains and lower losses. Specifically, when the realizations of θ and u have the same sign, the aggregate order flow from noise and non-fundamental traders moves the price toward θ, thereby reducing the unit loss. Conversely, when θ and u are of opposite signs, the combined order flow drives the price away from θ, yielding higher unit profits.

Furthermore, under a symmetric probability distribution, the expected profit for non-fundamental traders is positive. Given the symmetry of the normal distribution, for every positive realization of u, there exists a corresponding negative realization with equal density. For any such pair, the loss incurred when u and θ share the same direction is strictly less than the profit earned when they are opposed. Since profits are zero at $u = 0$, the expected profit conditional on $\theta \neq 0$ is strictly positive.

Finally, this logic extends to all realizations of the fundamental value. When $\theta < 0$, non-fundamental traders earn profits for $u > 0$ and incur losses for $u < 0$, again yielding a positive expected value. In the singular case where $\theta = 0$, profits and losses are perfectly symmetric, resulting in an expected profit of zero. Consequently, across the joint distribution of θ and u, non-fundamental traders earn positive unconditional expected profits.

Owing to the linear additivity of trading strategies, the market can be decomposed into two independent components: a fundamental market and a non-fundamental market. The fundamental market consists of fundamental-informed traders, noise traders, and a market maker who sets prices based on aggregate order flow. The non-fundamental market features noise traders demanding liquidity and non-fundamental informed traders committing to supply liquidity at the prices determined in the fundamental market. Viewed through this lens, non-fundamental data functions as an instrument that grants access to the ability to provide liquidity to noise traders. This decomposition offers a simple interpretation of the emergence of payment for order flow (PFOF). Supplying liquidity to noise traders at prevailing market prices can be ex ante profitable, even in the absence of fundamental information or active portfolio management.

These results offer new insights for the policy debate on payment for order flow (PFOF). Existing discussions and regulatory interventions primarily focus on institutional design, such as restricting off-exchange retail trading or strengthening best-execution and disclosure requirements (SEC, 2022, 2024; EU, 2024). Our analysis highlights a complementary channel: the sale of data feeds on retail order flow can generate effects analogous to off-exchange execution under PFOF. From the perspective of fundamental traders, knowing that noise orders are systematically partially offset by non-fundamental traders operating on such data has qualitatively similar—though more frictional—effects to knowing that retail order flow is partially internalized off exchange. In both cases, open access to order-flow information reduces the effective exposure of noise trading to the lit market, reshaping liquidity provision and price formation even when all trades formally remain on exchange.

This perspective suggests that open data policies targeting transaction-level or order-flow data can affect market quality in ways analogous to execution-based reforms such as

payment for order flow. While greater data openness may promote competition and access, it can also replicate liquidity and price-impact consequences traditionally attributed to off-exchange execution. In particular, even when retail brokers do not physically route orders off exchange, the sale or dissemination of retail order-flow data to liquidity providers can generate similar effects. In our model, access to non-fundamental order-flow data improves noise trader welfare but harms fundamental traders and leads to a thinner market. These results underscore that data openness can be as consequential as institutional design, and that regulatory frameworks should jointly consider data access and trading architecture.

Regulating data is inherently more challenging than regulating trading mechanisms, as the scope and content of data products are difficult to delineate. While regulators can explicitly restrict the execution of payment for order flow (PFOF), it is considerably harder to prohibit the sale of data products that embed retail order-flow information. Data vendors can repackage economically equivalent information through alternative formats or secondary aggregation, making enforcement along product boundaries impractical. Moreover, beyond traditional order-flow data, a growing set of non-fundamental data sources has emerged. For example, in China's markets, some platforms construct data products from user-level digital traces—such as asset-specific clicks, follows, and discussion intensity—often segmented by investor asset size. These data convey information about trading pressure without directly revealing transactions.

A more feasible regulatory margin is therefore the concentration of data markets rather than the precise definition of data products. This observation motivates our analysis. In the next section, we study how data sellers, operating under monopoly or competition, endogenously choose data production and sales strategies, which in turn shape the distribution and use of information in financial markets. While a monopolistic seller may restrict data diffusion, it also internalizes negative externalities that competitive data markets tend to ignore. As a result, the effects of data monopoly on market quality are ambiguous and vary across dimensions, revealing a fundamental trade-off for open data policy.

4 Endogenous Data Sales, Financial Market Quality, and Data Regulation

In this section, we analyze the data sellers' strategic choice of quantities for fundamental and non-fundamental information and characterize the resulting optimal market structure. At $t = 0$, sellers determine the mass of traders, M and K, who acquire each signal type. To

evaluate how ownership structures impact market outcomes, we consider six regimes defined in Section 2, including: (i) integrated monopoly, (ii) segmented monopoly, (iii) segmented competition, (iv) integrated competition, (v) asymmetric oligopoly with overlap in fundamental data, and (vi) asymmetric oligopoly with overlap in non-fundamental data. These regimes are distinguished by whether data provision is integrated or segmented and by the degree of competition across sellers.

For each regime, we derive the equilibrium (M, K) and compare the resulting market performance. These comparisons provide a theoretical framework for evaluating open data policies—such as those promoting entry or data sharing—by illustrating how upstream concentration shifts information acquisition and price formation. We find that the optimal regime is sensitive to the specific measure of market quality and depends critically on the scale of the financial market, N.

4.1 Data Sales, Externality, and Financial Market Quality

This subsection examines three benchmark market structures: integrated monopoly (Regime I), segmented monopoly (Regime II), and segmented competition (Regime III). These cases serve as theoretical anchors to isolate the effects of cross-market internalization and structural segmentation. By comparing these regimes, we identify the fundamental economic forces that govern the boundaries of data provision and determine the equilibrium supply of fundamental versus non-fundamental information.

Integrated monopoly (Regime I). Consider a monopolist who jointly controls the supply of both fundamental data (s_f) and non-fundamental data (s_n). The monopolist chooses quantities (M, K) to maximize total profit $\Pi(M, K) = M \cdot v_f(M, K) + K \cdot v_n(M, K)$, subject to the capacity constraints $0 \leq M, K \leq N$.[7]

The monopolist's optimality condition reflects two distinct forces. First, **intra-market competition** dictates that increasing the quantity of one data type reduces its marginal value to traders—a standard quantity-setting problem. Second, a **cross-market externality** exists: the sale of non-fundamental data s_n imposes a negative externality on the value of fundamental data s_f. An integrated monopolist internalizes this effect. For a given M,

[7]Here, we assume the data vendor captures the full surplus; results are robust to Nash bargaining over the trading surplus.

the marginal profit with respect to K is:

$$\frac{\partial \Pi}{\partial K} = \underbrace{\left(v_n(M,K) + K\frac{\partial v_n}{\partial K}\right)}_{\text{Marginal Revenue from } s_n} + \underbrace{\left(M\frac{\partial v_f}{\partial K}\right)}_{\text{Externality on } s_f} < 0. \tag{5}$$

The monopolist recognizes that s_n sales degrade market depth, which in turn reduces the profitability of fundamental trading and the achievable revenue from s_f. Because s_f is the primary driver of price discovery and surplus, the cannibalization cost outweighs the direct revenue gains from s_n. Consequently, the integrated monopolist finds it optimal to suppress the non-fundamental market entirely. Setting $K^I = 0$, the problem simplifies to:

$$\max_{0 \le M \le N} \Pi(M, 0) = M \cdot v_f(M, 0).$$

This collapses to the classic endogenous information sales problem. As in Admati and Pfleiderer (1988), the seller maximizes the value of the signal by minimizing competition among informed traders, leading to the following proposition:

Lemma 4 (Optimal Data Sales under Integrated Monopoly)**.** *The integrated monopolist provides fundamental data to a single trader and withholds non-fundamental data from the market. The unique equilibrium is:*

$$M^I = 1, \quad K^I = 0. \tag{6}$$

Lemma 4 highlights that while an integrated monopoly effectively "bans" non-fundamental data to protect the value of fundamental information, it does so at the cost of extreme restriction in data circulation. This benchmark suggests that while integration may eliminate noise data, it maximizes market illiquidity by concentrating information in the hands of a single trader.

Segmented monopoly (Regime II). In a segmented market structure, two independent monopolists separately control the supply of fundamental and non-fundamental data. The fundamental data monopolist chooses M to maximize $\Pi_f(M,K) = M \cdot v_f(M,K)$, while the non-fundamental monopolist chooses K to maximize $\Pi_n(M,K) = K \cdot v_n(M,K)$. They choose M and K simultaneously, forming mutual best response.

For the fundamental data monopolist, the objective function $\Pi_f(M,K)$ increases in $\frac{\sqrt{M}}{M+1}\sqrt{h(K)}$. Because $\sqrt{h(K)} > 0$, the best-response for the fundamental monopolist is

to set $M^{II} = 1$. Furthermore, given $M = 1$, the non-fundamental data monopolist solves:

$$\max_{0 \le K \le N} \Pi_n(1, K) = K \cdot v_n(1, K). \tag{7}$$

Critically, the segmented monopolist for non-fundamental data fails to internalize the negative externality $M\frac{\partial v_f}{\partial K}$ imposed on the fundamental data market. Instead, the seller expands K until marginal revenue is exhausted. From Lemma 2, $v_n(1, K) > 0$ for $K > 0$ given $M = 1$, while $\Pi_n(1, 0) = 0$. The first-order condition is:

$$v_n(1, K) + K\frac{\partial v_n(1, K)}{\partial K} = 0. \tag{8}$$

We show that there exists a unique solution, denoted by $\widehat{K}^{II} \ge 1$, to Eq. (8).

Lemma 5 (Optimal Data Sales under Segmented Duopoly)**.** *Let $\rho_n \equiv \tau_n/\tau_u > 0$, and define*

$$A \;\equiv\; 2^{-1/3}\left(27\rho_n + 16 - 3\sqrt{3}\,\sqrt{27\rho_n^2 + 32\rho_n}\right)^{1/3}, \qquad \widehat{K}^{II} \;\equiv\; \frac{1}{3}\left(\frac{4}{A} + A - 1\right).$$

Under a segmented monopoly, the fundamental data seller optimally restricts supply to a single trader, while the non-fundamental data seller supplies a strictly positive quantity of data. The unique equilibrium allocation is given by

$$M^{II} = 1, \qquad K^{II} = \min\{N, \widehat{K}^{II}\} \ge 1. \tag{9}$$

The optimal data allocation in a segmented data market illustrates how the failure to internalize cross-market externalities induces the sale of non-fundamental data. Even this limited form of duopoly competition promotes the dissemination of non-fundamental information that an integrated monopolist would optimally suppress, highlighting a potential trade-off faced by open data policies that promote entry and unbundling in data markets.

Segmented competition (Regime III). Suppose there are $J_f \ge 2$ sellers of fundamental data s_f and $J_n \ge 2$ sellers of non-fundamental data s_n. Sellers engage in Cournot competition within their respective markets and do not cross-sell. A fundamental data seller $j \in \{1, \ldots, J_f\}$ chooses quantity m_j to maximize $m_j\, v_f(m_j + m_{-j}, K)$, taking rivals' quantities $m_{-j} = \sum_{i \ne j} m_i$ as given. Similarly, a non-fundamental data seller $j \in \{1, \ldots, J_n\}$ chooses k_j to maximize $k_j\, v_n(M, k_j + k_{-j})$, where $k_{-j} = \sum_{i \ne j} k_i$. Aggregate quantities are $M = \sum_{j=1}^{J_f} m_j$ and $K = \sum_{j=1}^{J_n} k_j$.

Lemma 6 (Optimal Data Sales under Segmented Competition)**.** *Under segmented competition with $J_f \geq 2$ and $J_n \geq 2$, the unique symmetric equilibrium outcome features full market saturation:*

$$M^{III} = N, \qquad K^{III} = N. \tag{10}$$

This outcome stands in sharp contrast to monopoly. Because each Cournot seller internalizes only a fraction of the negative externality from expanding sales, every seller has a strong incentive to deviate from the monopoly allocation and sell additional units. Introducing competition—even with only two sellers—therefore induces maximal dissemination of both fundamental and non-fundamental data.

Unlike standard linear Cournot models, where equilibrium quantities interpolate between monopoly and perfect competition, segmented competition here admits no interior equilibrium. With at least two sellers, the market outcome coincides with perfect competition and delivers full saturation. As $N \to \infty$, equilibrium prices are driven to zero.

This result reflects the non-rivalry nature of data and the nonlinear structure of information demand. For any aggregate quantity $M < N$, there exists an unserved trader who values an additional unit at $v_f(M+1, K) > 0$ (and analogously for v_n), allowing any firm to profitably expand sales. Moreover, the inverse demand functions v_f and v_n are sufficiently concave, so that the marginal negative externality from additional sales diminishes with market saturation. As a result, private marginal revenue remains positive for all $M < N$, eliminating any interior fixed point. The only Nash equilibrium occurs at full coverage (i.e., $\sum_j m_j = N$). While firm-level allocations $(m_1, \ldots, m_{J_f})$ and $(k_1, \ldots, k_{J_n})$ can be indeterminate, all equilibria yield the same market-level outcome.

Data Market Structure and Financial Market Quality. The preceding analysis characterizes the unique equilibrium information structures under three distinct data market regimes. We now synthesize these results to assess how data monopoly affects financial market quality. Substituting the equilibrium allocations into the market quality metrics derived in Lemma 3 allows us to rank the regimes and highlight the trade-offs faced by a regulator. The results are summarized in the following proposition.

Proposition 3 (Endogenous Data Sales and Financial Market Quality)**.**

Define $\rho_n \equiv \tau_n/\tau_u$, $\widehat{N} \equiv 1 + 2\rho_n + \sqrt{4\rho_n^2 + 3\rho_n}$, and $\overline{N} \equiv \rho_n + \sqrt{\rho_n^2 + \rho_n + 1}$. Then, the effects of data market structure on financial market quality can be summarized as follows:

(i) Market depth (MD): $MD^{II} \leq \min\{MD^I, MD^{III}\}$, and $MD^{III} \leq MD^I$ if and only if $N \leq \widehat{N}$.

(ii) *Informational efficiency* *(IE):* $IE^I = IE^{II} \leq IE^{III}$.

(iii) *Order-flow volatility* *(*$\text{Var}(y)$*):* $\text{Var}^{II}(y) \leq \min\{\text{Var}^I(y), \text{Var}^{III}(y)\}$, *and* $\text{Var}^{III}(y) \leq \text{Var}^I(y)$ *if and only if* $N \leq \overline{N}$.

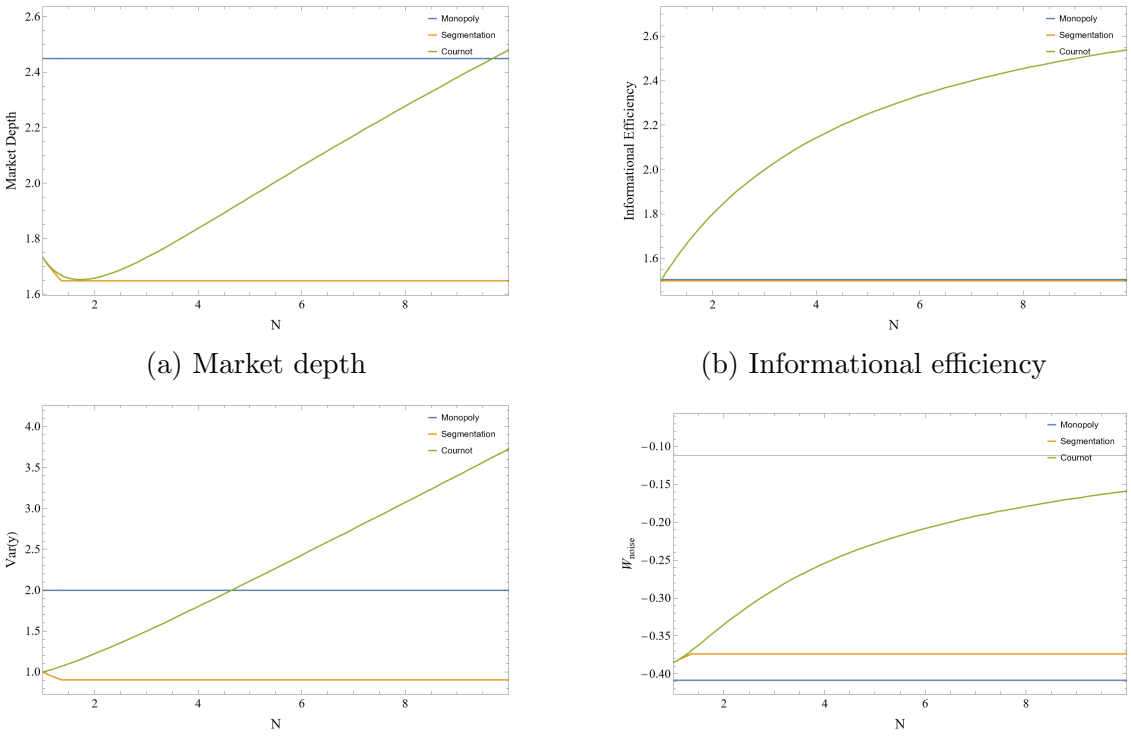
(iv) *Noise-trader welfare* *(*W_{noise}*):* $W^I_{noise} < W^{II}_{noise} \leq W^{III}_{noise}$.

Proposition 3 highlights that no single data-market regime simultaneously optimizes all dimensions of financial market quality. Instead, the optimal regulatory intervention is entirely contingent on the regulator's primary objective. We now provide economic intuition for each of the four comparative-static results.

Market depth. For a regulator aiming to maximize market depth, the optimal policy depends critically on market size N, as illustrated in Figure 3a. The segmented monopoly (MD^{II}) is uniformly dominated. It combines the minimal fundamental competition of an integrated monopoly with the entry of non-fundamental traders, which strictly harms liquidity (Lemma 3). The relevant trade-off is therefore between the integrated monopoly (MD^I) and perfect competition (MD^{III}).

An integrated monopoly suppresses liquidity-reducing non-fundamental trading by setting $K = 0$, but at the cost of minimal fundamental competition ($M = 1$). In contrast, perfect competition maximizes both M and K, generating two opposing effects on market depth. As N increases, the impact of competition is non-monotonic. When N is small, the liquidity loss induced by non-fundamental traders—who render order flow excessively informative for the market maker—dominates the benefit of additional fundamental competition. When N is large, the pro-liquidity effect of intensified fundamental trading prevails. Consequently, when $N \leq \widehat{N}$, the optimal regime is an integrated data monopoly; when $N > \widehat{N}$, competition dominates. Moreover, the threshold $\widehat{N}$ is strictly increasing in ρ_n, the relative precision of non-fundamental data. Higher-quality non-fundamental signals exacerbate liquidity losses, making monopoly—which uniquely suppresses s_n—optimal over a wider range of market sizes.

Informational efficiency. If the regulator's objective is to maximize informational efficiency, perfect competition is unambiguously optimal, as shown in Figure 3b. By Lemma 3, informational efficiency is strictly increasing in the number of fundamental traders and is independent of non-fundamental participation. Any restriction on fundamental data competition therefore strictly reduces price informativeness.



(a) Market depth

(b) Informational efficiency

(c) Order-flow volatility

(d) Noise-trader welfare

Figure 3: Trader Mass N and Financial Market Quality

Notes. Panels (a)–(d) report market depth, informational efficiency, order-flow volatility, and noise-trader welfare, respectively. Parameters are $\tau_\theta = \tau_n = 1$ and $\tau_f = 2$.

Order-flow volatility. For minimizing order-flow volatility, the segmented monopoly is optimal (Figure 3c). Lemma 3 shows that $Var(y)$ increases in M but decreases in K. The segmented monopoly uniquely combines low fundamental entry ($M = 1$), which suppresses informed-trading volatility, with positive non-fundamental participation ($K^{II} > 0$), which hedges noise and further dampens order-flow fluctuations. No other regime simultaneously limits speculative intensity while promoting noise absorption.

Noise-trader welfare. Finally, if the objective is to maximize noise-trader welfare, perfect competition dominates (Figure 3d). Noise-trader welfare is strictly increasing in both M and K. Greater fundamental competition reduces market-power rents and lowers price impact, while increased non-fundamental participation absorbs noise trades and shields noise traders from adverse execution prices.

Taken together, these results reveal a fundamental policy trade-off in data-market regulation. The optimal market structure depends not only on the regulator's objective but also on the underlying market environment. A uniform mandate to promote data competition can be misguided. In particular, when markets are small, a data monopoly—despite being maximally detrimental to price informativeness—can be optimal for a regulator focused

solely on market depth.

4.2 Cross-market Competition, Data Allocation and Regulation

In this subsection, we extend the analysis to strategic regimes in which data sellers compete across overlapping markets. These environments capture interactions between generalist vendors and specialized entrants, highlighting how cross-market strategies shape equilibrium outcomes. When sellers operate across data types, strategic interactions arise across markets, and equilibrium allocations need not be extreme (i.e., 0, 1, or N). Instead, outcomes may be interior. Cross-market sellers partially internalize negative externalities, though to a lesser extent than an integrated monopolist.

4.2.1 Cross-Data Competition and Asymmetric Oligopoly

Integrated competition (Regime IV). Under this regime, there are J sellers, each of whom simultaneously supplies both fundamental and non-fundamental data. Seller $j \in \{1, \ldots, J\}$ solves

$$\max_{0 \leq m_j, k_j \leq N} \Pi_j(m_j, k_j) = m_j v_f(m_j + m_{-j}, k_j + k_{-j}) + k_j v_n(m_j + m_{-j}, k_j + k_{-j}).$$

Relative to Regime III, integration forces competitive sellers to internalize cross-data externalities. This moderates aggressive data provision and may lead to interior outcomes. The equilibrium is characterized as follows.

Lemma 7 (Integrated competition)**.** *Under integrated competition, there exists a symmetric competitive equilibrium. The equilibrium allocation of fundamental and non-fundamental data is given by*

$$M^{IV} = \begin{cases} \min\{\widehat{M}^{IV}, N\}, & \text{if } J < \frac{N+3}{2} \text{ and } \rho_n > \tilde{\rho}_n; \\ N, & \text{otherwise.} \end{cases} \qquad \text{and} \qquad K^{IV} = N.$$

Lemma 7 shows that, in integrated markets, competitive sellers partially internalize cross-data negative externalities. Unlike segmented competition, sellers do not always supply both data types to all traders. While non-fundamental data is always sold to the full market, access to fundamental data may be restricted when non-fundamental signals are sufficiently precise. Intuitively, higher-quality non-fundamental data raises the opportunity cost of expanding fundamental data sales, as additional fundamental information dilutes non-fundamental data

revenues. When this marginal loss exceeds the marginal gain from fundamental data sales, sellers optimally restrict fundamental access.

However, this internalization remains weaker than under an integrated monopoly. Each competitive seller captures only a $\frac{1}{J}$ share of total revenue and thus internalizes only a fraction of the externality. As a result, equilibrium allocations differ sharply across regimes. Whereas an integrated monopolist supplies one unit of fundamental data and no non-fundamental data, integrated competitors always supply non-fundamental data fully and restrict fundamental data only under specific conditions. In the absence of exogenous market segmentation, competition therefore has a disproportionately strong expansionary effect on non-fundamental data provision.

Asymmetric oligopoly with overlapped fundamental data sales (Regime V). This regime features two vendors. Vendor 1 (the generalist) sells both data types, choosing (M_1, K_1). Vendor 2 (the specialist) sells only fundamental data, choosing M_2. Aggregate sales are $M = M_1 + M_2$ and $K = K_1$. Vendor profits are

$$\Pi_1 = M_1 v_f(M_1 + M_2, K_1) + K_1 v_n(M_1 + M_2, K_1), \qquad \Pi_2 = M_2 v_f(M_1 + M_2, K_1).$$

This regime extends Regime II by allowing the non-fundamental data vendor to operate across markets.

Lemma 8 (Asymmetric oligopoly with overlapped fundamental data sales)**.** *Under asymmetric oligopoly with overlap for fundamental data:*

1. *If $\rho_n > \widehat{\rho}_n$, there exists an endogenous segmented monopoly equilibrium with*

$$M^V = 1, \quad K^V = \min\{\widehat{K}^{II}, N\}.$$

2. *If $\rho_n \leq \widehat{\rho}_n$, there exists an asymmetric overlapping equilibrium with*

$$M^V = \min\{\widehat{M}^V, N\}, \quad K^V = \min\{\widehat{K}^V, N\},$$

where $\widehat{M}^V \geq 1$ and $\widehat{K}^V \leq \widehat{K}^{II}$.

Lemma 8 shows that two equilibrium patterns may arise. When non-fundamental data is sufficiently precise, the economy exhibits endogenous segmentation. Despite being permitted to operate across markets, the generalist optimally refrains from selling fundamental data,

leaving that market to the specialist. High cross-market externalities make fundamental data entry unprofitable for the generalist, reproducing the segmented monopoly outcome.

When non-fundamental data precision is lower, both markets exhibit interior allocations. The generalist enters the fundamental market, intensifying competition and increasing total fundamental data sales above the monopoly level, though not necessarily to the competitive benchmark. The generalist internalizes the dilution effect on non-fundamental data and therefore restricts fundamental supply. Anticipating this behavior, the specialist adjusts its own sales. In the non-fundamental market, the generalist also limits supply relative to the segmented monopoly outcome.

Asymmetric oligopoly with overlapped non-fundamental data sales (Regime VI). This regime also features two vendors. Vendor 1 (the generalist) sells both data types, choosing (M_1, K_1), while Vendor 2 (the specialist) sells only non-fundamental data, choosing K_2. Aggregate sales are $M = M_1$ and $K = K_1 + K_2$. Profits are

$$\Pi_1 = M_1 v_f(M_1, K_1 + K_2) + K_1 v_n(M_1, K_1 + K_2), \qquad \Pi_2 = K_2 v_n(M_1, K_1 + K_2).$$

Lemma 9 (Asymmetric oligopoly with overlapped non-fundamental data sales)**.** *Under asymmetric oligopoly with overlap for non-fundamental data, there exists an equilibrium with*

$$M^{VI} = 1, \quad K^{VI} = N.$$

Lemma 9 implies that fundamental data provision remains monopolistic, while non-fundamental data becomes fully competitive. Unlike Regime V, allowing the fundamental data seller to enter the non-fundamental market never induces endogenous segmentation. The cross-market externality from non-fundamental to fundamental data is insufficiently strong to deter entry. As a result, the generalist supplies fundamental data to a single trader, mirroring the integrated monopoly outcome, while competition among non-fundamental sellers drives full market coverage.

4.2.2 Data Allocation Comparison and Financial Market Quality

Data allocation comparison. Finally, we compare the equilibrium data allocations under all regimes. The different consequences arise from the competition effects within the two markets and the cross-market negative externality effects.Tracing the sources of these differences helps to provide insights for policy implications regarding the introduction of competition in data markets and the allowance of horizontal integration.

Proposition 4 (Data Allocation Comparison)**.** *The endogenous data allocation under different regimes is ranked as follows:*

$$1 = M^I = M^{II} = M^{VI} \leq \min\{M^{IV}, M^V\} \leq \max\{M^{IV}, M^V\} \leq M^{III} = N.$$

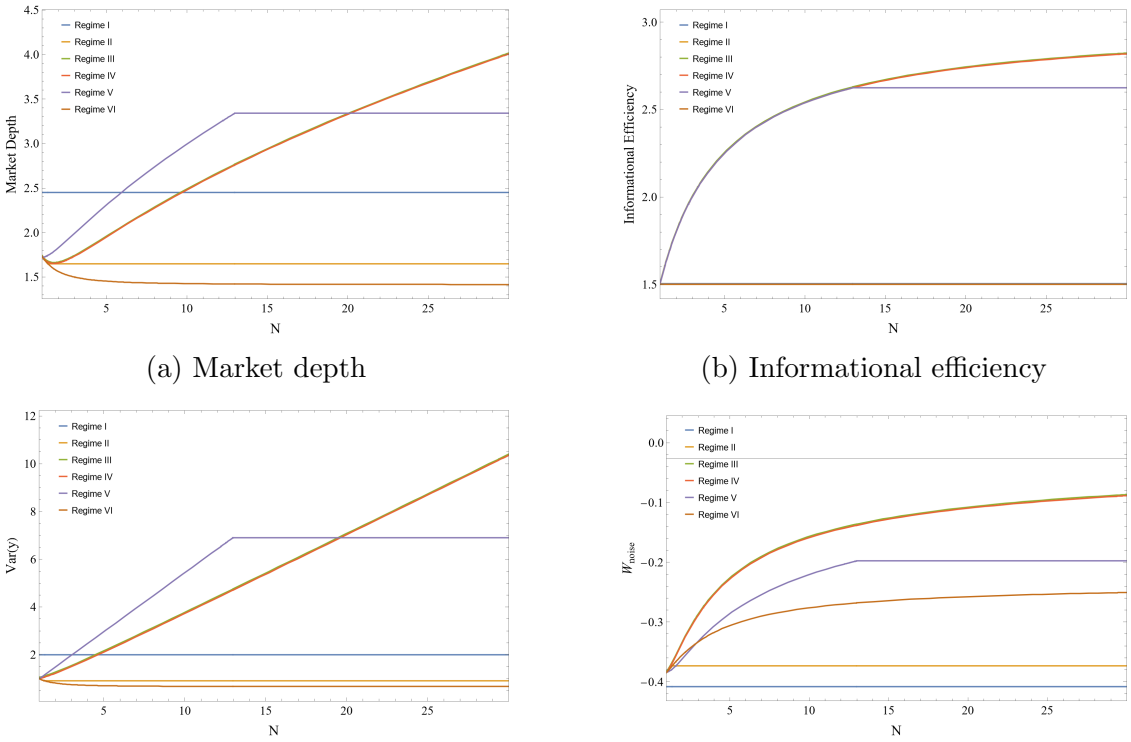
$$0 = K^I \leq K^V \leq K^{II} \leq K^{III} = K^{IV} = K^{VI} = N.$$

Comparing different regimes, we find that the market structure for data sales has some important characteristics. First, if there is only one seller in the fundamental data market, this seller always sells fundamental data to only one trader. Regardless of how many traders receive non-fundamental data or whether this seller enters the non-fundamental data market, this result always holds. When comparing Regime II and Regime IV, we find that when a monopolist of fundamental data is allowed to sell non-fundamental data, the amount of non-fundamental data sold does not affect the strategy of selling only one unit of fundamental data.

Second, when no sellers are engaged in cross-market sales, in any data market with two or more sellers, they always sell as much data as possible until the market reaches its upper limit, N. This result differs from standard linear Cournot competition, primarily due to the nonlinearity in the value of data. The second derivative of data value is greater than zero, meaning that as more data is sold, the dilution of value per unit of data becomes smaller. On the other hand, compared to the monopoly, when competition is introduced, each seller holds only a portion of the market share. As a result, the dilution of existing revenue from selling additional data is smaller, and thus they always seek to occupy as much market share as possible.

Third, allowing horizontal integration strengthens the internalization of cross-market negative externalities. Fundamental data and non-fundamental data are strategic substitutes. When a seller operates in both markets simultaneously, selling one type of data always results in a loss of revenue from the other type. Horizontal integration helps internalize this negative externality, as integrated sellers have the incentive to limit data sales. However, sometimes the externality is so large that the seller will endogenously decide not to engage in cross-market sales, as in the case of Regime V.

Fourth, horizontal integration under symmetric competition does not reduce the sale of non-fundamental data but may limit the sale of fundamental data. Comparing Regime III and Regime IV, we find that if both markets are competitive, allowing horizontal integration may likely limit the sale of fundamental data but does not affect the full sale of non-fundamental data, especially when competition is not intense and non-fundamental



(a) Market depth

(b) Informational efficiency

(c) Order-flow volatility

(d) Noise-trader welfare

Figure 4: Trader Mass N and Financial Market Quality

Notes. Panels (a)–(d) report market depth, informational efficiency, order-flow volatility, and noise-trader welfare, respectively. Parameters are $\tau_\theta = \tau_n = 1$, $\tau_f = 2$ and $J = 2$.

data is sufficiently valuable. This suggests that the negative externality of fundamental data on non-fundamental data is more significant than the reverse externality. Given that non-fundamental data reduces market depth without improving information efficiency, excessive non-fundamental data may not be desirable. Therefore, integration in a competitive environment may lead to worse outcomes.

Data market structure and financial market quality. To visualize the complex trade-offs across the six data market regimes, we perform numerical simulations by varying the market size N. Figure 4 plots the equilibrium outcomes for market quality. Since the baseline effects of monopoly and segmentation have been discussed, we focus here on the strategic implications of the oligopoly structures. First, we find that Integrated Competition (Regime IV) generates the same outcome as the Segmented Competition (Regime III). It implies that when ρ_n is not sufficiently large and there are only 2 sellers, the competitive pressure to capture market share still overwhelms the incentive to internalize cross-market externalities, driving the market toward maximum data dissemination. Second, Regime V exhibits a unique pattern. For small and intermediate N, it tracks the competition regimes

closely, delivering rapid improvements in depth and informational efficiency and an increase in W_{noise}. However, once N exceeds a threshold, key endogenous choices hit an upper bound, after which market depth, informational efficiency, and order-flow variance become flat. Consequently, Regime V is attractive in smaller markets where it unlocks large marginal improvements, but it becomes increasingly dominated as market size grows. Finally, Regime VI performs uniformly worst: depth and informational efficiency remain low and rise only slowly with N, while W_{noise} stays more negative. This confirms that introducing competition solely in the non-fundamental data market yields the dominated outcome: high adverse selection without the compensatory benefit of improved price efficiency.

5 Empirical Evidence on Data Antitrust Regulation

To empirically assess the real-world consequences of data monopoly, we exploit a quasi-natural experiment arising from China's first antitrust intervention in the bond data market. This section describes the institutional background, data, empirical strategy, and the results from a difference-in-differences (DiD) analysis.

5.1 Institutional Background

Our empirical setting centers on an antitrust action against one of the largest bond data vendors in China, hereafter referred to as Firm S.

China's bond market is segmented into an interbank market and an exchange market. The interbank market accounts for approximately 80% of total bond trading volume and is restricted to qualified institutional investors. Relative to the exchange market, it features a smaller number of participants, lower transparency, and more decentralized trading mechanisms, relying heavily on market makers and brokers. In contrast, the exchange market is centralized, accommodates a broader set of participants—including non-bank financial institutions and retail investors—and exhibits higher liquidity. Some bonds are cross-listed across both markets, and certain institutional investors trade across venues. These institutional features align closely with the small-N environments emphasized in our theoretical framework (Amstad and He, 2019).

An important intermediary in the interbank market is the currency broker. Brokers do not take proprietary positions or hold inventories; instead, they match buyers and sellers and earn commissions rather than bid–ask spreads. Broker-matched transactions account for roughly 40% of spot bond trading volume in the interbank market.[8]

[8]See the National Association of Financial Market Institutional Investors, *China Bond Market Develop-*

There are six major bond brokers operating in China. Given limited electronic trading and public quote dissemination, broker quote data constitute a primary source of market information. Firm S's monopolistic position arose from an exclusive agreement with one broker, hereafter Broker X, which prevented other data vendors from accessing Broker X's real-time quotes and transaction records. As a result, Firm S became the sole platform aggregating data from all six brokers, securing a dominant position in the broker data market.

In March 2023, Chinese regulators initiated an antitrust investigation into Firm S's practices. On March 14, authorities prohibited exclusivity clauses between brokers and data vendors, required Firm S to remove restrictive contractual terms, and mandated access to Broker X's data for competing vendors. By March 20, multiple rival platforms had integrated real-time quotes from all six brokers, marking a sharp increase in data market competition.

Regulatory disclosures emphasized three features of Firm S's monopoly: (i) a decade-long exclusive agreement with Broker X, (ii) its position as the sole provider of comprehensive broker data, and (iii) the imposition of restrictive trading terms, including monopoly pricing and product bundling. These practices raised information acquisition costs for investors and limited access to best available quotes, thereby constraining downstream data product development.

We treat this intervention as an exogenous shock to data market concentration. Notably described as China's first antitrust enforcement action involving data resources, the reform followed a prolonged period of effective monopoly and thus represents a structural break. Since competitive conditions changed rapidly after March 20, 2023, we use this date as the policy cutoff in our DiD design.

Importantly, the intervention plausibly affects only data market competition. It did not alter brokers' matching technologies or client relationships, nor did it directly target investor behavior. Data quality remained unchanged, and investor access to broker data was restored within days. Consequently, the reform did not disrupt upstream (broker) or downstream (investor) markets, allowing us to attribute observed effects primarily to reduced data concentration.

5.2 Data and Variable Description

Our analysis uses bond-level data from January 2021 to June 2025. Bond characteristics and trading information are obtained from Choice, a leading Chinese financial data platform, supplemented with data from Wind and the China Bond Information Network. The merged

ment Report (2024).

Table 1: Summary statistics for treated and control bonds.

	Treated		Control	
Variable	Mean	Std. dev.	Mean	Std. dev.
Price deviation (volume-weighted)	1.21%	2.80%	0.93%	2.40%
Turnover	18.21%	25.71%	20.21%	24.51%
outstanding mkt cap (CNY billion)	77.18	89.45	70.63	89.71
Net close price (CNY)	102.04	4.54	101.74	4.00
Adjusted duration	4.28	3.85	4.03	3.70
Convexity	43.69	101.66	39.55	91.49

dataset includes bond classifications, net price valuations, durations, and credit ratings. Choice aggregates transaction data from multiple venues, including CFETS, all six bond brokers, and bond exchanges, and reports daily net prices, trading volumes, and interbank quote statistics.

The empirical strategy relies on a DiD design exploiting differential exposure to broker data. Our sample focuses on interbank market bonds, some of which are cross-listed on exchanges. We exclude bonds issued within six months prior to the intervention or maturing within six months afterward, and we require bonds to trade at least once per month throughout the sample. Due to infrequent daily trading, we aggregate observations to the monthly level. The final sample consists of 700 bonds and 24,025 bond-month observations.

Treatment bonds are those that were ever matched by Broker X prior to the intervention, identified through manual verification. These bonds are most directly affected by the liberalization of data access. Control bonds were never matched by Broker X and serve as a baseline group. We define Post = 1 for observations from March 2023 onward and Post = 0 otherwise. The interaction Treat $\times$ Post captures the causal effect of increased data competition.

We examine two dimensions of market quality. Price informativeness is measured by the relative deviation between market transaction prices and valuation prices:

$$\text{Price deviation}_{i,t} = \frac{\left|\text{Close net price}_{i,t} - \text{Valuation}_{i,t}\right|}{\text{Valuation}_{i,t}}.$$

Market liquidity is proxied by turnover, defined as trading volume relative to outstanding amount:

$$\text{Turnover}_{i,t} = \frac{\text{Volume}_{i,t}}{\text{Outstanding amount}_{i,t}}.$$

All regressions control for bond characteristics, including credit rating, time to maturity, duration, convexity, and bond type, and include bond and month fixed effects. Table 1

reports summary statistics.

5.3 Data Monopoly and Market Quality

We estimate the following DiD specification:

$$Y_{it} = \alpha + \beta(\text{Treat}_i \times \text{Post}_t) + \gamma X_{it} + \mu_i + \lambda_t + \varepsilon_{it},$$

where Y_{it} denotes a market quality outcome for bond i in month t, X_{it} is a vector of controls, μ_i and λ_t are bond and time fixed effects, and ε_{it} is an error term. The coefficient β captures the causal effect of reduced data monopoly.

The results reveal a clear trade-off between informational efficiency and liquidity, consistent with the model's predictions. Panel regressions in Table 2 show that the DiD coefficient is negative and statistically significant at the 1% level across specifications. Following the intervention, price deviations for treated bonds decline by approximately 30–40 basis points relative to control bonds. This improvement reflects enhanced price discovery resulting from broader access to broker quote information.

Event-study estimates in Figure 5 support the identifying assumptions. Pre-intervention coefficients are statistically indistinguishable from zero, while post-intervention effects are persistently negative, indicating a sustained increase in price informativeness rather than a transitory adjustment.

In contrast, Table 3 documents a significant decline in trading activity. DiD coefficients for turnover are negative and robust across specifications, and Figure 6 shows a persistent post-intervention drop for treated bonds. This pattern aligns with our theoretical mechanism: increased dissemination of broker-based, non-fundamental data encourages non-fundamental trading and reduces liquidity, particularly in markets with a limited number of participants. Given the relatively small size of China's interbank bond market, wider data access may also weaken incentives for certain traders to remain active, leading to endogenous exit, as formalized in Wu (2024).

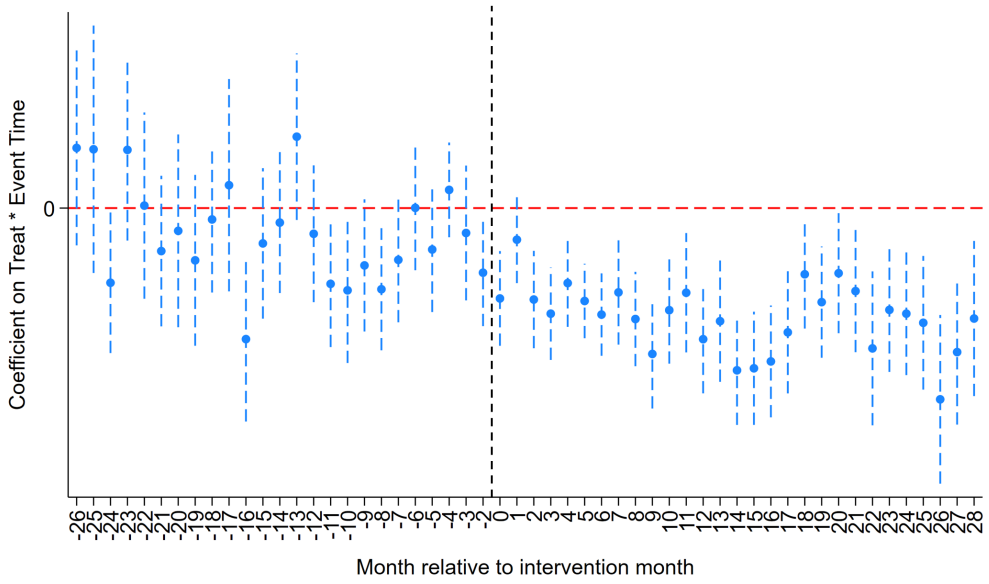


Figure 5: Testing the parallel-trend assumption of the impact on bond price deviation.

Notes: This figure plots event-time coefficients from an event-study specification that interacts the treatment indicator with time dummies relative to the antitrust intervention. The outcome variable is bond price deviation. Pre-intervention coefficients are statistically indistinguishable from zero. Following the intervention, price deviations for treated bonds decline significantly, consistent with improved price informativeness.

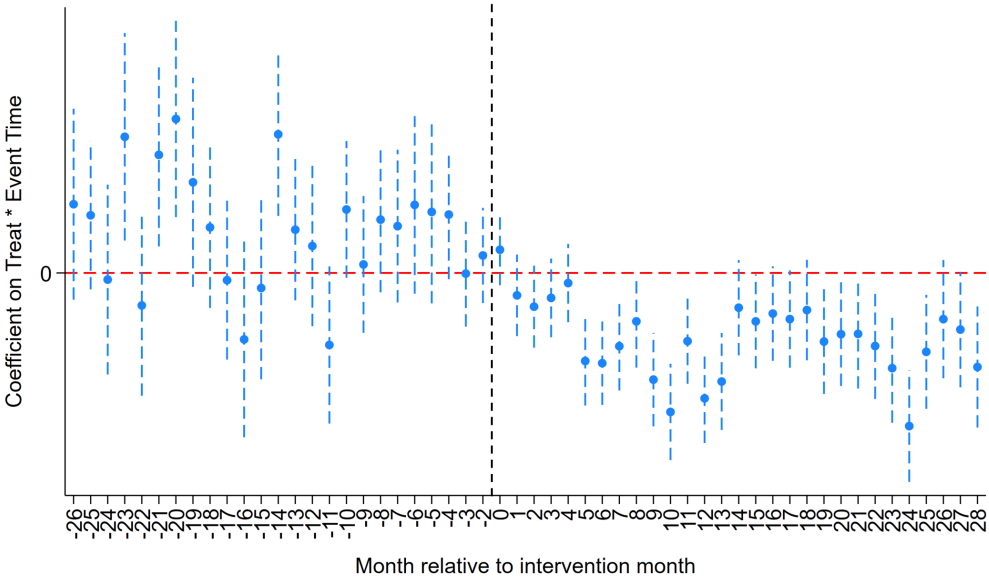


Figure 6: Testing the parallel-trend assumption of the impact on bond turnover.

Notes: This figure plots event-time coefficients from an event-study specification that interacts the treatment indicator with time dummies relative to the antitrust intervention. The outcome variable is bond turnover. Pre-intervention coefficients are statistically indistinguishable from zero. After the intervention, turnover for treated bonds declines significantly.

Table 2: Antitrust Intervention and Bond Price Informativeness

	(1)	(2)	(3)	(4)
Treat × Post	-0.011***	-0.007***	-0.003***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
Controls	No	Yes	Yes	Yes
Time FE	No	No	Yes	Yes
Bond FE	No	No	No	Yes
Observations	24,025	24,025	24,025	24,025
R^2	0.039	0.051	0.063	0.079

Notes: This table reports difference-in-differences estimates of the effect of the antitrust intervention on bond price informativeness, measured by the relative deviation between transaction prices and estimated valuation prices. Treat × Post equals one for treated bonds after March 2023 and zero otherwise. Columns (2)–(4) progressively include bond-level controls, time fixed effects, and bond fixed effects. Standard errors clustered at the bond level are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 3: Antitrust Intervention and Bond Market Liquidity

	(1)	(2)	(3)	(4)
Treat×Post	-0.158***	-0.116***	-0.051***	-0.042***
	(0.010)	(0.009)	(0.011)	(0.014)
Control	No	Yes	Yes	Yes
Date FE	No	No	Yes	Yes
Bond FE	No	No	No	Yes
Observations	24,025	24,025	24,025	24,025
R-sq	0.077	0.145	0.156	0.237

Notes: This table reports the DiD estimates of the effect of the antitrust intervention on bond market liquidity, measured by the turnover, trading volume / market cap of outstanding bonds. The key variable did is an interaction term equal to 1 for treated bonds after March, 2023, and 0 otherwise. Columns (2)–(4) progressively include bond-level controls, bond fixed effects, and time fixed effects. Standard errors clustered at the bond level are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

6 Conclusion

This paper studies how market power in data markets shapes downstream financial market outcomes. Theoretically, we show that fundamental and non-fundamental data have distinct and interacting effects on liquidity and price informativeness. While fundamental data improves price discovery and market depth, non-fundamental data can reduce liquidity without enhancing informativeness. Competition among data sellers can therefore generate sharp, discontinuous changes in market quality.

Empirically, exploiting China's first antitrust intervention in the interbank bond data market, we find that breaking a long-standing data monopoly improves pricing efficiency but reduces turnover. These findings confirm the model's central trade-off and suggest that, in small or thin markets, moderate data concentration may enhance liquidity, whereas competition maximizes information aggregation at the cost of trading activity.

Overall, our analysis highlights that data liberalization policies are not unambiguously welfare improving. Effective regulation must balance information efficiency against liquidity provision and account for market size and structure. The results also provide microfoundations for non-fundamental trading practices, such as payment-for-order-flow, and underscore the importance of considering both informational and liquidity channels in the design of data market regulation.

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar, "Too much data: Prices and inefficiencies in data markets," *American Economic Journal: Microeconomics*, 2022, *14* (4), 218–256.

Admati, Anat R and Paul Pfleiderer, "A monopolistic market for information," *Journal of Economic Theory*, 1986, *39* (2), 400–438.

_ and _, "Selling and trading on information in financial markets," *The American Economic Review*, 1988, *78* (2), 96–103.

_ and _, "Direct and indirect sale of information," *Econometrica: Journal of the Econometric Society*, 1990, pp. 901–928.

Amstad, Marlene and Zhiguo He, "Chinese bond market and interbank market," Technical Report, National Bureau of Economic Research 2019.

Begenau, Juliane, Maryam Farboodi, and Laura Veldkamp, "Big data in finance and the growth of large firms," *Journal of Monetary Economics*, 2018, *97*, 71–87.

Bergemann, Dirk and Alessandro Bonatti, "Data, competition, and digital platforms," *arXiv preprint arXiv:2304.07653*, 2023.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, "Equilibrium fast trading," *Journal of Financial economics*, 2015, *116* (2), 292–313.

Blackwell, David, "Equivalent comparisons of experiments," *The annals of mathematical statistics*, 1953, pp. 265–272.

Cespa, Giovanni, "Information sales and insider trading with long-lived information," *The Journal of Finance*, 2008, *63* (2), 639–672.

_ **and Thierry Foucault**, "Sale of price information by exchanges: does it promote price discovery?," *Management Science*, 2014, *60* (1), 148–165.

Chen, Zhaohui and William J Wilhelm Jr, "Sell-side information production in financial markets," *Journal of Financial and Quantitative Analysis*, 2012, *47* (4), 763–794.

Choi, Jay Pil, Doh-Shin Jeon, and Byung-Cheol Kim, "Privacy and personal data collection with information externalities," *Journal of Public Economics*, 2019, *173*, 113–124.

Dugast, Jérôme and Thierry Foucault, "Data abundance and asset price informativeness," *Journal of Financial economics*, 2018, *130* (2), 367–391.

_ **and** _ , "Equilibrium data mining and data abundance," *HEC Paris Research Paper No. FIN-2020-1393, Université Paris-Dauphine Research Paper*, 2023, (3710495).

Easley, David, Maureen O'Hara, and Liyan Yang, "Differential access to price information in financial markets," *Journal of Financial and Quantitative Analysis*, 2016, *51* (4), 1071–1110.

Eeckhout, Jan and Laura Veldkamp, "Data and market power," *NBER Working Paper*, 2022, (w30022).

EU, "Amendments to the Markets in Financial Instruments Regulation (MiFIR)," *data.europa.eu/eli/reg/2024/791/oj*, 2024.

Ganguli, Jayant Vivek and Liyan Yang, "Complementarities, multiplicity, and supply information," *Journal of the European Economic Association*, 2009, *7* (1), 90–115.

Garcia, Diego and Francesco Sangiorgi, "Information sales and strategic trading," *The Review of Financial Studies*, 2011, *24* (9), 3069–3104.

Goldstein, Itay, Chester S Spatt, and Mao Ye, "Big data in finance," *The Review of Financial Studies*, 2021, *34* (7), 3213–3225.

_ **,** _ **, and** _ , "The next chapter of big data in finance," *The Review of Financial Studies*, 2025, *38* (3), 605–622.

Huang, Shiyang, Yan Xiong, and Liyan Yang, "Skill acquisition and data sales," *Management Science*, 2022, *68* (8), 6116–6144.

Ichihashi, Shota, "Competing data intermediaries," *The RAND Journal of Economics*, 2021, *52* (3), 515–537.

_ , "The economics of data externalities," *Journal of Economic Theory*, 2021, *196*, 105316.

Kyle, Albert S, "Continuous auctions and insider trading," *Econometrica: Journal of the Econometric Society*, 1985, pp. 1315–1335.

Liu, Ernest, Song Ma, and Laura Veldkamp, "Data sales and data dilution," *Working paper*, 2024.

Madrigal, Vicente, "Non-fundamental speculation," *The Journal of Finance*, 1996, *51* (2), 553–578.

Manzano, Carolina and Xavier Vives, "Public and private learning from prices, strategic substitutability and complementarity, and equilibrium multiplicity," *Journal of Mathematical Economics*, 2011, *47* (3), 346–369.

O'Hara, Maureen, Yihui Wang, and Xing Alex Zhou, "The execution quality of corporate bonds," *Journal of Financial Economics*, 2018, *130* (2), 308–326.

Prüfer, Jens and Christoph Schottmüller, "Competing with big data," *The Journal of Industrial Economics*, 2021, *69* (4), 967–1008.

SEC, "Order Competition Rule," *www.sec.gov/rules/proposed/2022/34-96495.pdf*, 2022.

_ , "Disclosure of Order Execution Information," *www.sec.gov/files/rules/final/2024/34-99679.pdf*, 2024.

Veldkamp, Laura L, "Media frenzies in markets for financial information," *American Economic Review*, 2006, *96* (3), 577–601.

Wu, Xian, "Dynamic Market Choice," Technical Report, Working Paper 2024.

Yang, Liyan and Haoxiang Zhu, "Back-running: Seeking and hiding fundamental information in order flows," *The Review of Financial Studies*, 2020, *33* (4), 1484–1533.

Appendix

A Derivations and Proofs

Proof of Proposition 1 and Lemma 1

Proof. First, uninformed traders do not submit orders. Suppose an uninformed trader chooses a constant order x_u to maximize $E[x_u(\theta - p(y))]$. The expected profit is $E[x_u(\theta - \lambda y)] = x_u E[\theta] - \lambda x_u E[y]$. Given our prior $E[\theta] = 0$ and the fact that all signals and orders are linear in zero-mean variables, $E[y] = 0$. Thus, the expected profit for any constant order x_u is zero. $x_u = 0$ is therefore an optimal strategy, and we assume uninformed agents do not trade.

Given information structure (N_{fn}, N_f, N_n), informed trader i's problem is to maximize $E[x_i(\theta - p)|\mathcal{I}_i]$ by choosing x_i. Anticipating the linear pricing rule $p = \lambda y$ and $y = y_{-i} + x_i$, the first-order condition (FOC) requires

$$x_i = \frac{1}{2\lambda}\left(E[\theta|\mathcal{I}_i] - \lambda E[y|\mathcal{I}_i]\right). \tag{A.1}$$

And the second order condition is satisfied if $\lambda > 0$.

Case 1: $M > 0, K > 0$**.** We first solve the equilibrium when there exist both fundamental-informed traders and non-fundamental-informed traders. Here, we do not discuss whether some traders have both types of data, i.e., $N_{fn} > 0$ or $N_{fn} = 0$. We can show that the use of data is consistent regardless of the existing data for traders. Therefore, we only need to care about the total number of traders who have two types of data separately, i.e, $M \equiv N_f + N_{fn}$ and $K \equiv N_n + N_{fn}$.

F-type traders: For $\mathcal{I}_i = \{s_f\}$, $x_i = \beta_F \hat{\theta}_f$. The FOC is

$$2\lambda x_i = E[\theta|s_f] - \lambda E[y_{-i}|s_f],$$

where $E[\theta|s_f] = \hat{\theta}_f$, and $E[y_{-i}|s_f] = E[(N_f - 1)x^f + N_n x^n + N_{fn}x^{fn} + u|s_f]$. Due to the independence of s_f from s_n and u, $E[x^n|s_f] = E[x^{fn}|s_f] = \beta_{fn}\hat{\theta}_f$, and $E[u|s_f] = 0$.

$$E[y_{-i}|s_f] = (N_f - 1)\beta_f \hat{\theta}_f + N_{fn}\beta_{FN}\hat{\theta}_f.$$

Substituting it into the FOC and dividing by $\hat{\theta}_f$:

$$\lambda(N_f + 1)\beta_f + \lambda N_{fn}\beta_{fn} = 1 \tag{A.2}$$

N-type traders: For $\mathcal{I}_j = \{s_n\}$, $x_j = \gamma_n \hat{u}_n$. The FOC is

$$2\lambda x_j = E[\theta|s_n] - \lambda E[y_{-j}|s_n],$$

where $E[\theta|s_n] = E[\theta] = 0$, and $E[y_{-j}|s_n] = E[N_f x^f + (N_n - 1)x^n + N_{fn}x^{fn} + u|s_n]$. Due to independence, $E[x^f|s_n] = 0$, $E[x^n|s_n] = \gamma_n \hat{u}_n$, $E[x^{fn}|s_n] = \gamma_{fn}\hat{u}_n$, and $E[u|s_n] = \hat{u}_n$.

$$E[y_{-j}|s_n] = (N_n - 1)\gamma_n \hat{u}_n + N_{fn}\gamma_{fn}\hat{u}_n + \hat{u}_n.$$

Substituting it into the FOC and dividing by $\hat{u}_n$ and $\lambda > 0$:

$$(N_n + 1)\gamma_n + N_{fn}\gamma_{fn} = -1 \tag{A.3}$$

FN-type traders: For $\mathcal{I}_k = \{s_f, s_n\}$, $x_k = \beta_{fn}\hat{\theta}_f + \gamma_{fn}\hat{u}_n$. The FOC is

$$2\lambda x_k = E[\theta|s_f, s_n] - \lambda E[y_{-k}|s_f, s_n],$$

where $E[\theta|s_f, s_n] = \hat{\theta}_f$, and $E[y_{-k}|s_f, s_n] = E[N_f x^f + N_n x^n + (N_{fn} - 1)x^{fn} + u|s_f, s_n]$.

$$E[y_{-k}|s_f, s_n] = (N_f\beta_f + (N_{fn} - 1)\beta_{fn})\hat{\theta}_f + (N_n\gamma_n + (N_{fn} - 1)\gamma_{fn} + 1)\hat{u}_n$$

We substitute this into the FOC. By the orthogonality of $\hat{\theta}_f$ and $\hat{u}_n$, we can match coefficients for each term independently.

$$\lambda(N_{fn} + 1)\beta_{fn} + \lambda N_f\beta_f = 1 \tag{A.4}$$

$$(N_{fn} + 1)\gamma_{fn} + N_n\gamma_n = -1 \tag{A.5}$$

The equations (A.2)-(A.5) gives the optimal trading intensities:

$$\beta_{fn} = \beta_f = \frac{1}{\lambda(N_f + N_{fn} + 1)} \equiv \frac{1}{\lambda(M + 1)} \equiv \beta \tag{A.6}$$

$$\gamma_{fn} = \gamma_n = -\frac{1}{(N_n + N_{fn} + 1)} \equiv -\frac{1}{K + 1} \equiv \gamma \tag{A.7}$$

Then, we solve the pricing rule of the competitive market maker. Due to properties of normal distribution and linear pricing, $\lambda = \frac{Cov(\theta,y)}{Var(y)}$. The aggregate order flow can be re-written as

$$y = M\beta\hat{\theta}_f + K\gamma\hat{u}_n + u.$$

$$\lambda = \frac{\sqrt{M}}{M + 1}\sqrt{\frac{\tau_f}{\tau_\theta(\tau_f + \tau_\theta)}}\frac{1}{\sqrt{h(K)}} \tag{A.8}$$

where $h(K) \equiv Var(K\gamma\frac{\tau_n}{\tau_n+\tau_u}s_n + u) = \frac{\tau_n}{\tau_u(\tau_n+\tau_u)}(\frac{K}{K+1})^2 - 2\frac{\tau_n}{\tau_u(\tau_n+\tau_u)}\frac{K}{K+1} + \frac{1}{\tau_u}$.

Case 2: $M > 0, K = 0$**.** This is the standard version of Kyle (1985) with multiple informed traders. Actually, it is equivalent to the solution of Case 1 with $K = 0$.

Case 3: $M = 0, K > 0$**.** When there is no fundamental information in the financial market, N-type traders do not submit orders. As in case 1, in this case, N-type traders do not have any fundamental information. But the difference is that in case 1, N-type traders can earn rent from order uncertainty by submitting orders opposite to noise trades. This rent comes from the information disadvantage of the market maker. Due to the inability to distinguish whether orders come from noise traders or informed traders, the market maker always charge higher prices for more orders. N-type traders, on the other hand, know that prices are partially, mistakenly pushed up by noise trades. This generates information rent.

In case 3, however, the market maker no longer uses the upward pricing rule. The total order flow is $y = Kx^n + u$. The pricing constraint for the competitive market maker is $p = E[\theta|y]$. As $x^n = \gamma_n \hat{u}_n$ and u are both independent of θ, $p = E[\theta|y] = E[\theta] = 0$. It is intuitive. Now, the market makers know any order is uninformative. It is fair to always price as the prior mean of the asset's payoff, normalized as 0.

Given the price is 0 and $E[\theta|s_n] = 0$, the expected profit for N-type traders is always 0 for any order x^n. It is weakly dominant to submit $x^n = 0$. In this case, there is infinite equilibria with $p = 0$ and $x^n \in \mathbb{R}$. Taking any value of x^n does not affect the characteristics of equilibrium, therefore, for the convenience of discussion, we only consider the equilibrium with $x^n = 0$.

□

Proof of Proposition 2

Proof. We calculate the ex-ante expected profit for each type of traders. The ex-ante expected profit for j-type is

$$\pi_j = E[E[x^j(\theta - p)|\mathcal{I}_j]].$$

By the law of iterated expectations, $E[E[X|Y]] = E[X]$ for any random variables X, Y with finite mean. Applying this law, we have $E[E[\theta|s_f]] = E[E[u|s_n]]0$, $E[E[\theta^2|s_f]] = E[\theta^2] = \frac{1}{\tau_\theta}$, $E[E[u^2|s_n]] = E[u^2] = \frac{1}{\tau_u}$, $E[E[s_f\theta|s_f]] = \frac{1}{\tau_\theta}$, and $E[E[s_n u|s_n]] = \frac{1}{\tau_u}$. Substituting equations **??**, **??**, **??**, we get the expression for each type's ex-ante expected profit for the case $M > 0, K > 0$:

For a F-type trader,

$$\begin{aligned}
\pi_f &= E[E[\beta\hat{\theta}_f(\theta - \lambda(M\beta\hat{\theta}_f + K\gamma\hat{u}_n + u))|s_f]] \\
&= E[\beta E[\hat{\theta}_f\theta|s_f] - \lambda\beta^2 M E[\hat{\theta}_f^2|s_f]] \\
&= \frac{1}{\lambda(M+1)}\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)} - \frac{M}{\lambda(M+1)^2}\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)} \\
&= \frac{1}{\lambda(M+1)^2}\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}
\end{aligned} \tag{A.9}$$

For a N-type trader,

$$\begin{aligned}
\pi_n &= E[E[\gamma\hat{u}_n(\theta - \lambda(M\beta\hat{\theta}_f + K\gamma\hat{u}_n + u))|s_n]] \\
&= E[-\lambda\gamma(E[\gamma K\hat{u}_n^2|s_n] + E[\hat{u}_n u|s_n])] \\
&= \frac{\lambda}{(K+1)^2}\frac{\tau_n}{\tau_u(\tau_n+\tau_u)}.
\end{aligned} \tag{A.10}$$

For a FN-type trader,

$$\begin{aligned}
\pi_{fn} &= E[E[(\beta\hat{\theta}_f + \gamma\hat{u}_n)(\theta - \lambda(M\beta\hat{\theta}_f + K\gamma\hat{u}_n + u))|s_f, s_n]] \\
&= \frac{1}{\lambda(M+1)^2}\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)} + \frac{\lambda}{(K+1)^2}\frac{\tau_n}{\tau_u(\tau_n+\tau_u)}.
\end{aligned} \tag{A.11}$$

Therefore, we have $v_f(M,K) = \pi_f = \pi_{fn} - \pi_n$, $v_n(M,K) = \pi_n = \pi_{fn} - \pi_f$ when $M > 0, K > 0$. It implies the additivity.

As we discussed before, the N-type traders do not submit orders when $M = 0$. So, $v_n = 0$ when $M = 0, K > 0$. By the definition, the value of data is calculated by the difference between existing traders' ex-ante profit. So, the value of one data does not exist when there is no trader has this data, i.e, $M = 0$ or $K = 0$. □

Proof of Lemma 3

Proof. **Market depth:** The market depth (MD) is defined as the inverse of price impact:

$$MD = \frac{1}{\lambda}.$$

We analyze the impact of the number of fundamental (M) and non-fundamental traders (K) on the market depth.

$$\frac{\partial MD}{\partial M} \propto \frac{\partial (M+1)/\sqrt{M}}{\partial M} = \frac{M-1}{2M^{3/2}}.$$

This derivative is zero at $M = 1$ and strictly positive for all $M > 1$. We focus on the environment where exists at least one fundamental-informed trader, $M > 1$. So, we say

$$\frac{\partial \mathcal{D}(M,K)}{\partial M} > 0 \quad (\text{for } M > 1)$$

. For the number of non-fundamental traders K,

$$\frac{\partial MD}{\partial K} \propto \frac{\partial h(K)}{\partial K} < 0.$$

Informational efficiency: By definition, informational efficiency is $\frac{1}{Var(\theta|p)}$. Since $p = \lambda y$ and λ is a known equilibrium constant, observing p is informationally equivalent to observing y. Thus, it is equivalent to $(\mathrm{Var}(\theta|y))^{-1}$. We use the formula for the posterior variance of a Gaussian variable:

$$\mathrm{Var}(\theta|y) = \mathrm{Var}(\theta) - \frac{\mathrm{Cov}(\theta, y)^2}{\mathrm{Var}(y)}$$

From the proof of Proposition 1, we derive $\lambda = \mathrm{Cov}(\theta, y)/\mathrm{Var}(y)$, which implies $\mathrm{Var}(y) = \mathrm{Cov}(\theta, y)/\lambda$. Substituting this into the variance formula:

$$\mathrm{Var}(\theta|y) = \mathrm{Var}(\theta) - \frac{\mathrm{Cov}(\theta, y)^2}{\mathrm{Cov}(\theta, y)/\lambda} = \mathrm{Var}(\theta) - \lambda \cdot \mathrm{Cov}(\theta, y)$$

We also know from the proof of Proposition 1 that $\mathrm{Cov}(\theta, y) = M\beta \frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}$.

$$\mathrm{Var}(\theta|y) = \tau_\theta^{-1} - \lambda(M\beta V_f)$$

Now, substitute the equilibrium trading intensity $\beta = \frac{1}{\lambda(M+1)}$:

$$\mathrm{Var}(\theta|y) = \tau_\theta^{-1} - \lambda M \left(\frac{1}{\lambda(M+1)}\right) \frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)} = \tau_\theta^{-1} - \frac{M}{M+1} \frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}$$

The posterior variance $\mathrm{Var}(\theta|p)$ depends only on M, not on K. Taking the inverse gives the

expression for informational efficiency.

It is easy to see that the informational efficiency strictly increases in the number of fundamental-informed traders M:

$$\begin{aligned}\frac{\partial IE}{\partial M} &\propto -\frac{\partial \text{Var}(\theta|y)}{\partial M} \\ &\propto -\frac{1}{(1+M)^2} < 0.\end{aligned}$$

And the informational efficiency does not depend on the number of non-fundamental traders K.

Order flow volatility: The total (net) order flow y is expressed as:

$$y = M\beta\hat{\theta}_f + K\gamma\hat{u}_n + u.$$

By the orthogonality of the components (as $\hat{\theta}_f$ is independent of $\hat{u}_n$ and u):

$$\text{Var}(y) = \text{Var}(M\beta\hat{\theta}_f) + \text{Var}(K\gamma\hat{u}_n + u)$$

The first term is $\text{Var}(M\beta\hat{\theta}_f) = M^2\beta^2\text{Var}(\hat{\theta}_f) = M^2\beta^2\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}$. The second term is $\text{Var}(K\gamma\hat{u}_n+u)$, which we defined in Proposition 1 as $h(K)$. Now we substitute the equilibrium trading intensity $\beta = \frac{1}{\lambda(M+1)}$ and the expression for λ^2:

$$\begin{aligned}Var(y) &= Var\left(M\beta\hat{\theta}_f + K\gamma\hat{u}_n + u\right) \\ &= \left(\frac{M}{\lambda(M+1)}\right)^2 \frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)} + Var(K\gamma\hat{u}_n + u) \\ &= Mh(K) + h(K) \\ &= (M+1)h(K).\end{aligned} \tag{A.12}$$

Therefore, it is trivial to see

$$\frac{\partial Var(y)}{\partial M} > 0, \quad \frac{\partial Var(y)}{\partial K} < 0.$$

Noise trader welfare: Noise trader welfare, denoted by W_{noise}, is the ex-ante expected profit for noise traders:

$$W_{noise} = E[u(\theta - p)] = E[u\theta] - \lambda E[uy].$$

By assumption, u and θ are independent, so $E[u\theta] = E[u]E[\theta] = 0$. We substitute the full expression for $y = M\beta\hat{\theta}_f + K\gamma\hat{u}_n + u$, and use the linearity of expectations:

$$\begin{aligned}W_{noise} &= -\lambda E[uy] \\ &= -\lambda E[u\cdot(M\beta\hat{\theta}_f + K\gamma\hat{u}_n + u)] \\ &= -\lambda M\beta E[u\hat{\theta}_f] + \lambda K\gamma E[u\hat{u}_n] + \lambda E[u^2] \\ &= -\lambda\left(\frac{1}{\tau_u} - \frac{K}{K+1}\frac{\tau_n}{\tau_u(\tau_n+\tau_u)}\right) < 0.\end{aligned} \tag{A.13}$$

For impact of the number of fundamental traders M,

$$\frac{\partial W_{noise}}{\partial M} \propto -\frac{\partial\lambda}{\partial M} > 0.$$

For impact of the number of non-fundamental traders K, some algebra gives

$$\begin{aligned}\frac{\partial W_{noise}}{\partial K} &= -\frac{\partial \lambda}{\partial K}\left(\frac{1}{\tau_u} - \frac{K}{K+1}\frac{\tau_n}{\tau_u(\tau_n+\tau_u)}\right) + \frac{\lambda\tau_n}{\tau_u(\tau_n+\tau_u)}\frac{1}{(K+1)^2} \\ &= \frac{K\sqrt{M}\,\tau_n\sqrt{\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}}}{(1+K)^3(1+M)\,\tau_u\,(\tau_n+\tau_u)^2\left(\frac{\tau_n+(1+K)^2\tau_u}{(1+K)^2\tau_u(\tau_n+\tau_u)}\right)^{3/2}} > 0.\end{aligned}$$

For convenience, we summarize the expression of these four metrics as follows:

$$\begin{aligned} MD &= 1/\lambda \\ IE &= \left(\frac{1}{\tau_\theta} - \frac{M}{M+1}\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}\right)^{-1} \\ Var(y) &= (M+1)h(K) \\ W_{noise} &= -\lambda\left(\frac{1}{\tau_u} - \frac{K}{K+1}\frac{\tau_n}{\tau_u(\tau_n+\tau_u)}\right). \end{aligned} \tag{A.14}$$

□

Proof of Lemma 4

Proof. The profit of the monopolistic data seller is

$$\Pi = M\,v_f + K\,v_n.$$

The first order derivative for M:

$$\frac{\partial \Pi}{\partial M} = \frac{(1-M)(1+K)\sqrt{\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}}(\tau_n+\tau_u+K\tau_u)\sqrt{\frac{\tau_n+(1+K)^2\tau_u}{(1+K)^2\tau_u(\tau_n+\tau_u)}}}{2\sqrt{M}(1+M)^2(\tau_n+(1+K)^2\tau_u)}.$$

The sign of the FOC for M only depends on $(1-M)$. The profit increases in M when $0 < M \leq 1$ and decreases in M when $M > 1$. It achieves the maximum at $M = 1$. The optimal choice of M is

$$M^I = 1.$$

The first order derivative for K:

$$\frac{\partial \Pi}{\partial K} = -\frac{K\sqrt{M}\,\tau_n\sqrt{\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}}}{(1+K)^3(1+M)\,\tau_u\,(\tau_n+\tau_u)^2\left(\frac{\tau_n+(1+K)^2\tau_u}{(1+K)^2\tau_u(\tau_n+\tau_u)}\right)^{3/2}} \leq 0.$$

The profit always decreases in $K \geq 0$. It is optimal not to sell non-fundamental data, i.e,

$$K^I = 0.$$

□

Proof of Lemma 5

Proof. The profit of the data seller in the fundamental data market is

$$\Pi_f = M\, v_f.$$

The first order derivative relative to M is

$$\frac{\partial \Pi_f}{\partial M} = \frac{(1-M)\sqrt{\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}}\sqrt{\frac{\tau_n+(1+K)^2\tau_u}{(1+K)^2\tau_u(\tau_n+\tau_u)}}}{2\sqrt{M}(1+M)^2}.$$

The optimal decision is the same as the monopoly problem:

$$M^{II} = 1 = M^I.$$

The profit of the data seller in the non-fundamental data market is

$$\Pi_n = K\, v_n.$$

The first order derivative relative to K is

$$\frac{\partial \Pi_n}{\partial K} = \frac{\sqrt{M}\,\tau_n\sqrt{\frac{\tau_f}{\tau_\theta(\tau_f+\tau_\theta)}}\sqrt{\frac{\tau_n+(1+K)^2\tau_u}{(1+K)^2\tau_u(\tau_n+\tau_u)}}\,(\tau_n-(1+K)^2(K-1)\tau_u)}{(1+K)(1+M)(\tau_n+(1+K)^2\tau_u)^2}$$

We define $g(K) = \tau_n - \tau_u(1+K)^2(K-1)$. The discriminant of $g(K)=0$ is

$$\Delta_g \propto -27(1+\frac{\tau_n}{\tau_u})^2 + 22(1+\frac{\tau_n}{\tau_u}) + 5.$$

It is easy to find that $\Delta_g < 0$ for $\tau_n > 0, \tau_u > 0$. As the discriminant of $g(x)=0$ is negative, it only has one real root, which is

$$K^{II} = \frac{1}{3}\left(\frac{4 * 2^{1/3}\,\tau_u}{\sqrt[3]{27\,\tau_n\,\tau_u^2 + 16\,\tau_u^3 - 3\sqrt{3}\,\sqrt{27\,\tau_n^2\,\tau_u^4 + 32\,\tau_n\,\tau_u^5}}} + \frac{1}{2^{1/3}\,\tau_u}\sqrt[3]{27\,\tau_n\,\tau_u^2 + 16\,\tau_u^3 - 3\sqrt{3}\,\sqrt{27\,\tau_n^2\,\tau_u^4 + 32\,\tau_n\,\tau_u^5}} - 1\right) \geq 1.$$

Denote $A \equiv \frac{1}{2^{1/3}}\sqrt[3]{27\rho_n + 16 - 3\sqrt{3}\,\sqrt{27\,\rho_n^2 + 32\,\rho_n}}$ and $\rho_n = \frac{\tau_n}{\tau_u} > 0$.

$$\begin{aligned} 27\rho_n + 16 - 3\sqrt{3}\sqrt{27\rho_n^2 + 32\rho_n} &> 27\rho_n + 16 - 3\sqrt{3}\sqrt{27\rho_n^2 + 32\rho_n + \frac{16}{3\sqrt{3}}} \\ &= 27\rho_n + 16 - 3\sqrt{3}\left(3\sqrt{3}\rho_n + \frac{16}{3\sqrt{3}}\right) = 0. \end{aligned}$$

As $A > 0$, $K^{II} = \frac{1}{3}(4/A + A - 1) \geq \frac{1}{3}(4-1) \geq 1$.

□

Proof of Lemma 6

Proof. There are $J_f \geq 2$ competitive sellers in the fundamental data market. The number of potential rational traders is bounded by N. The inverse demand function is $p_f(M) = v_f(M)$ when $M \leq N$ and $p_f(M) = 0$ when $M > N$. We show that there are multiple equilibria with the same total supply $M^{III} = N$.

For any seller j, the problem is to maximize

$$\Pi_{f,j}(m_j, m_{-j}) = m_j \, p_f(m_j + m_{-j})$$

s.t. $m_j \geq 0$ and $m_j + m_{-j} \leq N$. The first order derivative relative to m_j is

$$\frac{\partial \Pi_{f,j}(m_j, m_{-j})}{\partial m_j} \propto -m_j^2 + m_j(1 + m_{-j}) + 2\, m_{-j}(1 + m_{-j}).$$

We consider the symmetric equilibrium, so $m_{-j} = (J_f - 1)m_j$. Substituting it into the first order derivative, it is

$$(2J_f - 3)J_f \, m_j + 2J_f - 1 > 0.$$

The complementary slackness requires

$$\frac{\partial \Pi_{f,j}(m_j, m_{-j})}{\partial m_j} \left(N - \sum_j m_j \right) = 0.$$

Therefore, in equilibrium, $\sum_j m_j = N$ for any $J_f > \frac{3}{2}$. It implies that, as long as there are more than 2 sellers in the fundamental data market, they would sell the data to all potential traders:

$$M^{III} = N.$$

There are $J_n \geq 2$ competitive sellers in the non-fundamental data market. The inverse demand function is $p_n(K) = v_n(K)$ when $K \leq N$ and $p_f(M) = 0$ when $K > N$.

For any seller j, the problem is to maximize

$$\Pi_{n,j}(k_j, k_{-j}) = k_j \, p_n(k_j + k_{-j})$$

s.t. $k_j \geq 0$ and $k_j + k_{-j} \leq N$. The first order derivative relative to k_j is

$$\frac{\partial \Pi_{n,j}(k_j, k_{-j})}{\partial k_j} \propto (1 + k_{-j})\, \rho_n - (-1 + k_j - k_{-j})\, (1 + k_j + k_{-j})^2.$$

We consider the symmetric equilibrium, so $k_{-j} = (J_n - 1)k_j$. Substituting it into the first order derivative, it is

$$(J_n - 2)J_n^2 k_j^3 + (3J_n - 4)J_n k_j^2 + (4J_n - 3)k_j + 1 + \rho_n > 0.$$

The complementary slackness requires

$$\frac{\partial \Pi_{n,j}(k_j, k_{-j})}{\partial k_j}\left(N - \sum_j k_j\right) = 0.$$

Therefore, in equilibrium, $\sum_j k_j = N$ for any $J_n \geq 2$. It implies that, as long as there are more than 2 sellers in the non-fundamental data market, they would sell the data to all potential traders:

$$K^{III} = N.$$

□

Proof of Proposition 3

Proof. The market depth MD under three different regimes of data sales:

$$\begin{aligned}
MD^I &= MD(M^I, K^I) = MD(1, 0) = 2\sqrt{\frac{\tau_\theta(\tau_f + \tau_\theta)}{\tau_u \tau_f}} \\
MD^{II} &= MD(M^{II}, K^{II}) = MD(1, K^{II}) \\
MD^{III} &= MD(M^{III}, K^{III}) = MD(N, N) = \frac{\sqrt{(\tau_n + (1+N)^2\tau_u)\,\tau_\theta(\tau_f + \tau_\theta)}}{\sqrt{N\,\tau_u(\tau_n + \tau_u)\,\tau_f}}
\end{aligned}$$

As we know $\partial MD/\partial K < 0$ from Lemma 3, it is obvious that $MD^I > MD^{II}$ as $K^{II} \geq 1$.

To compare MD^I and MD^{III}, we examine the difference $MD^I - MD^{III}$:

$$MD^I - MD^{III} \propto 2 - \sqrt{\frac{\tau_n + (1+N)^2\tau_u}{N(\tau_n + \tau_u)}}$$

$RHS \geq 0$ if and only if

$$\begin{aligned}
&\tau_n + (1+N)^2\tau_u \leq 4(\tau_n + \tau_u)N \\
\iff &1 + 2\rho_n - \sqrt{4\,\rho_n^2 + 3\,\rho_n} \leq N \leq 1 + 2\rho_n + \sqrt{4\,\rho_n^2 + 3\,\rho_n}
\end{aligned}$$

where $\rho_n \equiv \frac{\tau_n}{\tau_u}$. We can find that $1 + 2\rho_n - \sqrt{4\,\rho_n^2 + 3\,\rho_n} \leq 1$. As in our setting, $N \geq 1$, $MD^I \geq MD^{III}$ if and only $1 \leq N \leq \hat{N}$

$$\hat{N} \equiv 1 + 2\rho_n + \sqrt{4\,\rho_n^2 + 3\,\rho_n}.$$

For the informational efficiency IE, we know

$$IE^I = IE^{II} \leq IE^{III}$$

as IE is increasing in M and does not depend on K.

The order flow volatility $Var(y)$ under three different regimes of data sales:

$$Var^I(y) = \frac{2}{\tau_u}$$
$$Var^{II}(y) = Var(y)|_{M=1,K=K^{II}}$$
$$Var^{III}(y) = \frac{\tau_n + \tau_u(1+N)^2}{\tau_u(\tau_n+\tau_u)(1+N)}$$

For the order flow volatility $Var(y)$, we know

$$Var^I(y) > Var^{II}(y)$$

as $Var(y)$ is decreasing in K.

To compare $Var^{II}(y)$ and $Var^{III}(y)$, we compare their marginal change starting from $N = 1$. At $N = 1$, $Var^{II}(y) = Var^{III}(y)$ as $M^{II} = M^{III} = 1$ and $K^{II} = K^{III} = 1$.

$Var^{II}(y)$ decreases in N when $1 \leq N \leq K^{II}$ and remain constant when $N > K^{II}$. So, $Var^{II}(y) \leq Var^{II}(y)|_{M=1,K=1} = Var^{III}(y)|_{M=1,K=1}$.

Meanwhile, if $\rho_n > 4$, $Var^{III}(y)$ decreases in N when $1 \leq N \leq \sqrt{\rho_n} - 1$ and increases in N when $N > \sqrt{\rho_n}$, otherwise, $Var^{III}(y)$ always increases in N:

$$\frac{\partial Var^{III}(y)}{\partial N} \propto \tau_u(1+N)^2 - \tau_n.$$

We find that the decrease ratio of $Var^{III}(y)$ is slower than $Var^{II}(y)$ when $N \leq K^{II}$ as the former has a larger M to provide a positive growth effect. Therefore, $Var^{III}(y) \geq Var^{II}(y)$.

To compare $Var^I(y)$ and $Var^{III}(y)$, we examine the difference $Var^I(y) - Var^{III}(y)$:

$$Var^I(y) - Var^{III}(y) = \frac{(1+2N)\tau_u + (1-N^2)\tau_u}{(1+N)\,\tau_u\,(\tau_n+\tau_u)}.$$

$RHS \geq 0$ if and only if

$$\rho_n - \sqrt{\rho_n^2 + \rho_n + 1} \leq N \leq \rho_n + \sqrt{\rho_n^2 + \rho_n + 1}.$$

$\rho_n - \sqrt{\rho_n^2 + \rho_n + 1} < 0$ and in our setting $N \geq 1$. Therefore, $N \geq 1$, $Var^I(y) \geq Var^{III}(y)$ if and only $1 \leq N \leq \overline{N}$

$$\overline{N} \equiv \rho_n + \sqrt{\rho_n^2 + \rho_n + 1}.$$

For the noise trader welfare, we know W_{noise} increases in both M and K. As $M^I = M^{II} \leq M^{III}$ and $K^I < K^{II} \leq K^{III}$,

$$W^I_{noise} < W^{II}_{noise} \leq W^{III}_{noise}.$$

□

Proof of Lemma 7

Proof. Under the integrated competition regime, a seller $j \in \{1, \ldots, J\}$'s problem is

$$\max_{0 \leq m_j, k_j \leq N} \Pi_j(m_j, k_j) = m_j v_f(m_j + m_{-j}, k_j + k_{-j}) + k_j v_n(m_j + m_{-j}, k_j + k_{-j}).$$

We solve the symmetric equilibrium. For the optimal sales of non-fundamental data, the first order derivative of Π_j relative to k_j is

$$\begin{aligned}\frac{\partial \Pi_j}{k_j} =& - k_j(1+k_j)^2 m_j + k_{-j}^3(m_j + m_{-j}) + k_{-j}^2\Big((2+k_j)m_j + (3+k_j)m_{-j}\Big) \\ &+ m_{-j}\Big(\rho_n - (-1+k_j)(1+k_j)^2\Big) + k_{-j}\Big(m_j(\rho_n + 1 - k_j^2) + m_{-j}(\rho_n + 3 + 2k_j - k_j^2)\Big)\end{aligned}$$

For the symmetric equilibrium, each seller plays the same strategy, so $k_{-j} = (J-1)k_j$. Substituting it into the above equation and we get

$$\begin{aligned}\frac{\partial \Pi_j}{k_j} =& (J-2)J^2 k_j^3(m_j + m_{-j}) + J k_j^2\Big(2(J-2)m_j + (3J-4)m_{-j}\Big) \\ &+ m_{-j}(\rho_n + 1) + k_j\Big((J-1)m_j\rho_n + (J-1)m_{-j}\rho_n + (J-2)m_j + (3J-2)m_{-j}\Big) \geq 0.\end{aligned}$$

It is always positive and the sellers sell non-fundamental data to N traders.

The first order derivative of π_j relative to M_j is

$$\begin{aligned}\frac{\partial \Pi_j}{m_j} \propto\ & m_j\Big(\Big(1 + k_j + (-1+J)m_j - 2(-1+J)k_j m_j\Big)\rho_n \\ &+ (1 + k_j + k_{-j})^2\Big(1 + (-1+J)m_j\Big) - m_j\Big((1+k_j)\rho_n + (1+k_j+k_{-j})^2\Big) \\ &+ (-1+J)\Big(\Big(2 + 2(-1+J)m_j + k_j\Big(1 + m_j - Jm_j\Big)\Big)\rho_n \\ &+ 2(1 + k_j + k_{-j})^2\Big(1 + (-1+J)m_j\Big)\Big)\end{aligned}$$

Given $K = N$ and $k_j = \frac{N}{J}$, RHS is

$$m_j\left(-m_j J\left(\rho_n(N - 2J + 3) - (2J-3)(1+N)^2\right) + (2J + N - 1)\rho_n + (2J-1)(N+1)^2\right)$$

If $\rho_n(N - 2J + 3) - (2J-3)(1+N)^2 \leq 0$, the marginal effect of selling fundamental data is always positive, so they sell it until the upper bound N.

If $\rho_n(N - 2J + 3) - (2J-3)(1+N)^2 > 0$, then there exists an interior optimum

$$M_j^* = \frac{(2J-1)(1+N)^2 + \rho_n(N + 2J - 1)}{J\left(\rho_n(N - 2J + 3) - (2J-3)(1+N)^2\right)}.$$

If $J \geq \frac{N+3}{2}$, the polynomial $\rho_n(N - 2J + 3) - (2J-3)(1+N)^2$ is always negative, so the first order derivative is positive. If $J < \frac{N+3}{2}$ and $\rho_n > \frac{(2J-3)(1+N)^2}{N-2J+3} \equiv \tilde{\rho}_n$, the first order condition is satisfied when $M_j = M_j^*$. Therefore, the aggregate sales of fundamental data is $\hat{M}^{IV} = J \cdot M_j^*$:

$$\hat{M}^{IV} = \frac{(2J-1)(1+N)^2 + \rho_n(N + 2J - 1)}{\rho_n(N - 2J + 3) - (2J-3)(1+N)^2}$$

As the financial market size is limited as N, the optimal sales of fundamental data is

$$M^{IV} = \min\{\hat{M}^{IV}, N\}.$$

We can find that $\hat{M}^{IV}$ is continuous in $\rho_n > \frac{(2J-3)(1+N)^2}{N-2J+3} \equiv \tilde{\rho}_n$ and $J < \frac{N+3}{2}$, and

$$\frac{\partial \hat{M}^{IV}}{\partial J} > 0, \quad \frac{\partial \hat{M}^{IV}}{\partial \rho_n} < 0.$$

The first inequality implies that stronger competition effects can improve the sale of fundamental data, while the second inequality implies that stronger cross-market externalities can limit the sale of fundamental data.

It is easy to see that $\hat{M}^{IV} \to \infty$ as $\rho_n \to \tilde{\rho}_n$. As for the lower bound of $\hat{M}^{IV}$, we find

$$\lim_{J\to 2} \lim_{\rho\to\infty} \hat{M}^{IV} = \frac{N+3}{N-1} \geq 1.$$

Therefore, $\hat{M}^{IV}$ may take any value of $[0, \infty)$. It depends on the parameters ρ_n, J, and N.

□

Proof of Lemma 8

Proof. Denote the vendor for both data by vendor 1 and another vendor for fundamental data by vendor 2. The aggregate data sales are $M = M_1 + M_2$ and $K = K_1 + K_2$. The profit of vendor 1 is $\Pi_1 = M_1 v_f(M_1 + M_2, K_1) + K_1 v_n(M_1 + M_2, K_1)$, and the profit of vendor 2 is $\Pi_2 = M_2 v_f(M_1 + M_2, K_1)$.

sell $j \in \{1, 2\}$ is $\Pi_j = M_j v_f(M_j + M_{-j}, K_j + K_{-j}) + K_j v_n(M_j + M_{-j}, K_j + K_{-j})$.

The first-order derivative of π_2 relative to M_2:

$$\frac{\partial \pi_2}{\partial M_2} \propto -M_2^2 + (1 + M_1)M_2 + 2M_1(1 + M_2).$$

The first order condition has two real roots:

$$\frac{1}{2}\left(1 + M_1 - \sqrt{9M_1^2 + 10M_1 + 1}\right) \quad \text{and} \quad \frac{1}{2}\left(1 + M_1 + \sqrt{9M_1^2 + 10M_1 + 1}\right).$$

$\frac{1}{2}\left(1 + M_1 - \sqrt{9M_1^2 + 10M_1 + 1}\right) < 0$ for any $M_1 \geq 0$. So, the optimal decision of data vendor 2 is to set

$$M_2^* = \frac{1}{2}\left(1 + M_1 + \sqrt{9M_1^2 + 10M_1 + 1}\right) \geq 1.$$

The first order derivative of vendor 1's profit Π_1 relative to K_1 is

$$\frac{\partial \Pi_1}{\partial K_1} \propto -(M1 + M2)K_1^3 - (2M_1 + M_2)K_1^2 - (M_1 - M_2)K_1 + (1 + \rho_n)M_2$$

The discriminant of $RHS = 0$ is

$$-M_2\rho_n\Big(4M_1^3 + 3M_1^2M_2\Big(9\rho_n + 8\Big) + 6M_1M_2^2\Big(9\rho_n + 8\Big) + M_2^3\Big(27\rho_n + 32\Big)\Big) < 0.$$

Therefore, the first order condition for K_1 has one real root, and the second-order condition is always satisfied given positive parameters. The optimal K_1 is

$$K_1 = \hat{K}^V \equiv \frac{1}{3(M_1+M_2)}\left(B + \frac{(M_1+2M_2)^2}{B} - 2M_1 - M_2\right)$$

where

$$B \equiv \frac{1}{2^{1/3}}\Bigg(27M_2^3\rho_n + 2M_1^3 + 16M_2^3 + 3M_1^2M_2(9\rho_n+4) + 6M_1M_2^2(9\rho_n+4)$$
$$- 3\sqrt{3}\sqrt{M_2(M_1+M_2)^2\rho_n\left(4M_1^3 + 3M_1^2M_2(9\rho_n+8) + 6M_1M_2^2(9\rho_n+8) + M_2^3(27\rho_n+32)\right)}\Bigg)^{\frac{1}{3}}.$$

We have $\hat{K}^V \geq \frac{1}{3(M_1+M_2)}\left(2(M_1+2M_2) - 2M_1 - M_2\right) = \frac{M_2}{M_1+M_2} > 0$.

Substituting the optimal sales of vendor 2 into the first order derivative of vendor 1's profit relative to M_1, the first order condition requires that

$$M_1^* = \frac{(K_1+3)\rho_n^2 + (K_1+1)^2(K_1+6)\rho_n + 3(K_1+1)^4}{-(K_1^2+K_1-2)\rho_n^2 - (K_1-4)(K_1+1)^2\rho_n + 2(1+K_1)^2}.$$

If $K_1 > 1$ and $\rho_n > \frac{(1+K_1)^2}{K_1-1}$, then $M_1^* < 0$, the vendor 1 optimally does not sell fundamental data. It implies that $M_1 = 0$ and $M_2 = 1$. Then, we find that $B = A$ and $K_1 = \hat{K}^{II}$. $\hat{K}^{II}$ satisfies $\rho_n = (1+\hat{K}^{II})^2(\hat{K}^{II}-1)$. Substituting the constraint into the second inequality, it holds that $\hat{K}^{II} > 2$. We know that $\hat{K}^{II}$ is increasing in ρ_n and not bounded. There exists $\hat{\rho}_n$ such that $\hat{K}^{II}(\hat{\rho}_n) = 2$. Therefore, if $\rho_n > \hat{\rho}_n$, there exists an equilibrium, in which

$$M_1 = 0, K_1 = \min\{\hat{K}^{II}, N\}$$
$$M_2 = 1, K_2 = 0.$$

It is a segmented monopoly equilibrium, the same as the equilibrium in Regime II.

If $K_1 < 1$, or $K_1 > 1$ and $0 < \rho_n \leq \frac{(1+K_1)^2}{K_1-1}$, then $M_1^* \geq 0$. The aggregate data sales are $\hat{M}^V = M_1^* + M_2^*$ and $\hat{K}^V = K_1^*$. There exists an equilibrium, in which the endogenous data allocation is

$$M^V = \min\{\hat{M}^V, N\},\ K^V = \min\{\hat{K}^V, N\}.$$

Recall the interior solution $\hat{K}^V$ satisfies

$$h(\hat{K}^V) \equiv (M1+M2)(\hat{K}^V)^3 + (2M_1+M_2)(\hat{K}^V)^2 + (M_1-M_2)\hat{K}^V - (1+\rho_n)M_2 = 0.$$

By the implicit function theorem, the derivative $\frac{\partial \hat{K}^V}{\partial M_1}$ is proportion to $\frac{\partial h(\hat{K}^V)}{\partial M_1}$. Substituting $M_2 = M_2^*$,

$$\frac{\partial h(\hat{K}^V)}{\partial M_1} \propto \frac{1}{2}\left(\rho_n - (\hat{K}^V+1)^2(\hat{K}^V-1)\right) - \hat{K}^V(\hat{K}^V+1)^2$$

If $\hat{K}^V < 1$, we have $\hat{K}^{II} \geq 1$, so $\hat{K}^{II} > \hat{K}^V$.

If $K_1 > 1$ and $0 < \rho_n \leq \frac{(1+\hat{K}^V)^2}{\hat{K}^V - 1}$, RHS satisfies

$$RHS \leq (1+\hat{K}^V)^2 \left(\frac{(\hat{K}^V - 1)^2 - 1}{\hat{K}^V - 1} - \hat{K}^V \right) = -\frac{(1+\hat{K}^V)^2}{\hat{K}^V - 1} < 0.$$

So, the interior optimum $\hat{K}^V$ is decreasing in M_1. And we know that $\hat{K}^{II}$ is a special case of $\hat{K}^V$ with $M_1 = 0$. Therefore, it always holds that

$$\hat{K}^V \leq \hat{K}^{II}.$$

It implies that the generalist sells a smaller quantity of non-fundamental data than the segmented monopoly equilibrium. It is due to that the generalist internalizes the cross-market externality. □

Proof of Lemma 9

Proof. Denote the vendor for both data by vendor 1 and another vendor for non-fundamental data by vendor 2. The aggregate data sales are $M = M_1 + M_2$ and $K = K_1 + K_2$. The profit of vendor 1 is $\Pi_1 = M_1 v_f(M_1, K_1 + K_2) + K_1 v_n(M_1, K_1 + K_2)$, and the profit of vendor 2 is $\Pi_2 = K_2 v_n(M_1, K_1 + K_2)$.

The vendor 1 always sells fundamental data to one trader:

$$M^{VI} = M_1^* = 1.$$

For non-fundamental data, the first order derivative $\frac{\partial \Pi_1}{\partial K_1}$ is

$$\frac{\partial \Pi_1}{\partial K_1} \propto -K_1^3 - (K_2+2)K_1^2 + (K_2-1)^2 K_1 + K_2(\rho_n + 1) + 2K_2^2 + K_2^3.$$

The second order condition requires that $K_2 \leq 3K_1 + 1$.

For vendor 2, the first order derivative $\frac{\partial \Pi_2}{\partial K_2}$ is

$$\frac{\partial \Pi_2}{\partial K_2} \propto -K_2^3 - (K_1+1)K_2^2 + (K_1+1)^2 K_2 + K_1^3 + 3K_1^2 + (3+\rho_n)K_1 + \rho_n + 1.$$

The second order condition requires that $K_1 \leq 3K_2 - 1$.

The first order condition of vendor 1 for choosing K_1 gives that

$$K_1^3 = -(K_2+2)K_1^2 + (K_2-1)^2 K_1 + K_2(\rho_n + 1) + 2K_2^2 + K_2^3.$$

Substituting this equation into the term K_1^3 of $\frac{\partial \Pi_2}{\partial K_2}$:

$$\frac{\partial \Pi_1}{\partial K_1} \propto K_1^2 + K_1(4+\rho_n) + (1+K_2)(1+K_2+\rho_n) > 0.$$

It implies that if the first order condition for K_1 holds, the marginal revenue of vendor 2 for selling

more non-fundamental data is always positive. Therefore, he would sell as much as possible, i.e.,

$$K_2^* = N - K_1^*.$$

Given the optimal sale K_2^*, the first order condition for K_1 becomes

$$2K_1^2 - K_1(6N + 2N^2 + \rho_n) + n(1 + 2N + N^2 + \rho_n) = 0.$$

It has two real roots:

$$\hat{K}_1^* = \frac{1}{4}\left(2N^2 + 6N + \rho_n - \sqrt{(2N^2 + 6N + \rho_n)^2 - 8N(N^2 + 2N + 1 + \rho_n)}\right), \text{and}$$
$$\hat{K}_1^{**} = \frac{1}{4}\left(2N^2 + 6N + \rho_n + \sqrt{(2N^2 + 6N + \rho_n)^2 - 8N(N^2 + 2N + 1 + \rho_n)}\right).$$

It is easy to find that the second root $\hat{K}_1^{**}$ is larger than N:

$$\hat{K}_1^{**} \geq \frac{1}{4}\left(2N^2 + 6N + \rho_n\right) > \frac{3}{2}N > N.$$

So, $\frac{\partial \Pi_1}{\partial K_1}$ increases in $K_1 \in [0, \hat{K}_1^*]$ and decreases in $K_1 \in [\hat{K}_1^*, \infty)$.

By the Bernoulli's inequality, we have

$$\hat{K}_1^* < \frac{1}{4}\frac{8N(N^2 + 2N + 1 + \rho_n)}{2N^2 + 6N + \rho_n} = \frac{2N(N^2 + 2N + 1 + \rho_n)}{2N^2 + 6N + \rho_n}.$$

$RHS < N$ if $\rho_n < 2(N-1)$. If the inequality holds, vendor 1 sell $\hat{K}_1^*$ and vendor 2 share the remaining part $N - \hat{K}_1^*$. Otherwise, both sellers try to sell as much non-fundamental data as possible, and any quantity $K1, K2$ such that $K_1 + K_2 = N$ constitutes equilibrium. In aggregate, the equilibrium sale of non-fundamental data is always $K^V = N$.

□