China Walls

Tomy Lee Daniel Nathan Chaojun Wang*

May 17, 2025

Abstract

Regulators manage conflicts of interest within banking conglomerates by enforcing *China Walls*—internal information barriers around dealers. To evaluate if today's China Walls are effectively enforced, we map information sharing between dealers and funds using the universe of Israeli Shekel foreign exchange trades. Our design compares trading activities of affiliates against entirely unrelated firms around exceptionally large trades to detect information sharing. We document islands of informational autarky between dealers and their affiliate funds surrounded by a sea of information sharing. *(i)* A dealer never trades nor shares information with its affiliated funds. *(ii)* Dealers consistently share information with their client funds, including on days when they do not trade with each other. *(iii)* Affiliated funds, which are free to share information with each other, intensely do so among themselves. From a back-of-the-envelope calculation, establishing China Walls between affiliated funds would eliminate $16.1 billion in trades, comprising 37% of their trades on the event dates. Our results hold during crisis and noncrisis periods, and across granular cells of firm and asset characteristics. We reveal remarkable regulatory capacity to control information flows.

JEL classification: G21, G28, G14, G15

Keywords: Banking conglomerates, financial networks, information barriers, information sharing, regulatory capacity

*First version: November 20, 2024. Lee is at the Central European University. Nathan is at the Bank of Israel. Wang is at the Wharton School, University of Pennsylvania. We thank Markus Bak-Hansen, David Card, Andras Danis, Xavier D'Haultfoeuille, Felix Fattinger, Thomas Gehrig, Sasha Indarte, Simon Jurkatis, Attila Lindner, Florian Nagler, Gabor Pinter, Adam Szeidl, Toni Whited, Adam Zawadowski, and Josef Zechner for valuable comments. Emails: leeso@ceu.edu; daniel.nathan@boi.org.il; wangchj@wharton.upenn.edu.

1 Introduction

Banking conglomerates are rife with conflicts of interest. They manage funds and run broker-dealers that intermediate financial markets, all while investing on their own accounts. To limit these conflicts, regulators in the US increasingly enforce *China Walls*—blunt information barriers around broker-dealers, which are particularly exposed to conflicts of interest.[1]

Enforcing China Walls is a formidable challenge. Information sharing among affiliates occurs in private, is plausibly deniable, and yields large conglomerate-wide payoffs. More fundamentally, affiliates have tightly aligned incentives, precluding counterparty litigation that is often crucial to regulatory enforcement. As such, effectively enforced China Walls would reveal remarkable regulatory capacity to control information flows—especially relevant today, when concerns over privacy are widespread. Are today's China Walls effectively enforced within banking conglomerates?

We document that they are. Our empirical challenges mirror that of the regulators: information sharing is not directly observable, and compliance in one circumstance does not rule out violations at other times. We compare trading activities of dealers and funds around exceptionally large trades to overcome these challenges. Our difference-in-differences design detects information sharing between a dealer and its affiliate funds if the funds increase trading on days that the dealer makes an exceptionally large trade relative to funds that are entirely unrelated to the dealer. Three plausible assumptions underpin our design. First, exceptionally large trades pinpoint arrivals of especially valuable

[1] "China Walls," or the more common "Chinese Walls," is a reference to the Great Wall of China (Gozzi, 2003). "Information barriers," "firewalls," "ethical screens," and "insulation walls" are synonymous terms that appear later. We adopt "China Walls," because it is concise, does not have a common alternative meaning, and is the closest to the original reference.

private information, when there would be the strongest incentive to violate China Walls. Second, funds increase trading activity upon receiving valuable information. Third, dealers never share private information with unrelated funds that are neither their affiliates nor clients.

We implement this design on the near universe of foreign exchange trades involving the Israeli Shekel covering 21 million trades from 2019 to 2024. Among them, 87% are trades in the US dollar-Shekel currency pair. Moreover, the largest dealers in the Shekel market are identical to those in the broader US dollar market and Israeli financial regulations are mainly based on US regulations. An exception is that Israel does not impose China Walls, leaving the US regulators (whose jurisdiction reaches worldwide) as the main enforcers of China Walls in our setting. Their rules wall off dealers from their affiliate funds, while leaving funds affiliated to each other free to share information among themselves. Appendix A details the legal context.

Figure 1 illustrates our design. GS Dealer and GS Fund are affiliates. (GS, MS, and BoA are illustrative names.) Unrelated Fund is unaffiliated and never trades with the other firms in the figure. An event is an exceptionally large trade (event trade) by the GS Dealer (event firm) that belongs in the top 0.1 percentile of the GS Dealer's trades. We compare the daily gross dollar volumes of the GS Fund (affiliate firm) and the Unrelated Fund (control) around the event day. We conclude that the event dealers share information with their affiliate funds if the daily volumes of the affiliate funds increase relative to the unrelated funds around the event day.

This approach detects no information sharing from dealers to their affiliate funds nor, reversing their roles, from funds to their affiliate dealers. Richness of our setting provides two falsification tests. First, we verify whether our design reliably detects information

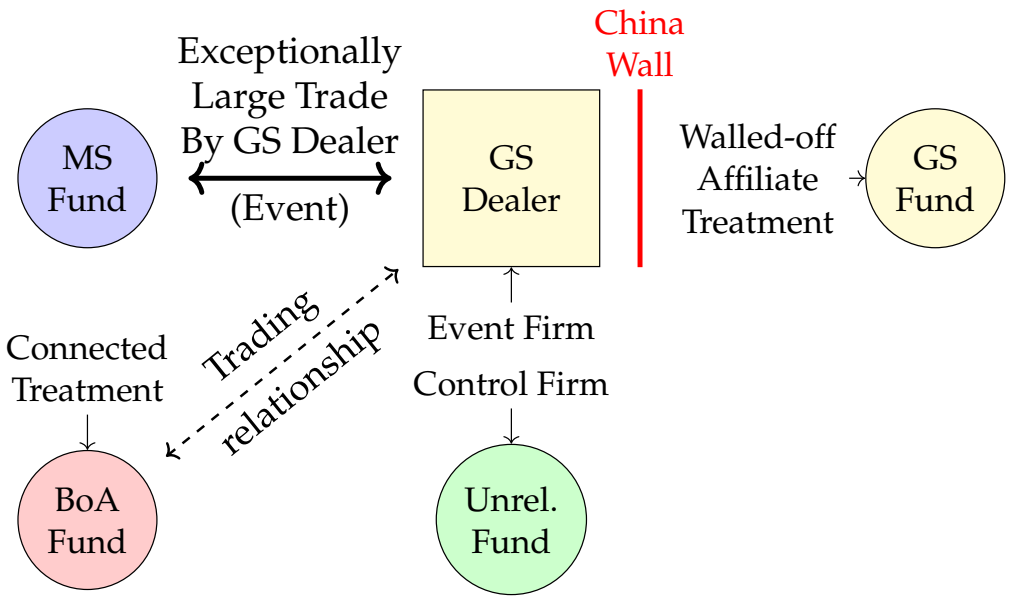


Figure 1: Identifying Information Sharing from Dealers to Affiliate Funds

sharing where it exists. Since dealers are well known to share information with their client funds (Barbon, Di Maggio, Franzoni, and Landier, 2019; Boyarchenko, Lucca, and Veldkamp, 2021), a reliable design must detect information sharing between such connected dealers and funds. In Figure 1, the BoA Fund is a client of the GS Dealer. Our first falsification test compares the daily volumes of the BoA Fund (connected firm) and the Unrelated Fund (control) around the day of the GS Dealer's exceptionally large trade. We consistently detect information sharing between connected dealers and funds. Second, we exploit funds that are affiliates but not walled off from each other to determine whether affiliates do share information with each other in the absence of China Walls. Affiliate funds intensely share information among themselves, and thus we infer that affiliate dealers and funds would share information if the China Walls were absent.

Section 2 develops the design. Our key identifying assumption is that especially valuable information prompts exceptionally large trades. This assumption is consistent with

standard theory (Kyle, 1985; Easley and O'Hara, 1987) and empirically holds in other markets (Kumar, Mullally, Ray, and Tang, 2020; Pinter, Wang, and Zou, 2024). A threat is the possibility that firms would split orders to disguise their private information. Appendix B jointly tests this assumption and the claim that our design isolates information sharing. Consistent with these claims, exceptionally large trades predict future returns, smaller trades do not, and we detect information sharing between connected dealers and funds only around the large trades.

We then strip away three sources of confounding variation in trade volumes. First, public news or aggregate shocks can simultaneously trigger the funds to increase trading and the dealers to make exceptionally large trades. Second, the liquidity and price impacts of the event trades, rather than information sharing, can induce funds to increase trading. Because no dealer would share information with a fund that is neither an affiliate nor a client—and yet these unrelated funds are as exposed to the aggregate shocks and the impacts of event trades as other funds—using the unrelated funds as controls removes the two confounders while preserving any variation due to information sharing. And third, we may be omitting relevant characteristics of events and funds. Our calendar date, days-relative-to-event, and event-by-fund fixed effects eliminate any confounding variation that is common across funds over time, or specific to a fund as long as it is invariant over the two weeks (the event window) around the event.

Section 3 describes the data. There are 7,700 funds, 46 conglomerates that control dealers, and 17,000 events in our sample. The dealers virtually *never* trade with their affiliate funds, perhaps due to the onerous constraints of the dealers' China Walls. Our main analyses test whether the China Walls preempt information sharing in addition to barring trades between affiliate dealers and funds.

Section 4 implements our design in stacked difference-in-differences specifications with never-treated controls of Cengiz, Dube, Lindner, and Zipperer (2019). Our analytical samples contain millions of observations across thousands of events and firms, providing the power to detect even tiny differences between treated and control groups. Despite the high power, in the 11 trading days around an exceptionally large trade by an event dealer, the daily gross dollar volumes of funds affiliated to this dealer are statistically indistinguishable from those of unrelated funds, differing by -0.02 standard deviation on the event day (clustered std. error: 0.04 sd). In stark contrast, the funds connected to the event dealer increase their volumes by a precisely estimated 1.9 sd (std. error: 0.007 sd) on the event day relative to the unrelated funds. Likewise, around a day when an event fund makes an exceptionally large trade, the gross volumes of its affiliate and unrelated dealers are indistinguishable from each other, whereas its connected dealers sharply increase their volumes on the event day relative to the unrelated dealers. All results remain when we replace gross volumes with net volumes signed in the direction of the event trade.[2]

Section 5 applies this design to funds affiliated to each other. On a day when an event fund makes an exceptionally large trade, the funds affiliated to the event fund increase their volumes by 1.7 sd (std. error: 0.2 sd) relative to the unrelated funds. Taken together, we reject information sharing between dealers and their affiliate funds—exactly where China Walls are present—and detect extensive information sharing elsewhere, both among affiliated funds and between dealers and their clients. We conclude that China Walls are effectively enforced.

[2]We do not observe who initiated each trade. To proxy the direction of each event trade, we assume that (i) any trade between a dealer and a fund is initiated by the fund, and (ii) for event trades between two dealers, the event dealer initiated the trade. We focus on gross volume, because these assumptions add noise to our net-volume estimates.

We address two key threats to this conclusion. First, trades between dealers and their client funds might generate mechanical increases in trading around event trades. We exclude, from each event, any fund or dealer that trades with the event firm on or up to five days after the event day, precluding such mechanical effects. Second, funds affiliated to each other may be exposed to common shocks through shared dealer connections. To remove these shocks, we flexibly control for overlaps in the sets of dealer connections between the event fund and its affiliate funds.

Section 6 scours granular cells of key event, asset, and firm characteristics for China Wall violations. We never detect information sharing between affiliate dealers and funds across crisis and noncrisis periods, asset classes, currency pairs, and fund types. We consistently detect information sharing among affiliated funds and between dealers and their clients. Moreover, hedge funds respond more intensely to event trades than other funds, and particularly so when the event trade was by another hedge fund (Table 4), echoing evidence that hedge funds are more informed and more sensitive to information than other funds (Di Maggio, Franzoni, Kermani, and Sommavilla, 2019; Kumar et al., 2020). Events are more likely during crisis periods, and yet treated firms' responses to crisis and noncrises events are precisely equal (Table 5)—our design fully strips away aggregate shocks and any variation that correlates with such shocks. Last, an event prompts the largest responses by connected firms that specialize in the currency pair and asset class of the event trade (Tables 6 and 7): homophily one would expect if the event trades indicate the type of information they embody. It is hard to imagine a confounder that can plausibly explain this rich combination of results.

Context and previous work on China Walls. The US regulators did not enforce China Walls before 2018. Previously, banking conglomerates voluntarily adopted China Walls to

protect against corporate liability from insider trading by their employees. (The employees themselves would remain liable.) The 2010 Dodd-Frank Act allowed US regulators to conduct "risk-based" enforcement, under which they can prosecute firms for practices that substantially raise the risk of a crime, even without evidence that the crime has actually occurred. The US Securities and Exchange Commission (SEC) began to exercise this power to enforce China Walls in 2018: insufficiently maintaining China Walls itself, not only insider trading, is now a prosecutable offense. Appendix A provides further detail.

Existing evidence on China Walls exploit samples that predate 2018. This evidence identifies extensive violations, as legal proceedings would eventually confirm.[3] We instead evaluate the China Walls during the recent period of their active enforcement. As importantly, we contribute a novel identification strategy that uses unrelated funds as controls to isolate the effects of information sharing. We validate our strategy in conditions where the China Walls are absent and information sharing is expected. Applying this design to a large and granular dataset yields precisely estimated and robust evidence that today's China Walls effectively preempt information sharing.

Broader contributions. We belong to the literature on the capacity of states to regulate firms.[4] In their settings, a large extent of regulatory enforcement occurs through private litigation by parties involved in the regulated activity (e.g., employer vs employee, insider vs outside shareholder; Glaeser and Shleifer (2003), La Porta, Lopez-De-Silanes, and Shleifer (2006)). In our setting, a China Wall violation involves affiliates under common

[3] Lehar and Randl (2006), Irvine, Lipson, and Puckett (2007), Seyhun (2008), Massa and Rehman (2008), Chen and Martin (2011), Ivashina and Sun (2011), Li (2018), Li, Mukherjee, and Sen (2021), Kondor and Pintér (2022), and Haselmann, Leuz, and Schreiber (2023) find evidence of China Wall violations in various settings. The latest in-sample year among them is 2017.

[4] Regulators have greatly reduced pollution (Keiser and Shapiro, 2019; Behrer, Glaeser, Ponzetto, and Shleifer, 2021), insider trading (Bhattacharya and Daouk, 2002), misleading financial disclosures (Greenstone, Oyer, and Vissing-Jorgensen, 2006), and discrimination in pay (Bailey, Helgerman, and Stuart, 2024) and access to accommodation (Cook, Jones, Logan, and Rosé, 2023).

corporate control, eliminating the threat of counterparty litigation. Moreover, bankers often communicate in plausibly deniable ways (Peluso, 2020). Therefore, our results reveal a remarkable regulatory capacity to control information flows beyond what is established in prior work.

We extend the empirical literature on information diffusion in financial markets. Dealers extract information from their clients' order flow (Hortaçsu and Kastl, 2012) and leak information to certain clients (Barbon et al., 2019; Boyarchenko et al., 2021; Chague, Giovannetti, and Herskovic, 2023). More broadly, dealers act as conduits through which information diffuses throughout their trading networks (Di Maggio et al., 2019; Hagströmer and Menkveld, 2019; Kumar et al., 2020). We identify a stark void in this informational network driven by regulatory intervention, thereby adding China Walls as a promising source of variation in information flows that is especially relevant today, when the financial sector is highly concentrated.

Roadmap. Section 2 develops the empirical design. Section 3 describes the data and performs motivating analyses. Sections 4 and 5 investigate the effectiveness of China Walls. Section 6 performs the heterogeneity analyses.

2 Design

2.1 Context

China Walls refer to a collection of rules and physical barriers that aim to preempt the flow of material private information (MPI) to or from the walled-off subsidiaries of banking conglomerates and their affiliates. An MPI is any information that (a) a reasonable investor would find important for her investment decisions and (b) is not publicly

disclosed.[5] For example, proprietary analysis, inside information, or private trade requests would constitute MPI. Typical China Walls require walled-off subsidiaries to be isolated via separate entrances, opaque and soundproof barriers, and the monitoring and recording of their employees' communications.

New regulations since the 2008 financial crisis established today's China Walls around broker-dealers within banking conglomerates (and bank-owned investment advisers, which we do not examine). Today, the US SEC routinely imposes large fines for deficiencies in the dealers' China Walls. Appendix A details relevant definitions, history and legal precedents, impacts of the Dodd-Frank Act, and recent enforcement cases.

Empirical setting. The foreign exchange market is an over-the-counter market, in which trades occur between dealers or a dealer and its client. The dealers are long-lived, trades are nonanonymous, and most firms rely exclusively on one or a few relationship dealers. Hence, reputation concerns preclude behavior frequently seen in centralized markets, such as repeated order submissions without the intent to trade or splitting a large trade quantity into a rapid sequence of small orders. This market operates at high frequency, where news is rapidly incorporated into exchange rates. Therefore, we do not expect private advantage from an MPI to last beyond a few trading days.

Our data covers the near universe of Israeli Shekel (ILS) foreign exchange transactions, which we obtain from the Bank of Israel. The ILS market structure is identical to the other foreign exchange markets. Indeed, 87% of ILS transactions are for the USD-ILS pair and the ILS and the USD markets have the same largest dealers.[6] More broadly, financial

[5] Material non-public information (MNPI) is the more commonly referred type of information in law. The MPI includes analyses based purely on public information, whereas the MNPI expressly excludes such analyses. We use MPI rather than MNPI since proprietary analysis is valuable private information.

[6] The share of USD in our sample is remarkably close to the 85% of all foreign exchange transactions that involve the USD (Somogyi, 2022).

regulations in Israel are largely based on the US. A peculiar Israeli law forbids Israeli holding companies from owning both a dealer and a nondealer investment firm, as the US Glass-Steagall Act did until its 1999 repeal. As such, the Israeli regulators neither mandate nor enforce the China Walls—the banking conglomerates do not incriminate themselves when reporting data at odds with their China Walls to the Bank of Israel. The enforcers of the China Walls in our setting are the nonIsraeli regulators, especially the US SEC whose jurisdiction extends to all banking conglomerates active in the US (every conglomerate in our sample).

2.2 Empirical Design

We must overcome three challenges to test the hypothesis that the China Walls are effectively enforced. First, the China Walls may be violated in circumstances that we do not examine. In particular, the test may neglect the circumstances where China Wall violations are the most likely to occur. Second, we need a proxy that isolates the variation specifically due to bilateral MPI sharing, all while reliably detecting any such information sharing. Third, the bank-owned dealers may not share MPI with their affiliates even if their China Walls were absent, in which case enforcement is moot.

Defining events. We seek events that pinpoint when a dealer or a fund receives especially valuable MPI. Under the plausible assumption that China Wall violations are the most likely to occur when there are the largest gains from sharing information with affiliates, rejecting violations during such events would also rule out violations at other times. Standard theory shows that an informed trader submits larger trades when she holds more valuable private information (Kyle, 1985; Easley and O'Hara, 1987). Empirically, the trades that are unusually large compared to its trader's other trades are particularly

predictive of returns (Kumar et al., 2020; Pinter et al., 2024). Appendix B presents concurring evidence in our setting. Therefore, we let an event be a dealer or a fund (a firm) and a day (event day) when the event firm makes a trade (event trade) that is exceptionally large compared to the its other trades.

Isolating information sharing. We consider an event firm i and a treated firm j such that, if i is a dealer, then j is either an affiliate fund or a client (i.e., connected) fund. A proxy for MPI sharing from the event firm i to the treated firm j must isolate information that is (i) material and (ii) bilaterally shared. Information is material only if it is important for determining the firms' optimal portfolios. Receiving an MPI would prompt firm j to rebalance its portfolio towards the new optimum, increasing its daily gross volume. Alternatively, j may become more likely to trade in the direction of the price change that the MPI predicts. Therefore, we choose increases in the gross volume of firm j to proxy for the sharing of MPI by the event firm i to j, and confirm that our results are robust to using net volumes signed in the direction of the event trades.

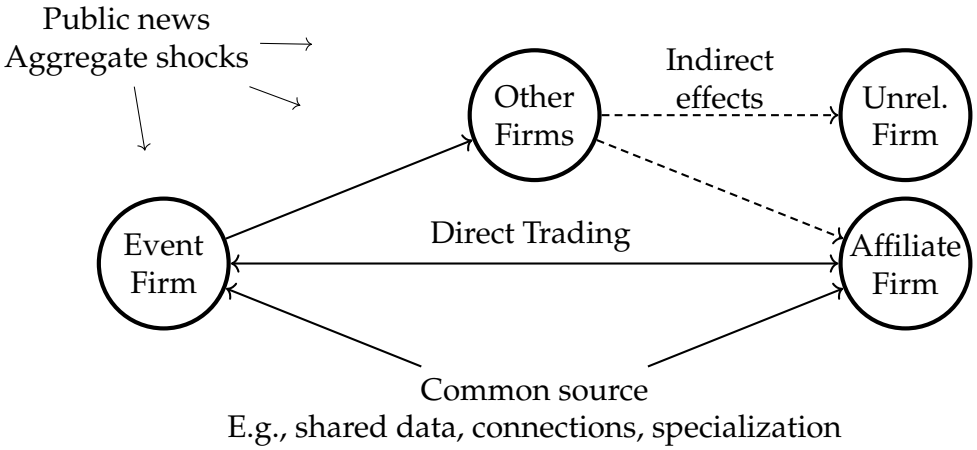


Figure 2: Potential Confounders to Measuring Bilateral Information Sharing

To isolate bilateral MPI sharing, we remove each of the four confounders that can also

generate the coincidence of firm i's exceptionally large trades and increases in firm j's gross volume. Figure 2 illustrates the confounders. First, any direct trade between i and j could mechanically induce both the event trade and the increased gross volume, for instance as the large trade itself causes the increase in gross volume. This confounder does not apply when firms i and j are affiliated or unrelated (i.e., neither affiliated nor ever trades with i in our sample), since the former do not trade in our data and the latter do not by definition. In the case where i and j are connected, we shut down the confounder by excluding firm j for the events in which j traded with i on or after the event day up to the end of the event window (five days after the event day).

Second, arrivals of public news or other aggregate shocks may trigger all firms to trade, including the event trades. Third, event firm i's MPI that corresponds to the event trade may indirectly induce firm j to increase its trading activity. Either the event trade itself or any sharing of the MPI by i to firms other than j could affect liquidity or prices throughout the market. As these liquidity or price impacts reach firm j, they may prompt j to increase trading. For example, if an event trade is between dealer i and another dealer, the second dealer might contact fund j to offload the newly gained inventory. Fund j's gross volume would then increase if it agrees to this trade, or if the contact reveals information to j.

We filter out the aggregate-shock and the indirect-impact channels by comparing the gross volume of firm j and those of the firms that are unrelated to the event firm i. The gross volumes of the unrelated firms would capture the aggregate shocks and the indirect impacts of firm i's event trade around the event day. At the same time, the event firm i would not share MPI with unrelated firms. Hence, our design detects bilateral MPI sharing from i to j if and only if the gross volume of j increases relative to unrelated firms on

or after the event day.

Fourth, a source common to firms i and j, but not to the unrelated firms, may simultaneously trigger i's event trade and heighten j's gross volume. We present three examples, in which we suppose that i and j are affiliated or connected with each other. Firms i and j may be more likely than two unrelated firms to have common data or research sources. They may also be more likely to specialize in the same assets, or be connected to the same third firm. In these examples, any information from a common data source would sometimes reach j before i, and likewise for those from asset-specific shocks or the shared connection. This noise in timing would generate pretrends before the event dates. We reject the presence of the common-sources channel if j and unrelated firms show parallel trends prior to the event day.

Connected treatment. A key remaining threat is the possibility that our design does not reliably detect bilateral MPI sharing where it exists. We exploit the stylized fact that a connected dealer and fund extensively share information (Barbon et al., 2019; Kumar et al., 2020; Chague et al., 2023) to falsify the reliability of our design to detect bilateral MPI sharing. If our design is reliable, then we will detect the sharing whenever i and j are connected to each other. Thus, we falsify the reliability of our design if the daily gross volumes of connected firms do not increase relative to the unrelated firms on or after the event day. We strengthen this falsification test by excluding the connected firms that trades with the event firm on or after the event day.

2.3 Implementation

We adopt the stacked difference-in-differences specification with never-treated controls of Cengiz et al. (2019).[7] An event is a firm and a date on which the firm made a trade in the 0.1 percentile of its trades by dollar value or its largest trade if the firm made fewer than 1,000 trades in our sample. All event trades by the same firm on the same day are combined into a single event. A firm is treated on or after the event day within the event window if the firm is an affiliate or a connection of the event firm. A firm is a control if it is unrelated to the event firm and not treated in any other event during the event window. Our event window is the 11 trading days around the event day, because exchange rates fully incorporate private information in about a trading week (Menkhoff, Sarno, Schmeling, and Schrimpf, 2016).

Our first regression specification is

$$Y_{ejt} = \sum_{\tau=-5}^{5} \alpha_\tau \mathbb{1}_{t=\ell_e+\tau} Affiliate_{ej} + \delta_{ej} + \varphi_t + \sum_{\tau=-5}^{5} \gamma_\tau \mathbb{1}_{t=\ell_e+\tau} + \varepsilon_{ejt}. \tag{1}$$

The dependent variable Y_{ejt} is the gross dollar volume of firm j on calendar date t and event e, standardized at the firm level. The affiliate treatment dummy $Affiliate_{ej}$ equals 1 if firm j is an affiliate of the event firm. The dummy $Affiliate_{ej} = 0$ if j is unrelated to the event firm and is not treated in any other event within the 21-day panel around event e. The indicator variable $\mathbb{1}_{t=\ell_e+\tau}$ equals 1 when t equals the event day ℓ_e shifted by τ days, and 0 otherwise. We control for event-by-firm, calendar date, and event date fixed effects δ_{ej}, φ_t, and γ_τ. These effects control for event-and-firm-specific factors as well as common

[7] This implementation yields average treatment-on-the-treated (ATT) effect estimates that always place positive weights on all groups (Gardner, 2022), unlike those of traditional staggered two-way fixed-effects difference-in-differences specifications (Roth, Sant'Anna, Bilinski, and Poe, 2023).

trends over calendar and event times. We cluster standard errors by event-and-firm and by calendar date, because our treatments are assigned event-by-firm and the incidence of events varies over time. Our data contains the near universe of transactions in the currency pairs we examine, as detailed in Section 3, implying a high sampling probability. Therefore, the clustered variances likely approximates the true variances (Abadie, Athey, Imbens, and Wooldridge, 2023).

The second specification repurposes Equation (1) to measure the MPI sharing between connected dealers and funds,

$$Y_{ejt} = \sum_{\tau=-5}^{5} \beta_\tau \mathbb{1}_{t=\ell_e+\tau} Connected_{ej} + \delta_{ej} + \varphi_t + \sum_{\tau=-5}^{5} \gamma_\tau \mathbb{1}_{t=\ell_e+\tau} + \varepsilon_{ejt}. \tag{2}$$

The connected treatment dummy $Connected_{ej}$ equals 1 if (a) firm j trades 10 or more times with the event firm in the sample, and (b) does not trade with the event firm on the event day and five days afterwards, $t = \ell_e, \ldots, \ell_e + 5$. Condition (a) restricts the connected firms to nonaffiliates, because exactly zero pair of affiliate dealer and fund trades 10 or more times. Condition (b) removes any mechanical increase in the gross volumes of the connected firms relative to the unrelated firms due to trades with the event firm. The conditions for $Connected_{ej} = 0$ and $Affiliate_{ej} = 0$ are identical, and the other elements in Equation (2) are the same as the corresponding elements in Equation (1).

We estimate each of Equations (1) and (2) twice. Either the dealers are the event firms and we examine the daily gross volumes of the funds, or the funds are the event firms and we examine the volumes of the dealers.

2.4 Identification Tests

We assume that a firm trades an exceptionally large size when especially valuable material private information arrives at the firm. The underlying claim is that firms submit larger orders when it has more valuable private information. This claim is consistent with standard theory (Kyle, 1985; Easley and O'Hara, 1987) and recent empirical evidence in over-the-counter markets that larger trades by each firm are more predictive of returns than its smaller trades (Kumar et al., 2020; Pinter et al., 2024). However, the theory on order splitting (Bernhardt and Hughson, 1997) and the lack of similar evidence specifically on the foreign exchange market question our claim.

Appendix B adjudicates our assumptions in the data. Placebo tests using small and medium trades as event trades jointly falsify two claims. (i) Exceptionally large (0.1 percentile) trades indicate arrivals of especially valuable MPI. (ii) Our design isolates bilateral MPI sharing, in that it yields significantly positive treatment coefficients if and only if the event firms bilaterally share MPI with the treated firms. We define a small event as a firm and a day when the firm makes a trade in the 99.9 to 100th percentile of its trades by dollar volume, and a median event as the same except in the 50 to 50.1st percentile. If informed firms do not split orders, and rather trade exceptionally large sizes to exploit especially valuable MPI, the large trades would predict returns and smaller trades would not. Moreover, if our design isolates information sharing, connected firms would increase their volumes relative to unrelated firms around the dates of trades that predict returns, and not for nonpredictive trades.

We find that the exceptionally large trades predict returns up to three days following the trade. The small and medium trades do not predict returns. Moreover, we find zero evidence of an increase in the gross volume of connected firms relative to unrelated firms

around the small-event or the median-event days. We conclude that our design isolates the sharing of especially valuable MPI.

3 Data and Descriptive Results

3.1 Data

We obtain the near universe of foreign exchange transactions involving the Israeli Shekel from the Bank of Israel (the Bank) in the sample period January 2019 to March 2024, spanning 1,368 trading days.[8] Each observation specifies the currency pair (ILS and another currency), price, quantity, date and time,[9] asset class (spot, forward, swap, or option), and the counterparty names. We exclude options due to insufficient observations and convert all nonUSD transaction values into USD at the contemporaneous official exchange rate published by the Bank.

Table 1 summarizes the samples we analyze. A three-step process generates samples whose observations are firm-by-date. First, we consolidate dealers up to the conglomerate by dropping all trades between dealers affiliated to each other and combining them under conglomerate-level labels. We do so, because a group of affiliate dealers are free to split incoming orders and transfer assets and capital among themselves, and are thus effectively a single economic entity. As such, consolidating affiliate dealers minimizes

[8] All Israeli firms, including the Israeli branches of conglomerates, must report each ILS transaction to the Bank. Non-Israeli firms fall under the same reporting requirement if their foreign exchange transactions in the previous year exceed $15 million per day on average, whether on their own accounts or on behalf of clients. This reporting requirement applies to practically all significant financial firms, because any foreign currency spot or derivative transaction is included in the reporting threshold, even if the firm rarely trades ILS. Rules can be retrieved from https://www.boi.org.il/en/economic-roles/statistics/reports-to-bank-of-israel/reporting-on-activity-in-the-foreign-currency-derivative/.

[9] We do not exploit intraday time stamps, because a large proportion of trades report 00:00:00 rather than the actual trade time. (The Bank only requires that firms report the correct date, not time.)

noise from nonmarket transactions that shift cash and inventory for tax or balance sheet purposes.[10] Second, we aggregate the daily gross volumes of each dealer and fund across asset classes (i.e., spot, forward, and swap). While aggregating, we keep the notional amount from each swap trade's first leg and ignore its second leg to avoid double counting. Third, we winsorize observations in the top 0.5 percentile by daily gross volume, calculated separately for dealers and for funds, because their daily volume distributions differ dramatically.

Affiliations. A four-step procedure identifies the affiliations of all firms. First, we determine the affiliations of most US-based firms using the quarterly organizational hierarchy data accessible via the National Information Center (https://www.ffiec.gov/npw/). We assign affiliations to firms as of last quarter, 2023, for the whole sample period, because financial firms rarely change their affiliations and typically change their legal names when they do. Second, all firms with obviously indicative names are linked to the corresponding conglomerate (e.g., "Deutsche Bank Luxembourg S.A."). Third, the remaining firm names are entered into `ChatGPT 4.0` as a query in the form, "as of `[date the firm last appears in the sample]`, is `[firm name]` independent? If not, which holding company does `[firm name]` belong to?" Fourth, we manually verify each answer generated in step three, by searching for the firm name paired with "independent" or the `ChatGPT`-suggested holding company name.

[10] Some 8% of foreign exchange spot trades are "back-to-back" trades between affiliate dealers for accounting or inventory rebalancing reasons (Bank for International Settlements, 2022). All trades by affiliate funds are market-based, since they only trade with nonaffiliate dealers.

Table 1: Sample Characteristics

			Final Sample	
	All trades	Fund trades	Dealer day	Fund day
Mean daily volume (USD millions)	29,510	2,843	19,940	2,843
Mean daily volume per firm (USD millions)	642	3.7	433	469
Dollar value per observation (USD millions)	2.7	1.7	635	0.34
Currency				
USD	0.87	0.76	0.94	0.76
JPY	0.07	0.22	0.004	0.22
EUR	0.02	0.02	0.03	0.02
Asset class				
Spot	0.36	0.50	0.32	0.50
Forward	0.13	0.40	0.11	0.40
Swap	0.50	0.10	0.58	0.10
Observations	20,832,686	2,762,406	62,974	10,643,975

All trades: Raw data set containing the near universe of Israeli Shekel transactions. *Fund trades*: Transactions involving a fund. *Dealer day*: Dealer transactions aggregated to the daily gross dollar volume in USD; excludes trades between dealers affiliated to each other and trades with nonfinancial firms. *Fund day*: Fund transactions aggregated to the daily gross dollar volume in USD. *Mean daily volume* is the average daily total dollar volume in USD billions. *Mean daily volume per firm* is the mean daily volume divided by the number of firms in the sample. All Currency and Asset class figures are weighted by dollar volume.

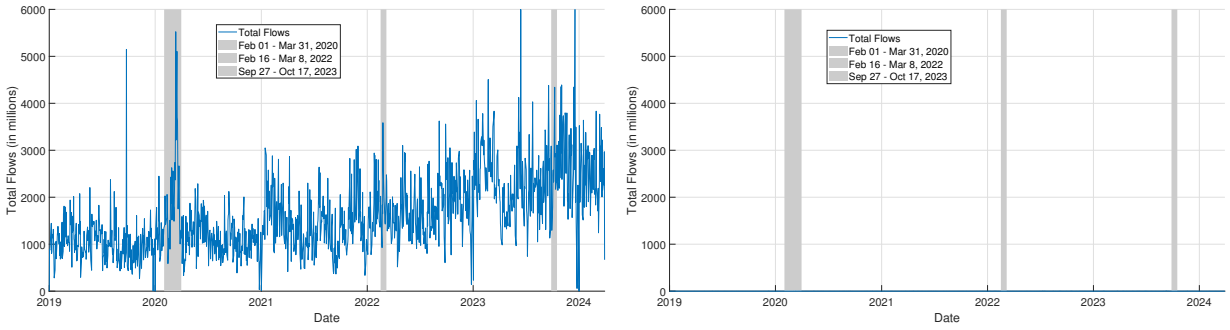
Table 2: Number of Unique Entities

	Conglomerates	Dealers	Funds
US	15	92	4,826
Israeli	11	15	192
Independent	–	11	6,660
Hedge funds	–	–	632
Total	46	229	7,775

A conglomerate is a holding company and the group of firms that the holding company ultimately controls. "Dealers" also include brokers and broker-dealers. All dealers in our sample are broker-dealers, which match client orders or trade on their own accounts at the their discretion. "Independent" denotes entities that do not belong to a conglomerate. All independent dealers are Israeli, due to Israeli law that forbids common ownership of banks and dealers.

3.2 Motivating Analyses

Three analyses motivate our main empirical design. First, Figure 3 plots the total daily dollar volume of transactions. The dealers trade USD2.8 billion with the funds daily, of which near precisely zero is with their affiliate funds—there are four trades between a dealer and an affiliate fund, worth a mere USD5.51, in our sample.



(a) Unaffiliated Dealers and Funds

(b) Affiliate Dealers and Funds

Figure 3: Daily Gross Dollar Volumes Traded Between Dealers and Funds

Figure 3a: The sum of daily gross dollar volume in USD millions across pairs of dealer and fund that are not affiliated with the same banking conglomerate. Figure 3b: The sum of daily gross dollar volume in USD millions across pairs of affiliate dealer and fund. Shaded regions mark the onsets of the Covid pandemic, the Russian Invasion of Ukraine, and the Hamas attack on Israel.

Second, Figure 4 computes the correlation in daily gross volumes within unrelated dealer-fund pairs. For each lag $l = -10 \ldots + 10$ and a pair of dealer i and fund j that are nonaffiliates and do not trade in the sample, we compute the correlation $CorrGV_{ijl}$ between the date t gross volume of i and date $t + l$ gross volume of j. We average this correlation across the unrelated dealer-fund pairs for each l. Figure 4a plots the results. There are strongly positive and significant correlations in trading activity among the unrelated dealers and funds. Absent a control group, the common shocks driving comovement among the unrelated firms may severely contaminate measures of bilateral information sharing.

Third, we estimate a simplified version of our main specifications (1)-(2). We compare the correlations $CorrGV_{ijl}$ within the affiliate and the connected dealer-fund pairs against the unrelated pairs. Doing so tests whether the trading activities of affiliates and connected firms correlate once stripped of common shocks. Our implementation uses the regression specification

$$CorrGV_{ijl} = a_l Affiliate_{ij} + b_l Connected_{ij} + c_i + d_j + \varepsilon_{ijl}. \tag{3}$$

The dummy variable $Affiliate_{ij}$ equals 1 if dealer i and fund j are affiliates and 0 if they are unrelated. The dummy $Connected_{ij}$ equals 1 if i and j trades 10 or more times in the sample and 0 if they are unrelated. We exclude the trades between i and j to compute $CorrGV_{ijl}$, which avoids mechanical correlations due to within-pair trades. The dealer and the fund effects c_i and d_j control for time-invariant factors specific to each dealer and each fund.

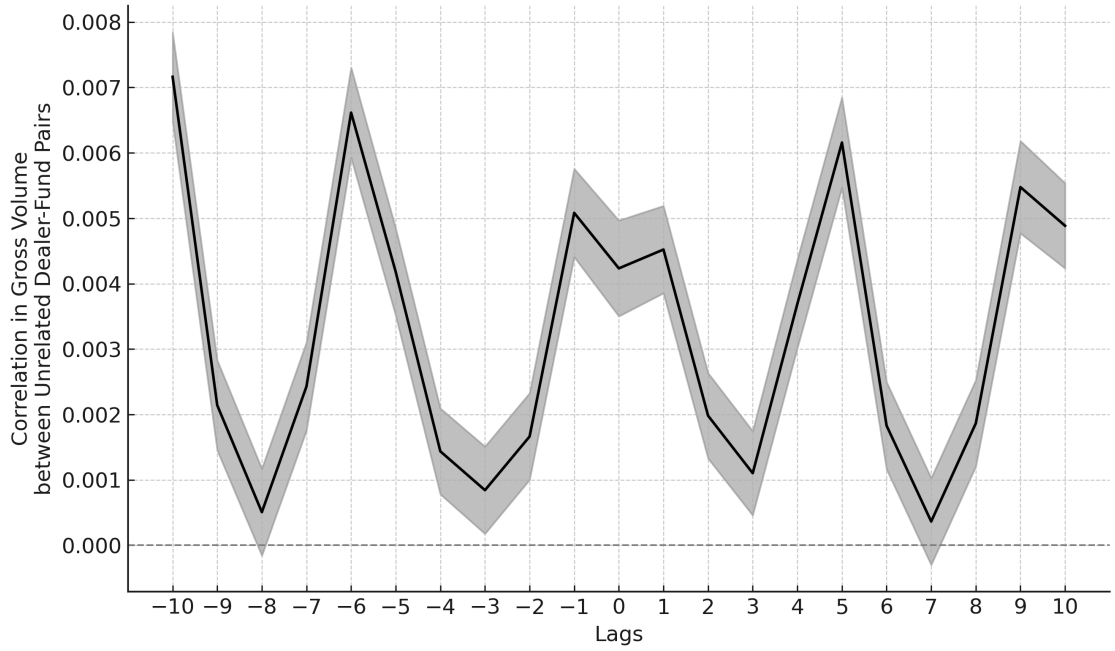
Figure 4b plots the coefficients c_l and d_l across $l = -10 \ldots 10$. Daily gross volumes

of the affiliate dealer-fund pairs are no more correlated than those of the unrelated pairs across all lags l. In contrast, the connected dealer-fund pairs are sharply more correlated contemporaneously than the unrelated pairs. These correlations suggest that China Walls effectively block material information flows between walled-off firms, while information flows freely among connected firms. Our main design isolates bilateral information sharing and focuses on the dates when there is the greatest incentive to violate the China Walls.

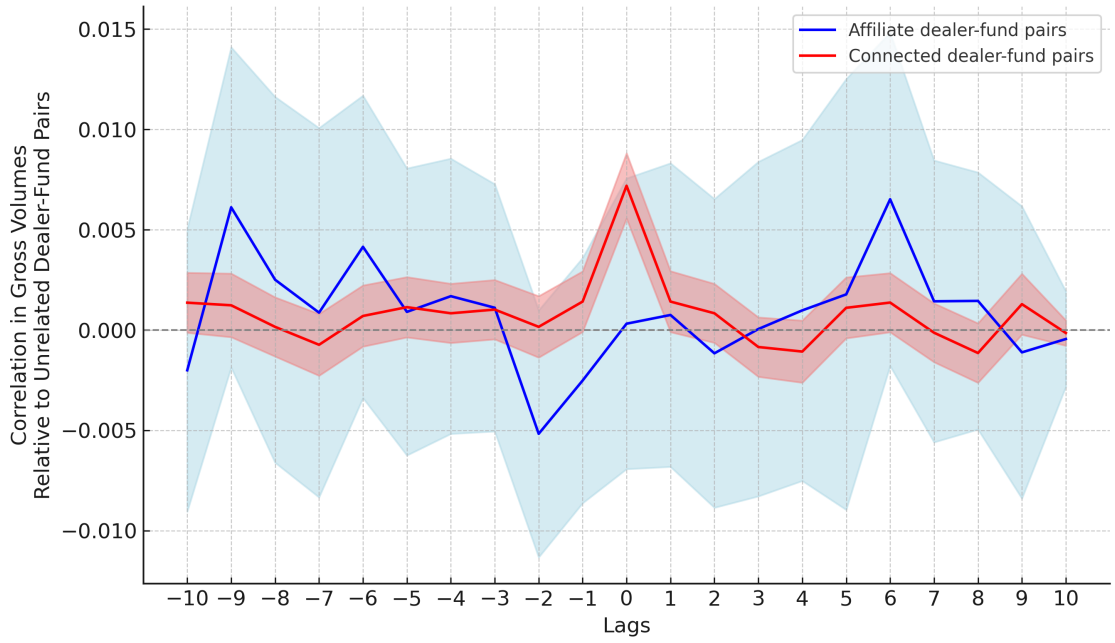
4 Are China Walls Effectively Enforced?

We first estimate Equations (1) and (2) selecting the dealers as the event firms and the funds as the treated and the control firms. Figure 5a plots in blue the differences α_τ in standardized gross volume between affiliate and unrelated funds around the days of exceptionally large trades by dealers, and in red the differences β_τ between the connected and the unrelated funds. The affiliate funds exhibit neither pretrends nor posttrends. The connected funds show no pretrends and a positive estimate on the event day. The event-day estimates are far apart: the affiliate funds increase their gross volumes on the event day by -0.02 standard deviation (std. error: 0.04 sd), whereas the connected funds increase theirs by 1.9 sd (std. error: 0.007 sd).

We interpret Figure 5a as follows. The exceptionally large trades pinpoint the arrivals of especially valuable MPI at the dealers, and receiving MPI would prompt increases in trading activity. The null posttrend of the affiliate funds implies that the dealers do not share the especially valuable MPI with their affiliate funds. The positive posttrend of the connected funds means that the dealers obtain the MPI on the days that their connected



(a) Unrelated Pairs



(b) Affiliate and Connected Pairs using Unrelated Pairs as Controls

Figure 4: Correlations in Daily Gross Volumes within Dealer-Fund Pairs

funds exhibit heightened trading activity. We partition how this coincidence of MPI and increase in gross volume could arise into the four channels other than bilateral MPI sharing. The MPI may induce the dealers to trade with the connected funds, in which case the coincidence would be mechanical. An aggregate shock affecting all firms could simultaneously cause both the event trade and the increase in gross volume. The MPI, the event trade, and related trading or information sharing by the event dealers may indirectly affect the connected funds as the dealers' actions percolate throughout the market. There may be common shocks specific to the connected dealers and funds, perhaps because they tend to share sources of information or common thirdparty connections.

The mechanical channel is ruled out by the exclusion of funds that traded with the event dealer on or after the event day for each event. The aggregate-shock and the indirect-effect channels are stripped away by the unrelated fund control group, since the unrelated funds would be exposed to the aggregate shocks and the indirect effects of the dealers' actions. This control would preserve any increase in gross volume due to bilateral MPI sharing, because the dealers would not share MPI with the unrelated funds. The common-shocks channel is rejected by the parallel pretrend, as the shocks common to the connected dealers and funds would sometimes cause the connected funds' gross volumes to increase before the event dealers make their exceptionally large trades. Altogether, only the bilateral sharing channel remains. We conclude that the dealers do not bilaterally share MPI to their affiliate funds.

Figure 5b presents the coefficient estimates of Equations (1) and (2) where we examine the standardized daily gross volumes of the dealers around the days when a fund makes an exceptionally large trade. In blue are the differences α_τ in the volumes between the affiliate and the unrelated dealers around the event days. In red are the differences β_τ

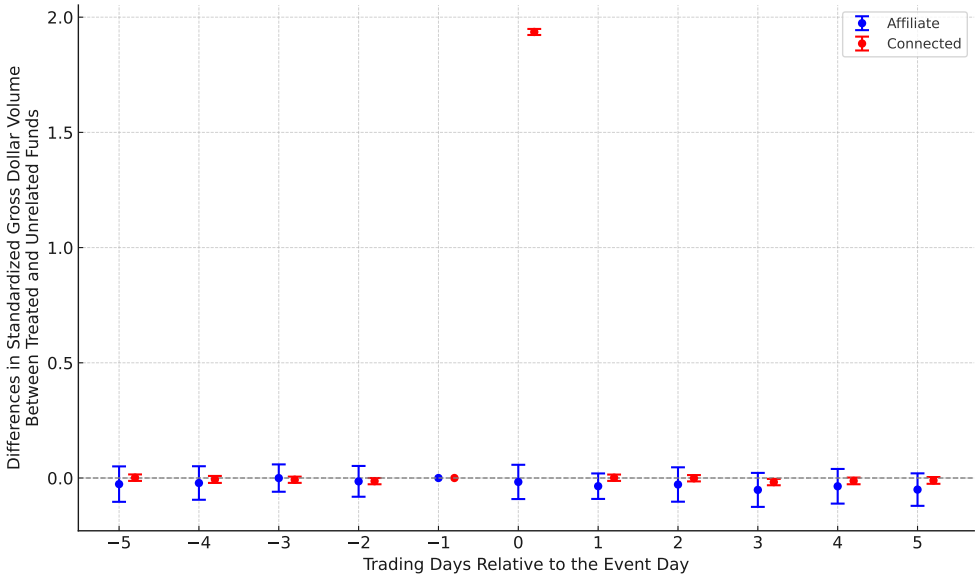
between the connected and the unrelated dealers. Neither the affiliate nor the connected dealers exhibit pretrends. The affiliate dealers do not show posttrends, and the estimated increase in their gross volumes is precisely nil. The connected dealers increase their gross volumes by 0.26 sd (std. error: 0.02 sd) on the event day. We conclude the funds do not bilaterally share MPI to their affiliate dealers.

Based on the results of Figures 5 and 7, we conclude that the China Walls are effective on the whole. Table 3 details the pooled regression counterparts to Figures 5 and 7. The affiliate funds have precisely null response to the arrival of especially valuable information at the dealers and the converse for the affiliate dealers to the funds.[11] In contrast, the connected dealers and funds respond strongly to each other's information, with estimated coefficients in the multiples of the affiliate coefficients. By far the most responsive are the funds to the information from their affiliate funds. Altogether, the pooled results confirm that the China Walls are effective enforced.

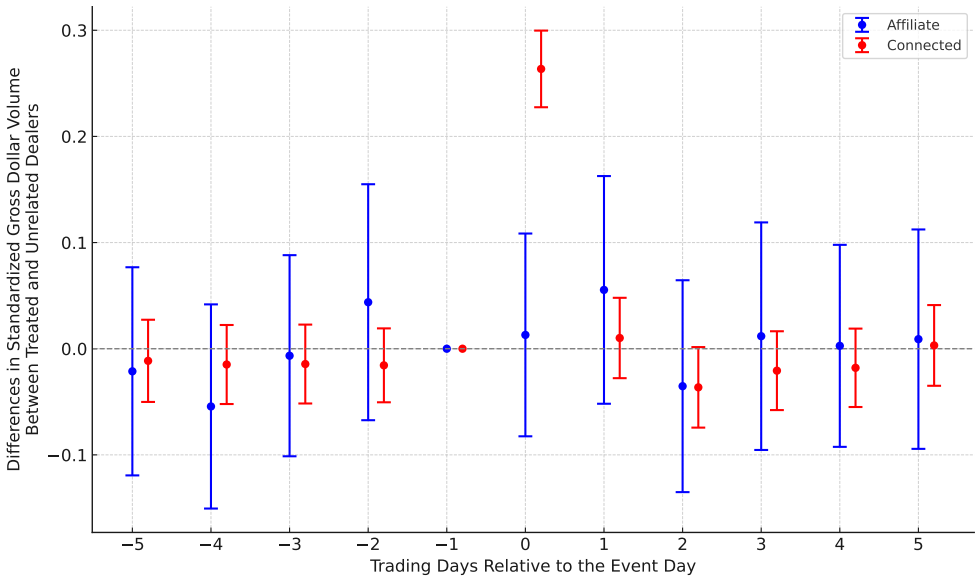
5 Do Affiliates Without China Walls Share Information?

We exploit that each banking conglomerate owns multiple funds to infer whether the affiliate dealers and funds would share MPI if their China Walls were absent. A pair of affiliate funds belong to the same entity, yet are not walled off. Where affiliate funds bilaterally share MPI with each other, we infer that dealers and their affiliate funds would also share MPI in the absence of China Walls.

[11] The fund-to-dealer specification has far fewer events and observations than the dealer-to-fund specification, because there are many more funds than dealers. Each event fund has no more than one affiliate dealer and a few connected dealers, whereas each event dealer has several affiliate funds and numerous connected funds.



(a) Fund Responses to Dealer Information



(b) Dealer Responses to Fund Information

Figure 5: Coefficient Estimates from Equations (1) and (2)

Table 3: Responses in Daily Volumes by Firms on and after the Event Day

	D2F Affiliate	F2D Affiliate	D2F Connected	F2D Connected	F2F Affiliate
$Post \times Affiliate$	-0.024	0.012			0.23***
	[0.017]	[0.028]			[0.021]
$Post \times Connected$			0.32***	0.034**	
			[0.0047]	[0.014]	
$Post \times DealerOverlap$					0.01***
					[0.004]
$Post \times Affiliate$					0.25***
$\times DealerOverlap$					[0.026]
Event×Firm FE	Yes	Yes	Yes	Yes	Yes
Calendar Date FE	Yes	Yes	Yes	Yes	Yes
Days-since-Event FE	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.022	0.41	-0.007	0.45	0.068
Within R-squared	0	0.0001	0.0005	0.0001	0.0001
Events	7,710	7,894	7,710	7,894	7,894
Observations	89,005,179	4,156,128	42,150,672	3,614,383	12,664,366

Coefficient estimates from the pooled counterparts to Equations (1), (2) and (4). The dependent variable is the standardized daily gross US dollar volume of a firm winsorized at the top 0.5 percentile. An event is a firm and a day when the firm made a trade in the 0.1 percentile among its trades. Each event window is 11 days around the event day. Affiliate treatment includes firms that belong to the same conglomerate as the event firm. Connected treatment includes firms that trade at least 10 times with the event firm in our sample, and do not trade with the event firm on or after the event day. Affiliate and Connected treatments are mutually exclusive, because no dealer trades with an affiliate fund in our sample. Controls includes firms that are unaffiliated and never trades with the event firm, and are not treated in another event throughout the event window. We include event-by-firm, calendar date, and days-since-event fixed effects. *D2F*: Dealers are the event firms and funds are the treated and the control firms. *F2D*: Funds are the event firms and dealers are the treated and the control firms. *F2F*: All firms are funds. *DealerOverlap* indicates a treated or control fund whose set of connected dealers overlaps with that of the event fund. Standard errors in square brackets are clustered at the event-by-firm and date levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.1 Design and Implementation

Figure 6 depicts the design. Dotted arrows indicate trading relationships. GS Hedge Fund's sole dealer connection is BoA Dealer. GS Mutual Fund and the GS Hedge Fund are affiliate funds whose dealer connections do not overlap. All funds that trade with the MS Dealer are dropped, such as Independent Fund, to remove any confounding variation due to overlapping dealer connections. We compare the daily gross dollar volume of the GS Hedge Fund (the affiliate fund) to the Unrelated Fund around an exceptionally large trade by the GS Mutual Fund (the fund event). We conclude that the enforcement of China Walls are necessary if the daily volumes of the affiliate funds increase relative to the unrelated funds on or after the fund event day.

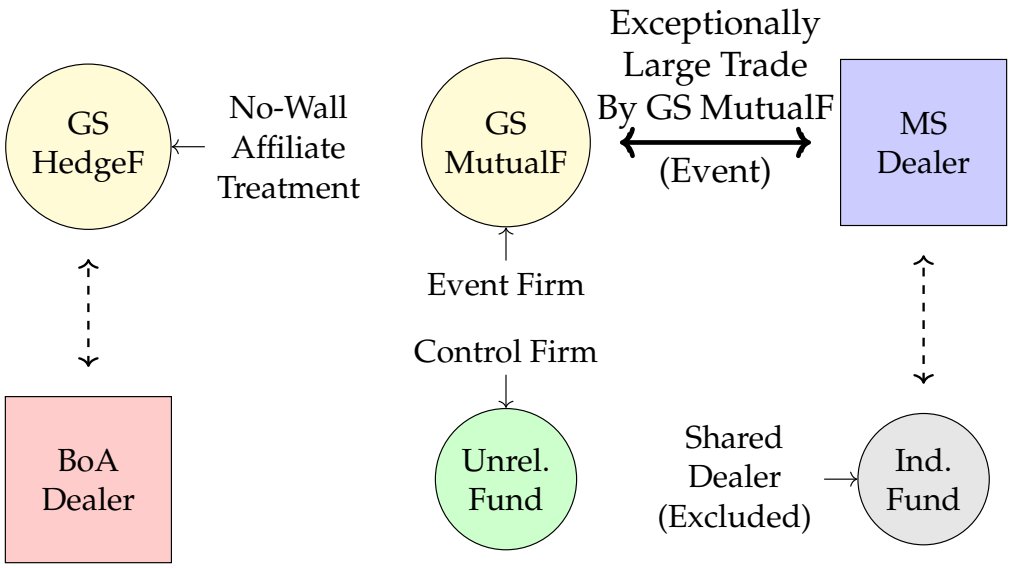


Figure 6: Identification: Information Sharing Between Affiliate Funds

Several conspicuous differences between the affiliate fund pairs and the dealer-fund pairs threaten the validity of this inference. Specifically, a dealer and a fund are likely farther apart in size and in trading strategy than two funds. We partition the affiliate fund

pairs into granular cells of similar or greatly differing sizes and trading strategies to help address this threat to inference. We reject that the China Walls are unnecessary if and only if the gross volumes of the affiliate funds increase relative to the unrelated funds on or after the day when a fund makes an exceptionally large trade consistently across the cells of fund-event fund characteristics. We exclude the affiliate funds that frequently trade with a dealer with whom the event fund also frequently trades. Removing the effects of overlapping dealers this way prevents confounding variation due to common dealer connections, strengthening our inference.

We apply the following specification to implement this design on the subsample of funds:

$$\begin{aligned} Y_{ejt} = & \sum_{\tau=-5}^{5} \nu_\tau \mathbb{1}_{t=\ell_e+\tau} Affiliate_{ej} + \delta_{ej} + \varphi_t + \sum_{\tau=-5}^{5} \gamma_\tau \mathbb{1}_{t=\ell_e+\tau} \\ & + \sum_{\tau=-5}^{5} \kappa_\tau \mathbb{1}_{t=\ell_e+\tau} Affiliate_{ej} DealerOverlap_{ej} \\ & + \sum_{\tau=-5}^{5} \eta_\tau \mathbb{1}_{t=\ell_e+\tau} DealerOverlap_{ej} + \varepsilon_{ejt}. \end{aligned} \tag{4}$$

The control dummy $DealerOverlap_{ej}$ equals 1 if the set of dealers with whom fund j trades at least 10 times in the sample overlaps with the event fund's analogous set of dealers, and equals 0 otherwise. Our focus is on the coefficients ν_τ, which measure the MPI sharing from the event funds to their affiliate funds without an overlapping dealer. Separate event-date effects, γ_τ and η_τ, flexibly control for any trend over event time specific to the funds with or without an overlapping dealer.

5.2 Results

Figure 5 establishes that the affiliate dealers and funds do not share material information. (And that, if they did, our design would reliably detect it.) One interpretation is that the China Walls are effectively enforced. The alternative is that the affiliate dealers and funds would not share MPI even if the China Walls were absent, rendering their enforcement unnecessary. We exploit that affiliate funds are not walled off from one another to infer whether the walled-off affiliates would share information absent the China Walls.

Figure 7 presents the results from Equation (4) estimated on the subsample of funds. In green are the differences ν_τ in standardized gross volume between the affiliate funds and the other funds whose dealer connections do not overlap with the event fund around exceptionally large trades by a fund. Despite removing the common shocks through any overlapping dealers, the affiliate funds increase their gross volumes by a precisely estimated 1.7 standard deviations (std. error: 0.2 sd) on the event date. The large size of this response is consistent with affiliated funds being eager to share information among themselves. In magenta are the differences $\nu_\tau + \kappa_\tau + \eta_\tau$ between the affiliate funds whose dealer connections do overlap with the event fund and the nonaffiliate nonoverlapping funds. As one might expect, incorporating overlapping dealer effects dramatically raises the event date response, to 2.6 sd (std. error: 0.3 sd).

6 Heterogeneity

Our heterogeneity exercises aim to test the robustness of the China Walls. It is particularly important to test the robustness of our affiliate fund-to-fund results: Where even the affiliate funds only share MPI under special contexts, there is high likelihood that the

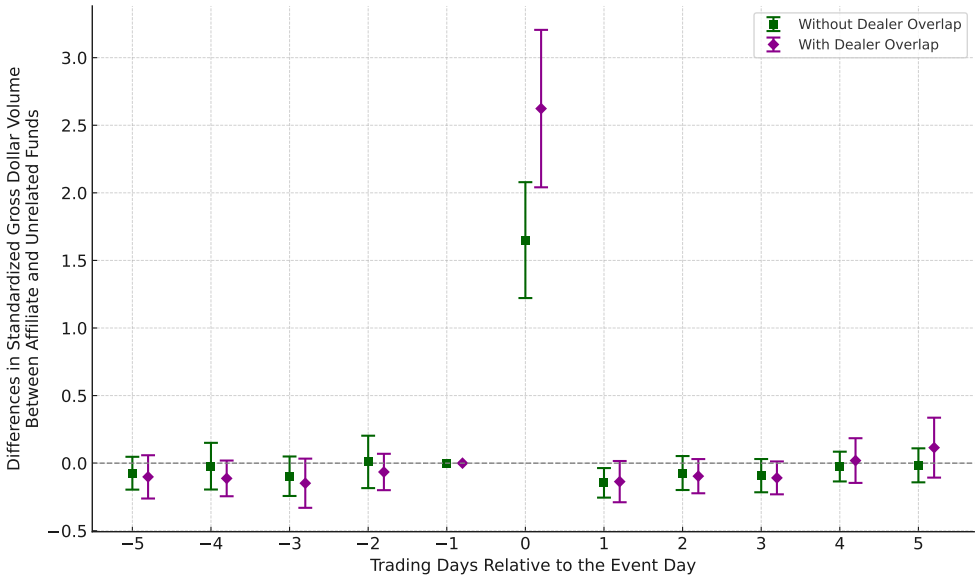


Figure 7: Affiliate Fund Response to Event Fund Information

affiliate dealers and funds would not share MPI absent the China Walls.

To do so, we repeat the analyses of Section 4 across cells of fund types, currency pair, and asset class (i.e., spot, forward, or swap), and for event during crisis and noncrisis periods. Since our main specification yields a null result, we maximize the power to detect deviations from the null by interacting dummy variables corresponding to each characteristic with the complete set of terms in the pooled counterparts of Equations (1), (2) and (4). (Rather than splitting our sample across those characteristics.) We examine both event-level and firm-level characteristics. The dummy $HedgeFund = 1$ if the treated or control firm is a hedge fund, and $HedgeFundEvent = 1$ if the event firm is a hedge fund. Other firm-level dummies indicate whether a firm's share of trades in a currency pair or asset class is greater than the median across firms, separately for dealers and for funds. Event-level dummies indicate whether the event trade was in a given currency pair or asset class, and whether the event occurred during the crisis periods following the

2020 Covid shock, the 2022 Russian invasion of Ukraine, and the 2023 Hamas attack.

6.1 Hedge Funds versus Other Funds

Table 4 separates the responses of hedge funds and nonhedge funds to events by hedge funds and nonhedge funds.

6.2 Crisis Periods

Table 5 compares the coefficient estimates for the events during crisis and noncrisis periods. The crisis periods span Covid (February 1st to March 31, 2020), the Russian Invasion of Ukraine (February 16 to March 8, 2022), and the Hamas Attack (September 27 to October 17, 2023).

6.3 Currency Pairs and Asset Classes

Tables 6 and 7 present the treated firms' responses split by currency pair and asset class. Each table cell is the increase in the treated firms' daily gross volumes relative to the unrelated firms on and after the event day, where the firms' specializations and the events belong to the currency pair or asset class specified for the row.

To arrive at each estimate in Table 6, we first run the pooled counterparts to Equations (1), (2) and (4) augmented with the complete set of interactions involving four dummies: $USDFirm_f = 1$ if the firm's dollar-value share of trades in USD-ILS pair over our sample exceeds the median across firms (i.e., the firm "specializes in USD"), $USDEvent_e = 1$ if the event trade (or any event trade for every event with multiple event trades) was for USD-ILS, and $NonUSDFirm_f$ and $NonUSDEvent_e$ are defined

Table 4: Responses by Fund Type

	D2F Affiliate	F2D Affiliate	D2F Connected	F2D Connected	F2F Affiliate
$Post \times Affiliate$	-0.023	0.026			0.157***
	[0.018]	[0.030]			[0.022]
$Post \times Affiliate$	-0.0068				0.176
$\times$ *Hedge Fund*	[0.045]				[0.32]
$Post \times Affiliate$		0.024			0.047
$\times$ *HF Event Trade*		[0.14]			[0.14]
$Post \times Connected$			0.122***	0.029*	
			[0.0036]	[0.016]	
$Post \times Connected$			0.364***		
$\times$ *Hedge Fund*			[0.012]		
$Post \times Connected$				0.112***	
$\times$ *HF Event Trade*				[0.033]	
$Post \times Affiliate$					2.21***
$\times$ *HF Event Trade*					[0.39]
$\times$ *Hedge Fund*					
Event$\times$Firm FE	Yes	Yes	Yes	Yes	Yes
Calendar Date FE	Yes	Yes	Yes	Yes	Yes
Days-since-Event FE	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.022	0.457	-0.007	0.449	0.049
Within R-squared	0	0.0001	0.0003	0.0002	0.0001
Events	7,710	7,894	7,710	7,894	7,894
Observations	89,005,179	4,156,128	42,150,672	3,614,383	12,664,366

Coefficient estimates from the pooled counterparts to Equations (1), (2) and (4). The dependent variable is the standardized daily gross US dollar volume of a firm winsorized at the top 0.5 percentile. An event is a firm and a day when the firm made a trade in the 0.1 percentile among its trades. Each event window is 11 days around the event day. Affiliate treatment includes firms that belong to the same conglomerate as the event firm. Connected treatment includes firms that trade at least 10 times with the event firm in our sample, and do not trade with the event firm on or after the event day. Affiliate and Connected treatments are mutually exclusive, because no dealer trades with an affiliate fund in our sample. Controls includes firms that are unaffiliated and never trades with the event firm, and are not treated in another event throughout the event window. We include event-by-firm, calendar date, and days-since-event fixed effects. *D2F*: Dealers are the event firms and funds are the treated and the control firms. *F2D*: Funds are the event firms and dealers are the treated and the control firms. *F2F*: All firms are funds. Below: F2F estimates exclude treated and control funds whose dealer connections overlap with the event fund. HF Event Trade is an event whose event trade was by a hedge fund. Standard errors in square brackets are clustered at the event-by-firm and date levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5: Responses to Events During Crisis and Noncrisis Periods

	D2F Affiliate	F2D Affiliate	D2F Connected	F2D Connected	F2F Affiliate
$Post \times Affiliate$	-0.021	0.012			0.226***
	[0.018]	[0.028]			[0.062]
$Post \times Affiliate$	-0.055	-0.032			-0.022
$\times Crisis$	[0.076]	[0.095]			[0.201]
$Post \times Connected$			0.319***	0.033**	
			[0.005]	[0.015]	
$Post \times Connected$			0.012	-0.006	
$\times Crisis$			[0.026]	[0.044]	
$Post \times DealerOverlap$					0.014
					[0.012]
$Post \times Affiliate$					0.264***
$\times DealerOverlap$					[0.073]
Event×Firm FE	Yes	Yes	Yes	Yes	Yes
Calendar Date FE	Yes	Yes	Yes	Yes	Yes
Days-since-Event FE	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.022	0.457	-0.007	0.449	0.033
Within R-squared	0	0.0001	0.0005	0.0002	0.0002
Crisis Events	440	3,303	440	3,303	3,303
Total Events	7,710	7,894	7,710	7,894	7,894
Observations	89,005,179	4,156,128	42,150,672	3,614,383	12,664,366

Coefficient estimates from the pooled counterparts to Equations (1), (2) and (4). The dependent variable is the standardized daily gross US dollar volume of a firm winsorized at the top 0.5 percentile. An event is a firm and a day when the firm made a trade in the 0.1 percentile among its trades. Each event window is 11 days around the event day. Affiliate treatment includes firms that belong to the same conglomerate as the event firm. Connected treatment includes firms that trade at least 10 times with the event firm in our sample, and do not trade with the event firm on or after the event day. Affiliate and Connected treatments are mutually exclusive, because no dealer trades with an affiliate fund in our sample. Controls includes firms that are unaffiliated and never trades with the event firm, and are not treated in another event throughout the event window. We include event-by-firm, calendar date, and days-since-event fixed effects. *D2F*: Dealers are the event firms and funds are the treated and the control firms. *F2D*: Funds are the event firms and dealers are the treated and the control firms. *F2F*: All firms are funds. *DealerOverlap* indicates a treated or control fund whose set of connected dealers overlaps with that of the event fund. *Crisis*: Event occured during the beginnings of Covid pandemic, the Russian Invasion of Ukraine, or the Hamas-Israeli War. Standard errors in square brackets are clustered at the event-by-firm and date levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

analogously. We separately compute the medians for dealers and for funds when assigning $USDFirm_f$ and $NonUSDFirm_f$. We further interact the NonUSD dummies with $MatchingNonUSD_{ef}$, which equals one if $NonUSDFirm_f = NonUSDEvent_e = 1$ and the firm's currency specialization is in the currency of the event trade. Table 6 displays the appropriate sums of the estimated coefficients, and uses the coefficients' covariance matrices to obtain the standard errors of those sums. The estimate in the USD Event Trade-USD Firm by D2F Connected cell, for example, is the sum of coefficients that remain when we set $USDFirm_f = USDEvent_e = 1$ and $NonUSDFirm_f = NonUSDEvent_e = 0$.

Table 7 estimates are calculated in the same way as Table 6, except using dummies corresponding to spot, forward, and swap asset classes, rather than currency pairs.

Table 6: Responses by Currency

	D2F Affiliate	F2D Affiliate	D2F Connected	F2D Connected	F2F Affiliate
USD Event Trade	-0.019	0.021	0.376***	0.034*	0.561***
-USD Firm	[0.035]	[0.073]	[0.006]	[0.019]	[0.219]
USD Event Trade	-0.029	-0.077	0.280***	0.038*	0.194***
-NonUSD Firm	[0.021]	[0.080]	[0.004]	[0.022]	[0.067]
NonUSD Event Trade	0.041	0.122	0.121***	-0.062	0.313***
-USD Firm	[0.033]	[0.108]	[0.007]	[0.052]	[0.097]
NonUSD Event Trade	-0.074	0.029	0.391***	-0.010	0.462***
-NonUSD Firm (Matching Currency)	[0.257]	[0.213]	[0.019]	[0.045]	[0.167]
NonUSD Event Trade	-0.028	-0.033	0.137***	0.007	0.225***
-NonUSD Firm (Nonmatching)	[0.030]	[0.063]	[0.005]	[0.070]	[0.087]
Adjusted R-squared	0.022	0.290	-0.016	0.449	0.033
Within R-squared	0	0.0001	0.0011	0.0002	0.0002
Events	7,710	7,894	7,710	7,894	7,894
Observations	89,005,179	4,156,128	42,150,672	3,614,383	12,664,366

Coefficient estimates from the pooled counterparts to Equations (1), (2) and (4). The dependent variable is the standardized daily gross US dollar volume of a firm winsorized at the top 0.5 percentile. An event is a firm and a day when the firm made a trade in the 0.1 percentile among its trades. Each event window is 11 days around the event day. Affiliate treatment includes firms that belong to the same conglomerate as the event firm. Connected treatment includes firms that trade at least 10 times with the event firm in our sample, and do not trade with the event firm on or after the event day. Affiliate and Connected treatments are mutually exclusive, because no dealer trades with an affiliate fund in our sample. Controls includes firms that are unaffiliated and never trades with the event firm, and are not treated in another event throughout the event window. We include event-by-firm, calendar date, and days-since-event fixed effects. *D2F*: Dealers are the event firms and funds are the treated and the control firms. *F2D*: Funds are the event firms and dealers are the treated and the control firms. *F2F*: All firms are funds. Below: F2F estimates exclude treated and control funds whose dealer connections overlap with the event fund. USD Event Trade is an event whose event trade was a USD trade. USD Firm is a treated or control fund (dealer) whose share of trades by dollar volume involving the USD is above the median across all funds (dealers). Standard errors in square brackets are clustered at the event-by-firm and date levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 7: Responses by Asset Class

	D2F Affiliate	F2D Affiliate	D2F Connected	F2D Connected	F2F Affiliate
Spot Event Trade	0.013	-0.031	0.520***	0.068	0.203***
-Spot Firm	[0.021]	[0.049]	[0.008]	[0.049]	[0.029]
Spot Event Trade	0.044	0.026	0.323***	0.049**	0.250***
-Forward Firm	[0.035]	[0.055]	[0.009]	[0.023]	[0.044]
Spot Event Trade	0.007	0.041	0.318***	0.028	0.174**
-Swap Firm	[0.044]	[0.097]	[0.015]	[0.054]	[0.088]
Forward Event Trade	-0.052	-0.104	0.226***	0.118**	0.248***
-Spot Firm	[0.040]	[0.081]	[0.010]	[0.050]	[0.038]
Forward Event Trade	-0.082	0.004	0.241***	0.099***	0.300***
-Forward Firm	[0.066]	[0.066]	[0.012]	[0.027]	[0.056]
Forward Event Trade	0.026	-0.200	0.134***	0.079	0.301**
-Swap Firm	[0.112]	[0.157]	[0.019]	[0.055]	[0.134]
Swap Event Trade	-0.039	0.149	0.219***	0.019	0.246***
-Spot Firm	[0.026]	[0.147]	[0.005]	[0.054]	[0.073]
Swap Event Trade	-0.029	0.240	0.237***	-0.000	0.224*
-Forward Firm	[0.027]	[0.258]	[0.006]	[0.031]	[0.137]
Swap Event Trade	-0.003	0.280	0.485***	-0.020	0.645*
-Swap Firm	[0.030]	[0.412]	[0.009]	[0.059]	[0.338]
Event×Firm FE	Yes	Yes	Yes	Yes	Yes
Calendar Date FE	Yes	Yes	Yes	Yes	Yes
Days-since-Event FE	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.030	0.489	-0.007	0.449	0.051
Within R-squared	0	0.0001	0.0007	0.0002	0.0002
Events	7,710	7,894	7,710	7,894	7,894
Observations	89,005,179	4,156,128	42,150,672	3,614,383	12,664,366

Coefficient estimates from the pooled counterparts to Equations (1), (2) and (4). The dependent variable is the standardized daily gross US dollar volume of a firm winsorized at the top 0.5 percentile. An event is a firm and a day when the firm made a trade in the 0.1 percentile among its trades. Each event window is 11 days around the event day. Affiliate treatment includes firms that belong to the same conglomerate as the event firm. Connected treatment includes firms that trade at least 10 times with the event firm in our sample, and do not trade with the event firm on or after the event day. Affiliate and Connected treatments are mutually exclusive, because no dealer trades with an affiliate fund in our sample. Controls includes firms that are unaffiliated and never trades with the event firm, and are not treated in another event throughout the event window. We include event-by-firm, calendar date, and days-since-event fixed effects. *D2F*: Dealers are the event firms and funds are the treated and the control firms. *F2D*: Funds are the event firms and dealers are the treated and the control firms. *F2F*: All firms are funds. Below: F2F estimates exclude treated and control funds whose dealer connections overlap with the event fund. Event Trades are separated by asset class. Spot Firm is a treated or control fund (dealer) whose share of trades by dollar value involving spot trades is above the median across all funds (dealers). Forward Firm and Swap Firm are defined analogously. Standard errors in square brackets are clustered at the event-by-firm and date levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix

A Detailed Context

This section provides detailed institutional context with a focus on the US.

A.1 Definitions

A *banking conglomerate* is a group of firms controlled by the same holding company and that includes a depository institution (i.e., a bank). A *financial conglomerate* is a broader term encompassing any such groups that includes firms offering financial services as its primary activity. We write "financial conglomerate" when discussing the period up to the 2000s, when most financial conglomerates became banking conglomerates, and "banking conglomerates" elsewhere.

Figure 8 summarizes the components of a banking conglomerate. Their services includes deposits, lending, insurance, asset management (i.e., investing clients' capital), proprietary trading (investing own capital), brokering (matching client orders) and dealing (absorbing client orders onto inventory), investment analysis and advising, underwriting (asset issuance), corporate advising (on mergers and acquisitions and other strategic decisions), and payments and trade finance. A conglomerate partitions these services into insurers, commercial banks (deposits, loans), investment banks (underwriting, corporate advising), investment funds (asset management), broker-dealers (brokering, dealing, analysis, proprietary trading), and investment advisers.

All regulations against the misuse or leakage of financial information target *material non-public information* (MNPI). Information is MNPI if its public disclosure would ap-

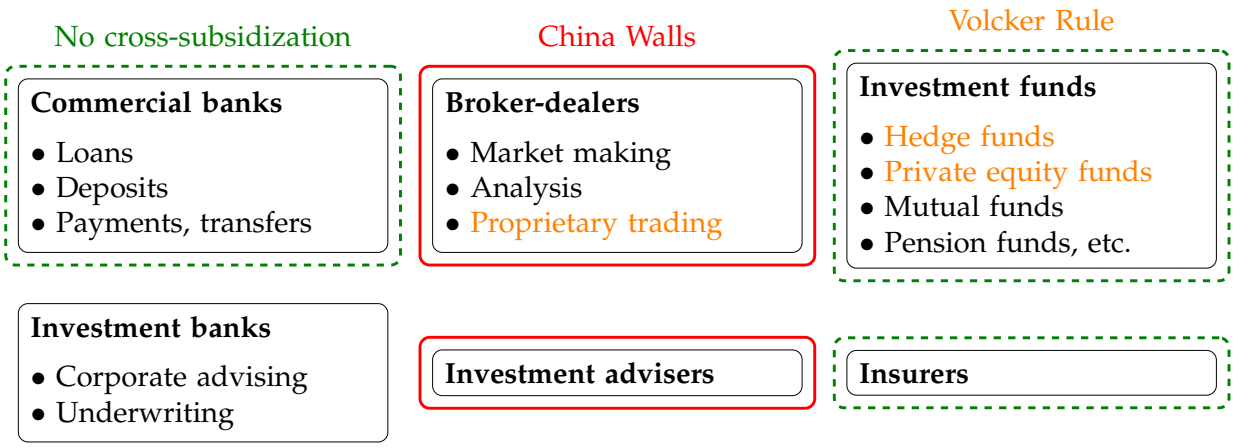


Figure 8: Stylized Banking Conglomerate and Relevant Legal Restrictions

Green dotted lines indicate restrictions on transactions and transfers: Banking laws, fiduciary duty to investors, and state-level insurance laws bar commercial banks, investment funds, and insurers from transferring capital to affiliates or trading with them at unfavorable terms. Red solid lines indicate the China Walls that aim to block the flow of information around subsidiaries in which conflicts of interest concentrate: Broker-dealers and investment advisers are required to prevent their employees interacting with the employees of affiliates. Orange fonts highlight the Volcker Rule restrictions on proprietary trading and ownership of hedge funds and private equity funds by banking conglomerates.

preciably affect market prices. In practice, common-law courts treat as MNPI any non-publicly disclosed information that reasonable investors in the relevant securities would find important for their investment decisions. For example, insider earnings information or outstanding order flows of clients are MNPI.[12] Possessing, sharing, or acting on MNPI is not generically illegal. However, financial intermediaries owe legal duties over MNPIs, as we soon elaborate.

The *China Walls* are blunt internal barriers set around subsidiaries with especially high risk of MNPI misuse. The Walls include both physical barriers and rules, typically:

- Separate offices, elevators, and entry ways for walled-off affiliates, with opaque and soundproof physical barriers when located on the same floor.

[12] Analyses of MNPI are MNPI, whereas analyses of publicly available information are not.

- Cool-down periods for employees transferring between walled-off affiliates.
- Watch lists that prohibit employees from trading or advising on the listed securities.
- Records of every instance where an "over-the-wall" executive (who oversees multiple affiliates walled off from each other) receives MNPI from any subsidiary, and requirement that the executive recuse themselves from any business related to the MNPI.
- Monitor and retain all business-related emails and messages sent by employees, and review those containing MNPI.
- Contingency plans when MNPI leaks through the China Walls, and the appointment of officers responsible for enforcing the Walls and handling the contingencies.

These restrictions on employee interactions effectively ban transactions between walled-off affiliates.

A.2 Key Regulations on Banking Conglomerates

The markings in Figure 8 indicate each key regulation on the banking conglomerates. Two concerns underlie the regulations. First, the conglomerates may divert publicly insured deposits or insurance premiums towards risky trades or to cross-subsidize affiliates, thereby shifting risk onto the state or the insureds. Second, the conflicts of interest inherent in combining intermediation, advisory, and trading functions could disadvantage retail investors and undermine trust in financial markets.

Three constraints on banking conglomerates address these concerns. First, a bank or an insurer cannot cross-subsidize affiliates. The US Regulation W (and similar rules

elsewhere) limit the outstanding value of bank-to-affiliate transactions to 20 percent of the bank's capital and 10 percent with any single affiliate.[13] These trades must occur at prevailing market prices and under punitive collateral requirements. Moreover, banks cannot trade securities issued by its affiliates, accept them as collateral, nor guarantee a trade, loan, or securities issuance that involves an affiliate. Analogous rules on insurers, which are harmonized across the US yet enforced by state authorities, prevent their capital being used to subsidize affiliates (Hamilton, 2011).

Second, the Volcker Rule restricts banking conglomerates from proprietary trading and owning risky investment funds. Specifically, a banking conglomerate cannot use its own capital to make short-term profit-seeking trades. The Rule also limits its ownership stake and exposure to hedge funds and private equity funds. Broad exemptions apply. The Rule exempts the trades linked to market making by broker-dealers and any trade held for more than 60 days. Further, hedge funds and private equity funds active entirely outside the US are exempt and, within the US, a conglomerate may sponsor and control such funds if it holds less than 3 percent of the funds' assets. Therefore, most banking conglomerates contain hedge funds and considerable scope remains for bank-affiliated broker-dealers to trade on private information using own capital.

Third, as we elaborate next, the China Walls around broker-dealers and around investment advisers seek to minimize information leakage surrounding these firms. Statutes single out investment advisers for their large potential impact on investment decisions. The broker-dealers are singled out, because their role as intermediaries provide constant stream of privileged information gleaned from their clients' orders. Under the argument

[13]Outstanding transaction value include loans, face value of guaranteed assets or liabilities, and gross purchases from affiliates. For example, purchasing $1 million of an asset from an affiliate would raise the outstanding value by $1 million until the bank sells $1 million of the same asset back to that affiliate. (Sales to other affiliates or of other assets do not affect the outstanding value generated by this purchase.)

that broker-dealers leaking this information to affiliate funds or receiving inside information from affiliates would place the investing public at a sharp disadvantage, preventing such information flows is necessary to maintain trust and participation in financial markets.

A.3 China Wall Enforcement Over Time

Origins. Under common-law tradition, insider trading on behalf of clients was encouraged. Brokers and dealers were expected to use all information that came into their possession, and further solicit inside information, to fulfill their fiduciary duty. This expectation was upended in 1961, when a landmark judgement held each conglomerate liable for damages incurred by the investing public due to trades based on its MNPI. The ruling demands that the intermediaries holding MNPI either publicly disclose or take no action whatsoever related to the MNPI. Subsequent court rulings placed the full burden of avoiding incompatible duties onto the conglomerates.[14]

Financial conglomerates were in an impossible legal jeopardy. Beyond fiduciary duty and the new duty to the investing public, the agency principle requires the firms acting as agents to safeguard the private information of their principal (Tuch, 2014). Suppose a conglomerate owns a dealer and a mutual fund, and the dealer receives a large trade request from a client hedge fund—an MNPI. By fiduciary duty, the dealer ought to share this MNPI with the mutual fund for the benefit of the fund's investors. Yet, doing so

[14]A typical case is *Black and Shearson v. Hammill Co.* (Black and Shearson, Hammill Co., 1968) which rules, "conflict in duties is the classic problem encountered by one who serves two masters. It should not be resolved by weighing the conflicting duties; it should be avoided in advance [..] or terminated when it appears." The judgement upheld awards of $25 thousand (1968 dollars) each to two customers of a dealer, which sold debentures of a failing firm whose board included a partner at the dealer's parent company. The conflicting duties were the dealer's fiduciary duty to its customers and the partner's duty to keep the inside information of the failing firm confidential.

would expose the conglomerate to liability if the mutual fund trading on the MNPI cause losses to some traders. This liability can be avoided only by publicly disclosing the hedge fund's trade request, in violation of the agency principle. These incompatible duties left financial conglomerates in near-permanent state of legal liability.

The China Walls provided a way out. In 1968, the US Securities and Exchange Commission (SEC) began offering safe harbor from liability for the conglomerates that implement sufficiently strict China Walls, as determined by the SEC.[15] The logic is that walled-off subsidiaries can be considered separate entities for the purpose of determining whether a legal duty has been breached. Continuing the example, the dealer would not owe fiduciary duty to the investors of the affiliate mutual fund if this fund were walled off from the dealer. The US financial conglomerates widely adopted the China Walls, which became broadly standardized according to SEC guidelines. Financial conglomerates in other jurisdictions followed, whether through their US operations or regulatory standardization (in Australia, Canada, France, Germany, Japan, Switzerland, and the UK).

Pre-2008 crisis legal status. A 1980 US Supreme Court case (*Chiarella v. United States*) replaced the constellation of duties with one overarching duty to "disclose or abstain." A person has the duty to disclose or abstain from acting on an MNPI when: (a) she owes fiduciary duty to the source of the MNPI; and (b) the action would give her a personal benefit.

The 1980s also saw the deregulation of financial conglomeration in the US and the UK. The arguments were that full-service financial conglomerates would generate economies of scope and be more competitive versus less regulated foreign competitors. Because the duty to disclose or abstain might render full-service conglomerates nonviable, new

[15] Alternative means to avoid incompatible-duty liabilities, such as obtaining client consent to waive fiduciary duties, are likely ineffective under most circumstances (Tuch, 2014).

statutes explicitly incorporated the China Walls as safe harbor and broadened their legal protections (Brooke, Burrows, Faber, Harpum, and Silber, 1995, p. 98).[16] Suppose a fund consistently earns large profits whenever an affiliate dealer receives large order flows. Under the new statutes, presence of a China Wall between the dealer and the fund would protect the conglomerate against liabilities to the dealer's clients and to the fund's counterparties.[17]

Pre-2008 crisis regulatory regime. The China Walls were initially an legal benefit available to the banking conglomerates—not a regulatory requirement. As such, the China Walls enforcement was purely reactive, occuring in the course of assigning liability upon the discovery of fraud or breach of duty. Indeed, no US regulator proactively evaluated the China Walls between 1990 and 2012, the years when the SEC reviewed the Walls within broker-dealers as a research exercise.[18] The prosecutions over the LIBOR scandal highlights the nonobligatory status of China Walls precrisis: While each settlement with an implicated banking conglomerate often delves into its China Walls, the sole purpose of doing so were to determine the degree of the conglomerate's legal liability for fraud and insider trading. Lacking sufficient China Walls was not an offence in itself.

Further, financial regulators had more limited enforcement powers. Imposition of large penalties or punishment of individuals required court judgement, with 5-year

[16]The UK removed most restrictions on financial conglomeration in 1986. The US gradually weakened the Glass-Steagall Act provisions throughout the 1980s and 90s, until largely repealing the Act in 1999. The UK Financial Services Act 1986 (FSA) and the US Insider Trading and Securities Fraud Enforcement Act 1988 (ITSFEA) explicitly provide safe harbor from a wide range of liabilities to the financial conglomerates that adopt China Walls.

[17]The China Walls grant similar protection elsewhere. For instance, in a landmark Australian case, *ASIC v. Citigroup (2007)*, Citigroup's trading arm purchased one million shares of a target firm one day before its acquisition announcement, in a deal where Citigroup's investment bank was advising the acquirer. The judge dismissed the case, on the basis that the China Wall between Citigroup's trading and investment bank arms was sufficient to preclude conflict of interest (Hanrahan, 2007).

[18]The 1990 review was in response to the 1998 ITSFEA Act that explicitly gave safe harbor to walled-off broker-dealers. The 2012 review was in response to the Dodd-Frank Act.

statute of limitations. A firm that aided a violator could only be prosecuted if the firm knowingly assisted in the violation, a high legal bar. Most importantly, regulatory action required the evidence of actual fraud or breach of duty. Engaging in transactions with a high risk of fraud or duty breach, or failing to maintain China Walls that could greatly suppress the misuse of MNPI were not themselves actionable by regulators.

Current Regulatory Regime. The US Dodd-Frank Act 2010, and partly coordinated laws elsewhere, dramatically reshape the enforcement of China Walls today. The key change is the "risk-based" enforcement powers granted to financial regulators. Rather than requiring actual illegality before the regulators can act, Dodd-Frank gave them the ability to prosecute behavior that raises the risk of fraud or duty breaches. Moreover, a regulator can now prescribe corporate organization and internal rules that the regulator believes necessary to cap the risk of illegality to a reasonable level.

Today's China Walls form a heavily enforced risk-based regulatory prescription. The landmark case is the SEC's 2018 settlement with Mizuho Securities in which Mizuho paid $1.25 million partly for failing to maintain information barriers between its broker-dealer and hedge fund trading desks (US Securities and Exchange Commission, 2018). This case began a series of prosecutions by the SEC where the key issue was the effectiveness of the China Walls itself (Barrack, Moskowitz-Hesse, Richards, and Cox, 2020). As an ongoing example, in 2021, the SEC began a proactive sweep of monitoring and retension of business-related communication among employees across all broker-dealers and investment advisors. The first consequent settlement included a $125 million fine on Morgan Stanley for their failure to retain all business-related messages sent by its broker-dealer employees *on their private devices* (US Securities and Exchange Commission, 2021). As of early 2024, over $2 billion in fines have been meted out to dozens of broker-dealers and

investment advisors over similar failures. Similarly, the SEC charged Virtu Financial in 2024 merely for having a database accessible to both broker-dealer and nonbroker-dealer employees—despite producing no evidence that any MNPI was leaked (US Securities and Exchange Commission, 2024). Therefore, following Dodd-Frank, the regulatory regime over China Walls morphed from reactive to proactive.

B Placebo Results

Two exercises jointly test two identifying assumptions that: (a) Exceptionally large trades pinpoint the arrivals of especially valuable MPI; and (b) our design yields a significant and positive coefficient at $t = 0$ if and only if the event firms bilaterally share MPI with the treated firms.

The first exercise is to compute the price impacts of exceptionally large, median (50 to 50.1st percentile among the event firm's trades by dollar value), and exceptionally small (99.9 to 100th percentile) trades. We do not observe who initiated each trade. Instead, under the intuition that net volumes determine prices (Kyle, 1985), we net all trades in the given percentile in each day separately for funds and dealers. To sign each event trade, we assume, for each dealer-fund trade, that the fund was the initiator. We assume that the event dealer was the initiator for each interdealer trade.

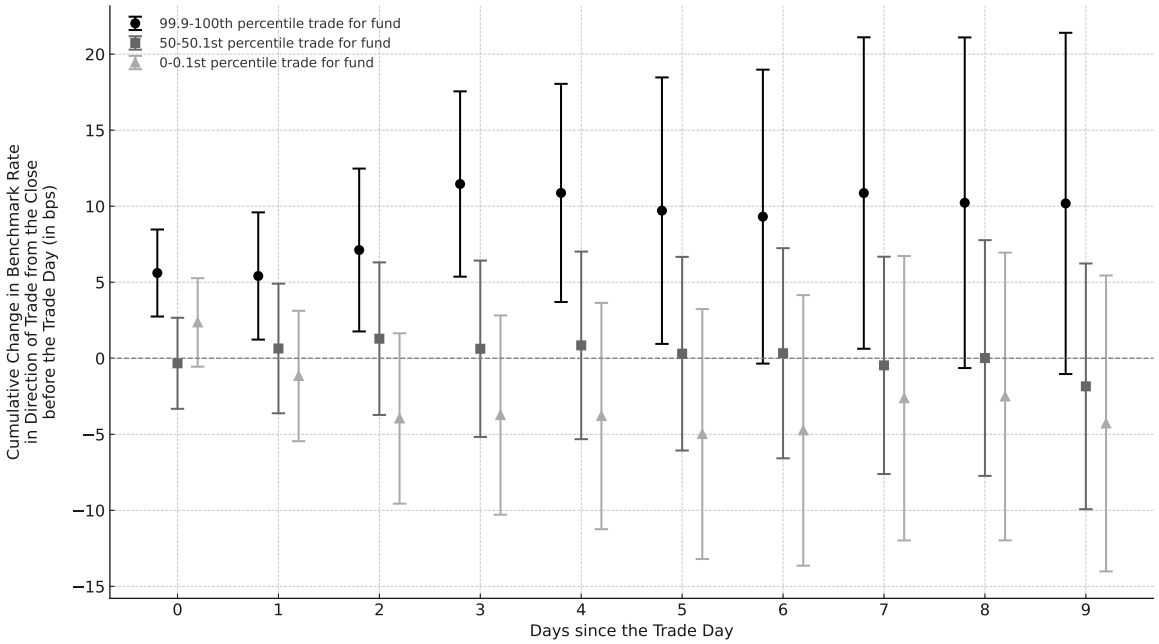
A three-step procedure obtains the price impact of firm type k, firm's trade-size percentile p, and cumulative return horizon ℓ. First, we convert the net dollar volumes on day t into trade-direction dummies $d_{t,k,p} \in \{-1, 0, 1\}$, for $k \in \{\text{fund}, \text{dealer}\}$ and percentile $p \in \{[0, 0.1], [50, 50.1], [99.9, 100]\}$. The dummy $d_{t,k,p} = -1$ if the day's net volume is negative, $d_{t,k,p} = 1$ if its positive, and zero otherwise. Second, we calculate the cu-

mulative returns $R_{t,t+\ell}$ between t and $t+\ell$, $\ell \in \{0,\ldots,9\}$, using Bloomberg benchmark exchange rates. Third, the price impact is the coefficient $\rho_{k,p,\ell}$ in the time-series regression (5):

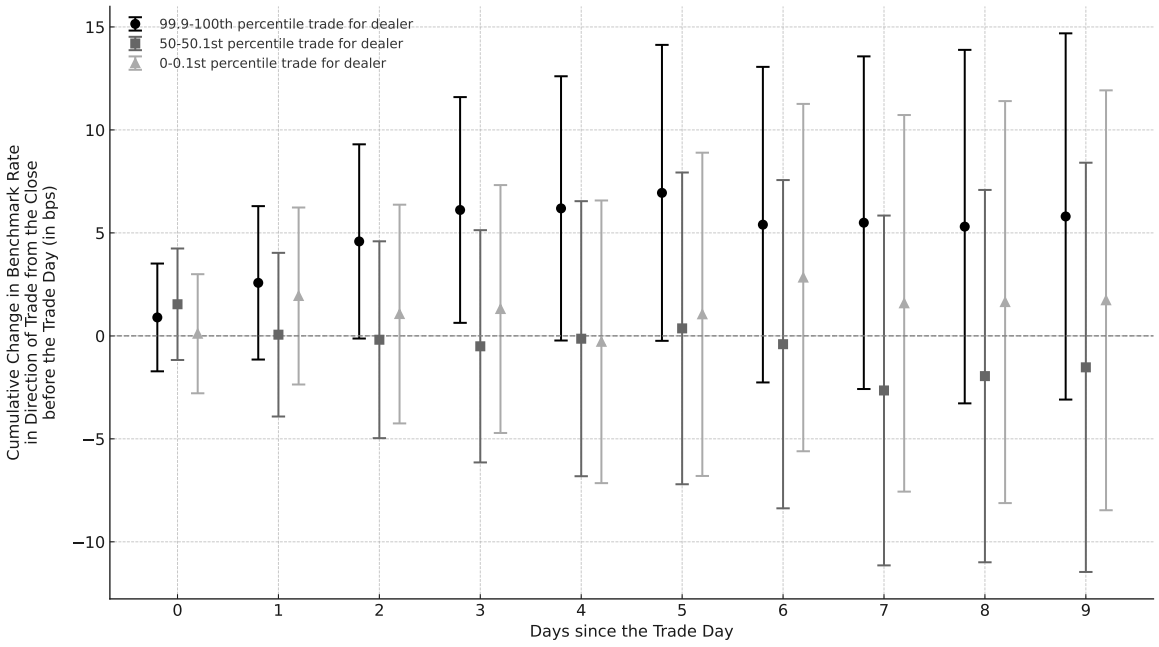
$$R_{t,t+\ell} = \alpha_{k,p,\ell} + \rho_{k,p,\ell} \cdot d_{t,k,p} + \varepsilon_{t,k,p,\ell}. \tag{5}$$

Figure 9 plots the price impact estimates. The net volumes from exceptionally large trades predict future returns, whereas the median and the small trades do not.

The second exercise replicates Figures 5 and 7, except redefining an event to be a day when a firm makes a median or a small trade. As Figure 10 depicts, across all specifications, every coefficient estimate is insignificant at the 95% confidence level. Combined with Figures 5 and 9, these results show that the daily gross volumes of connected firms and non-walled-off affiliate funds increase only in response to the trades that are predictive of returns. We conclude that the exceptionally large trades pinpoint the arrivals of valuable MPI, and that the bilateral sharing of the valuable MPI drives our results.

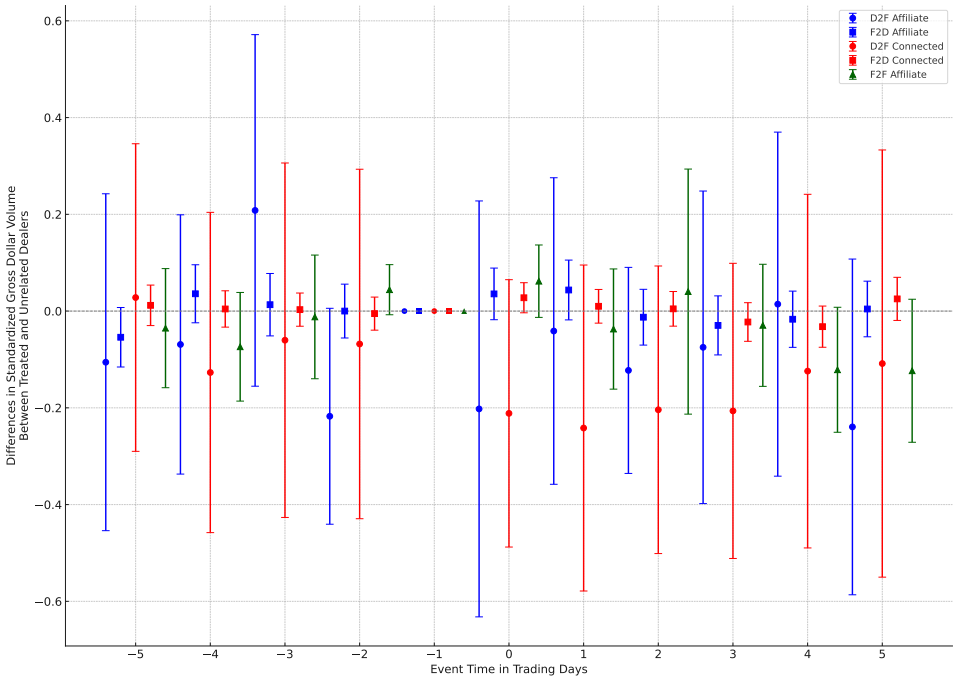


(a) Trades by Funds

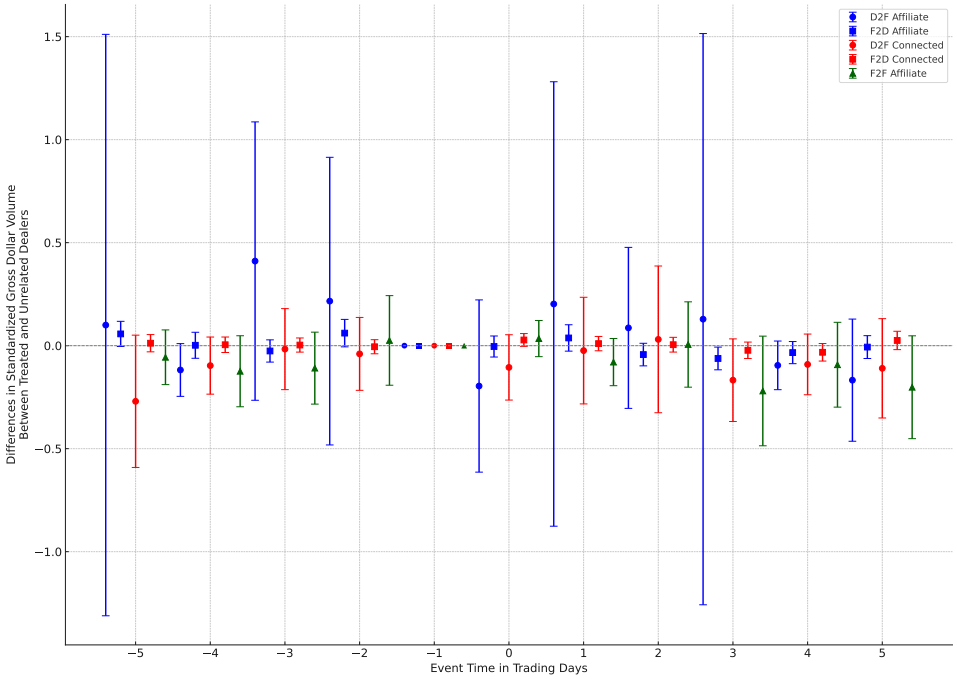


(b) Trades by Dealers

Figure 9: Price Impact Estimates



(a) Event Trade in 50 to 50.1st Percentile



(b) Event Trade in 99.9 to 100th Percentile

Figure 10: Placebo Estimates

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023, February). When Should You Adjust Standard Errors for Clustering? *Quarterly Journal of Economics 138*(1), 1–35.

Bailey, M. J., T. Helgerman, and B. A. Stuart (2024, August). How the 1963 Equal Pay Act and 1964 Civil Rights Act Shaped the Gender Gap in Pay. *Quarterly Journal of Economics 139*(3), 1827–1878.

Bank for International Settlements (2022, October). OTC foreign exchange turnover in April 2022. Technical report, Bank for International Settlements, Basel, Switzerland.

Barbon, A., M. Di Maggio, F. Franzoni, and A. Landier (2019). Brokers and Order Flow Leakage: Evidence from Fire Sales. *Journal of Finance 74*(6), 2707–2749.

Barrack, C., M. Moskowitz-Hesse, L. Richards, and P. Cox (2020, March). Protecting Firm and Client Information: MNPI and Client Confidentiality. In *Securities Industry and Financial Markets Association*.

Behrer, A. P., E. L. Glaeser, G. A. M. Ponzetto, and A. Shleifer (2021, April). Securing Property Rights. *Journal of Political Economy 129*(4), 1157–1192.

Bernhardt, D. and E. Hughson (1997). Splitting orders. *Review of Financial Studies 10*(1), 69–101.

Bhattacharya, U. and H. Daouk (2002). The World Price of Insider Trading. *Journal of Finance 57*(1), 75–108.

Black and Shearson, Hammill Co. (1968, October). Black v. Shearson, Hammill Co.

Boyarchenko, N., D. O. Lucca, and L. Veldkamp (2021, February). Taking Orders and Taking Notes: Dealer Information Sharing in Treasury Auctions. *Journal of Political Economy 129*(2), 607–645.

Brooke, J., A. Burrows, D. Faber, C. Harpum, and S. Silber (1995, December). Fiduciary Duties and Regulatory Rules. Technical Report LAW COM No. 236, The Law Commission, London.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019, August). The Effect of Minimum Wages on Low-Wage Jobs. *Quarterly Journal of Economics 134*(3), 1405–1454.

Chague, F., B. Giovannetti, and B. Herskovic (2023, February). Information Leakage from Short Sellers.

Chen, T. and X. Martin (2011). Do Bank-Affiliated Analysts Benefit from Lending Relationships? *Journal of Accounting Research 49*(3), 633–675.

Cook, L. D., M. E. C. Jones, T. D. Logan, and D. Rosé (2023, February). The Evolution of Access to Public Accommodations in the United States. *Quarterly Journal of Economics 138*(1), 37–102.

Di Maggio, M., F. Franzoni, A. Kermani, and C. Sommavilla (2019, November). The relevance of broker networks for information diffusion in the stock market. *Journal of Financial Economics 134*(2), 419–446.

Easley, D. and M. O'Hara (1987, September). Price, trade size, and information in securities markets. *Journal of Financial Economics 19*(1), 69–90.

Gardner, J. (2022, July). Two-stage differences in differences.

Glaeser, E. L. and A. Shleifer (2003, June). The Rise of the Regulatory State. *Journal of Economic Literature 41*(2), 401–425.

Gozzi, R. (2003). The Chinese Wall Metaphor. *ETC: A Review of General Semantics 60*(2), 171–174.

Greenstone, M., P. Oyer, and A. Vissing-Jorgensen (2006). Mandated Disclosure, Stock Returns, and the 1964 Securities Acts Amendments. *Quarterly Journal of Economics 121*(2), 399–460.

Hagströmer, B. and A. J. Menkveld (2019). Information Revelation in Decentralized Markets. *Journal of Finance 74*(6), 2751–2787.

Hamilton, L. (2011, February). US - NAIC Adopts Modified Insurance Holding Company System model Act and regulation. *Global Corporate Insurance & Regulatory Bulletin*.

Hanrahan, P. F. (2007, December). *ASIC v Citigroup: Investment Banks, Conflicts of Interest, and Chinese Walls*, pp. 117–142. London: IMPERIAL COLLEGE PRESS.

Haselmann, R. F. H., C. Leuz, and S. Schreiber (2023, March). Know Your Customer: Informed Trading by Banks.

Hortaçsu, A. and J. Kastl (2012). Valuing Dealers' Informational Advantage: A Study of Canadian Treasury Auctions. *Econometrica 80*(6), 2511–2542.

Irvine, P., M. Lipson, and A. Puckett (2007, May). Tipping. *Review of Financial Studies 20*(3), 741–768.

Ivashina, V. and Z. Sun (2011, May). Institutional stock trading on loan market information. *Journal of Financial Economics 100*(2), 284–303.

Keiser, D. A. and J. S. Shapiro (2019, February). Consequences of the Clean Water Act and the Demand for Water Quality. *Quarterly Journal of Economics 134*(1), 349–396.

Kondor, P. and G. Pintér (2022). Clients' Connections: Measuring the Role of Private Information in Decentralized Markets. *Journal of Finance 77*(1), 505–544.

Kumar, N., K. Mullally, S. Ray, and Y. Tang (2020, August). Prime (information) brokerage. *Journal of Financial Economics 137*(2), 371–391.

Kyle, A. S. (1985). Continuous Auctions and Insider Trading. *Econometrica 53*(6), 1315–1335.

La Porta, R., F. Lopez-De-Silanes, and A. Shleifer (2006). What Works in Securities Laws? *Journal of Finance 61*(1), 1–32.

Lehar, A. and O. Randl (2006, January). Chinese Walls in German Banks. *Review of Finance 10*(2), 301–320.

Li, F. W., A. Mukherjee, and R. Sen (2021, September). Inside brokers. *Journal of Financial Economics 141*(3), 1096–1118.

Li, T. (2018, June). Outsourcing Corporate Governance: Conflicts of Interest Within the Proxy Advisory Industry. *Management Science 64*(6), 2951–2971.

Massa, M. and Z. Rehman (2008, August). Information flows within financial conglomerates: Evidence from the banks–mutual funds relation. *Journal of Financial Economics 89*(2), 288–306.

Menkhoff, L., L. Sarno, M. Schmeling, and A. Schrimpf (2016, April). Information Flows in Foreign Exchange Markets: Dissecting Customer Currency Trades. *Journal of Finance 71*(2), 601–634.

Peluso, D. (2020, December). Turning a blind eye: The complicit trespassing of 'Chinese walls' in financial institutions in New York. *Critique of Anthropology 40*(4), 438–454.

Pinter, G., C. Wang, and J. Zou (2024, July). Size Discount and Size Penalty: Trading Costs in Bond Markets. *Review of Financial Studies 37*(7), 2156–2190.

Roth, J., P. H. C. Sant'Anna, A. Bilinski, and J. Poe (2023, August). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics 235*(2), 2218–2244.

Seyhun, H. N. (2007/2008). Insider Trading and the Effectiveness of Chinese Walls in Securities Firms. *Journal of Law, Economics & Policy 4*(2), 369–408.

Somogyi, F. (2022, August). Dollar Dominance in FX Trading.

Tuch, A. F. (2014). Financial Conglomerates and Information Barriers. *Journal of Corporation Law 39*(3), 563–616.

US Securities and Exchange Commission (2018, July). In the Matter of Mizuho Securities USA LLC.

US Securities and Exchange Commission (2021, December). In the Matter of J.P. Morgan Securities LLC.

US Securities and Exchange Commission (2024, January). Securities and Exchange Commission v. Virtu Financial Inc. and Virtu Americas LLC.

Cournot Competition, Informational Feedback, and Real Efficiency*

Lin William Cong† Xiaohong Huang‡ Siguang Li§ Jian Ni¶

First version: May 2024; this version: February 2025.

Abstract

We revisit the link between firm competition and real efficiency in a novel setting with informational feedback from financial markets. While intensified competition can decrease market power concentration in production, it reduces the value of proprietary information (on, e.g., market prospects) for speculators and discourages information production and price discovery in financial markets, with non-monotonic welfare effects. Market feedback can impact or even reverse the positive effects of competition on consumer welfare and real efficiency, especially when price becomes sufficiently informative for product decisions. The findings underscore the importance of considering the interaction between product market and financial market in antitrust policy, e.g., concerning the regulation of horizontal mergers. We demonstrate the robustness of the main results under dynamic trading, cross-asset trading and learning, etc.

JEL Classification: D61. D83. G14. G34. L40.

Keywords: Feedback Effects, Information Production, Horizontal Merger, Product Market Competition.

*The authors are especially grateful to Ehsan Azarmsa, Bradyn Breon-Drish, Maryam Farboodi, Yan Xiong, Liyan Yang and Anthony Zhang for helpful discussions and detailed feedback. They also thank Kevin Aretz, Jaden Chen, Liang Dai, Laurent Frésard, Pingyang Gao, Itay Goldstein, Peicong Hu, Yunzhi Hu, Chong Huang, Robert Jarrow, Dan Luo, Fred Sun, Haokun Sun, Jian Sun, Xi Weng, Sang Wu, Xiaoqi Xu, Mao Ye, Zi'ang Yuan, Yao Zeng, Hongda Zhang, Leifu Zhang, Yi Zhang, and participants at the 2024 HKU Accounting Theory Conference, the 37th Australasian Finance & Banking Conference (AFBC), and Wuhan University for constructive feedback. All errors are our own. This article received generous research support from the Guangzhou Municipal Science and Technology Project (2023A03J0691) and the Guangzhou-HKUST(GZ) Joint Funding Program (2024A03J0630). Send correspondence to Li.

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

‡School of Finance, SWUFE. Email: XiaohongHuang0718@163.com

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

¶School of Finance, SWUFE. Email: nijian@swufe.edu.cn

1 Introduction

The interaction and alignment between financial market efficiency and real efficiency constitute a long-standing topic in financial economics, as recently highlighted in studies on feedback effects (Goldstein et al., 2013; Goldstein and Yang, 2019; Goldstein, 2023). Unlike traditional theories on price formation (Grossman and Stiglitz, 1980; Hellwig, 1980; Glosten and Milgrom, 1985; Kyle, 1985), here the information flow is bi-directional: stock prices not only aggregate information from firms, but also contain new information effectively aggregated from traders, which real decision makers (e.g., managers) learn about and use to improve the efficacy of their decisions (e.g., investments and productions).

Against such a backdrop, we revisit the link between firm competition and real efficiency in the presence of stock market feedback. We show that the interaction between the financial market and the product market can undermine the positive effects of competition on real efficiency, contrary to conventional wisdom. Through a parsimonious model in which firm productions are endogenous to stock trading because of the informational feedback from stock prices, we provide new insights into competition and antitrust regulation.

Specifically, we consider a group of identical firms, each supervised by a manager, competing in a standard Cournot setting. The production decision of each firm depends on the assessment of uncertain market prospects, which managers can learn from stock prices. Meanwhile, stock prices aggregate the costly private information acquired by speculators who are incentivized by potential trading profits in financial markets. Firm managers then use the information extracted from stock prices to guide production decisions, which in turn affects firm valuation. The reliance of production decisions on stock prices establishes the feedback effect of the financial market on the real economy.

It is well known that firm competition increases total welfare by reducing market power concentration when firms engage in Cournot competition, which justifies the validity of antitrust regulations related to M&As, for example. However, when these firms are publicly traded, a countervailing force arises: intensified competition can reduce the information content of stock prices and decrease real efficiency. Therefore, intensified competition could generate a loss in total welfare rather than gains. Intuitively, with informational feedback, intensified competition generates both direct and indirect effects on total welfare. The direct effect entails the welfare gain as competition intensifies, reminiscent of that in conventional Cournot competition; the indirect effect comes from managerial learning from stock prices

that aggregate individual speculators' information. Because intensified competition generally curbs the incentive for speculators to produce information, this translates into reduced information acquisition and incorporation into real decisions. A negative relationship between product market competition and total welfare ensues when the indirect effect is dominant.

The key mechanism behind the potential negative relationship between competition and welfare stems from feedback effects that influence the allocative efficiency of resources in production in uncertain environments. Managers set the capacity based on their estimation of future market prospects, relying on information learned from the stock market. In cases of managerial underestimation of market prospects, weaker competition enhances the informativeness of stock prices, correcting managers' downward biases, boosting production, and eventually improving resource allocation. Welfare increases if this production boost outweighs reduced total output caused by market power concentration. In contrast, when managers overestimate market prospects, reduced competition similarly improves information quality but corrects upward biases. This leads to reduced production and amplifies allocative efficiency losses, thus intensifying the negative welfare impact of market concentration.

Note that the negative link between competition and welfare depends on the relative gap in information production, rather than the absolute intensity, as competition intensifies. For example, when the information acquisition cost is high or low, information production either ceases or is in full scale, leading to a minimal change in information production when competition intensifies. Therefore, the market concentration channel dominates and thus competition always improves total welfare. In contrast, for an intermediate level of information cost, welfare-reducing competition always arises in the sense that any market structure with the total number of competing firms exceeding an exogenous threshold becomes suboptimal due to welfare loss related to deteriorated managerial learning alone.

We identify product profitability and market uncertainty as two key determinants of the relative strength of the aforementioned competing forces. Both factors can contribute to the direct effect of product market competition, although the positive effect of market uncertainty is more nuanced. With fixed information production for each stock, an increase in the number of stocks reduces the probability that all order flows are uninformative. However, intensified competition decreases information production, which indirectly leads to a large loss of welfare when amplified by the uncertainty of market prospects. Thus, one would expect the indirect effect to be dominant with low product profitability and high market

uncertainty.

We extend the discussion in several important directions. First, we consider horizontal mergers by comparing the total welfare of a monopoly with that of a duopoly. Interestingly, a monopoly can dominate a duopoly in total welfare for an intermediate level of information production cost. When information production is too cheap or too costly, there is a small gap in the amount of information produced, and thus a monopoly is unlikely to be dominant.

Second, we consider cross-asset trading in which some traders with large investment opportunities (L-traders, including hedge funds, as introduced in Goldstein et al., 2014) can trade all stocks and the rest (S-traders such as individuals and some mutual funds) with small investment opportunities can only trade one stock. With cross-asset trading, the expected trading profits of L-traders, as competition intensifies, will first increase and then decrease, exhibiting an inverted U-shape pattern. Thus, the incentive for L-traders to acquire information will reach its maximum for a moderate level of competition. This differs sharply from S-traders, for whom the incentive of information production is always maximized in a monopoly. However, a negative relationship between competition and total welfare can still arise with L-traders, since the incentive of information production for L-traders will drop quickly after achieving its maximum level.

Third, we consider cross-asset learning in which market makers can observe the order flows of all stocks, rather than a single stock. This gives market makers more information advantages, reducing trading profits for both the S-traders and the L-traders. Actually, this makes S-traders more prone to competition compared to L-traders. Meanwhile, S-traders have a weaker incentive to acquire information compared to L-traders, implying that L-traders may "crowd out" S-traders due to cross-asset trading opportunities/abilities. Interestingly, we find that a negative relationship between product competition and total welfare can arise when S-traders are not fully crowded out by L-traders, which is more likely to occur if the cost of information production is relatively small.

Cochrane (2011) argues that discount rates mainly drive stock price movements instead of cash flows. We therefore also consider discount rates, and follow Dou et al. (2021) to assume that discount rates rise with competition. This further discourages speculators from acquiring information, exacerbating the negative effects of competition on information production and welfare.

Finally, we examine the impact of dynamic trading. Multiple trading rounds introduces market manipulation opportunities, especially on small firms (Edmans et al., 2015; Gold-

stein and Guembel, 2008; Banz, 1981; Acharya and Pedersen, 2005; Comerton-Forde and Putniņš, 2014). As competition reduces firm size, manipulation likelihood increases, further suppressing price informativeness and amplifying competition's negative welfare impact.

Our results have immediate implications for antitrust regulations in practice, where efficiency and welfare are the primary considerations. For example, regulators worry that M&A deals may substantially reduce competition and thus welfare costs by giving firms excessive market power to exploit other market participants and consumers (Guesnerie and Hart, 1985; Farrell and Shapiro, 1990; Landes and Posner, 1997). Horizontal mergers between direct competitors is particularly concerning. However, due consideration of the interaction between (financial) market efficiency and real efficiency is missing from existing antitrust rules.[1] The informational feedback from stock prices to real decisions generates a counterintuitive implication: reduced competition can improve social welfare when the feedback effect from the financial market is sufficiently large. Using data from the U.S. market, we illustrate the importance of incorporating feedback effects in assessing the welfare impacts implications of mergers. Overall, these results highlight that feedback effects from the stock market are a critical factor in analyzing the welfare impact of horizontal mergers and the efficiency of market competition. To avoid misinterpreting merger and acquisition outcomes, antitrust regulatory bodies should take into account the interaction between the financial market and the real economy.

Literature. Our study adds to the literature on the feedback effects of financial markets on real efficiency. Early studies include Fishman and Hagerty (1989), Leland (1992), Dow and Gorton (1997), and Subrahmanyam and Titman (1999). As reviewed by Bond et al. (2012), and recently by Goldstein (2023), real decision makers (e.g., firm managers) can collect new information from stock prices to improve investments and production decisions (Foucault and Frésard, 2014; Edmans et al., 2015; Lin et al., 2019; Goldstein et al., 2013; Edmans et al., 2017; Goldstein and Yang, 2019). Central to this strand of literature is the alignment of market efficiency (i.e., the prediction power of stock prices for future cash flows) and real efficiency (i.e., the usefulness of stock prices for investment and production

[1]Section 7 of the Clayton Act, amended by the Celler-Kefauver Act later, prohibits mergers and acquisitions when the effect "may be substantially to lessen competition or to tend to create a monopoly." Consequently, the US Department of Justice (DOJ) and the Federal Trade Commission (FTC) have developed the Horizontal Merger Guidelines, delineating key factors and analytical frameworks, as well as many specific examples of how these principles can be applied in actual merger reviews. See, e.g., `https://www.justice.gov/atr/horizontal-merger-guidelines-0`.

decisions). These two notions of efficiency typically diverge under feedback effects (Dow and Gorton, 1997; Bond et al., 2012). Bai et al. (2016) derive a welfare-based measure of price informativeness and find a revelatory component has contributed significantly to the efficiency of capital allocation since 1960. Goldstein and Yang (2019) reveal a stark difference between market efficiency and real efficiency by considering multiple dimensions of information, generating interesting insights for optimal design of disclosure systems.[2]

Our paper differs by focusing on the welfare implications of intensified competition on real efficiency. In our model, product market competition can increase real efficiency by reducing firms' market power and decrease real efficiency by reducing information production by speculators. The two competing forces of reducing market power concentration and reducing information production jointly determine the impact of product market competition on social welfare.

A closely related study is Xiong and Yang (2021), which emphasizes the strategic information disclosure of firms. Our paper differs from theirs in the following three aspects, including: First, in their model, competition reduces firms' voluntary disclosure, ultimately leading to a decrease in economic efficiency. In contrast, we stress the role of information production by speculators and show that this mechanism alone can generate a negative relationship between competition and total welfare. Second, their analysis mainly compares a monopoly product market with a perfect competition market, whereas we consider any arbitrary number of firms and characterize general conditions under which competition decreases total welfare. Third, speculators no longer exogenously possess private information, but instead endogenously choose whether to become informed in our model.[3] Huang and Xu (2023) also explore the secondary market and product market competition, but focus on how initial stock holdings affect arbitrageurs' buying and thus entry decisions of potential uninformed entrants through feedback effects. More broadly, our paper relates to the aggregate implications of information production (e.g., Han and Yang, 2013). In particular, Angeletos et al. (2023) show that the two-way feedback between startup activity and investors beliefs can generate excessive and non-fundamental influences on firm activities and asset prices.

[2]More literature focusing on optimal disclosures include: Chen et al. (2021); Edmans et al. (2015); Boleslavsky et al. (2017); Gao and Liang (2013) and Jayaraman and Wu (2019).

[3]More precisely, Xiong and Yang (2021) also consider endogenous information acquisition by speculators in their Section 5.3. A key difference is that when the number of firms increases, information acquisition decreases in the extensive margin in our paper, while Xiong and Yang (2021) document a different pattern in which the extensive margin of information acquisition increases while the intensive margin decreases. This further suggests that this insight is robust to different ways of modeling information acquisition.

Our study is also related to the long-standing literature investigating the relationship between competition and economic efficiency and its implications for antitrust regulations. Dating back to Smith (1776) and Cournot (1838), the traditional wisdom — the existence of market power can generate market inefficiencies and reduce welfare by raising price and suppressing output — has greatly influenced the evolution of the Horizontal Merger Guidelines (Nocke and Whinston, 2022).[4] On the one hand, the unilateral effect analysis emphasizes the trade-off between post-merger market power and potential synergies (see, e.g., Williamson, 1968; Farrell and Shapiro, 1990; Nocke and Whinston, 2022).[5] On the other hand, the coordinated effect analysis concerns implicit anti-competitive coordination from mergers in the absence of explicit communication (see, e.g., Compte et al., 2002; Miller and Weinberg, 2017; Porter, 2020). Röller et al. (2001) and Asker and Nocke (2021) offer comprehensive surveys of this vast literature before 2001 and more recent developments, respectively. In addition, Peress (2010) analyzes how product market competition influences stock price informativeness, which in turn affects capital allocation.

We examine not only the potential negative impact of firm competition on price informativeness but also the informational feedback from stock prices to production decisions, with novel welfare and policy implications. In particular, we show that without cost synergies that are commonly assumed in prior studies, informational feedback from stock market alone can affect and even reverse the welfare effects of a horizontal merger. Thus, our analysis reveals the feedback effect to be an important and indispensable factor in analyzing the welfare impact of horizontal mergers and the efficiency of market competition.

Finally, several recent studies explore direct evidence for merger-specific efficiency (Ashenfelter et al., 2015; Braguinsky et al., 2015), and characterize what counts as an efficiency (Hemphill and Rose, 2017; Geurts and Van Biesebroeck, 2019). Covarrubias et al. (2020) identify good and bad concentrations at the aggregate and industry level in the United States over the past three decades. Our paper contributes to the discussion of positive merger-specific efficiencies by exploring a new channel through feedback effects between the product market and the financial market. Two other related papers, Edmans et al. (2012) and Luo (2005), similarly explore the feedback effect in mergers and acquisitions. Both em-

[4]The Horizontal Merger Guidelines feature two key considerations: unilateral price effects and coordinated effects. Other concerns include pro-competitive forces such as market entry and dynamic considerations (see, e.g., Mermelstein et al., 2020; Nocke and Whinston, 2010).

[5]Recently, a growing literature evaluates"merger simulations" to quantify unilateral price effects and welfare impacts (Werden and Froeb, 1994; Weinberg, 2011; Björnerstedt and Verboven, 2016; Nevo, 2000).

phasize how learning by insiders from outsiders' information affects the decision for M&As but do not focus on the link between competition and efficiency as we do.

The remainder of the paper is organized as follows: Section 2 sets up the model. Section 3 characterizes the equilibrium. Section 4 revisits the relationship between production competition and real efficiency in the presence of feedback effects. Section 5 extends the baseline model and discusses the robustness of the main results. Finally, Section 6 concludes. All proofs are relegated to the appendix.

2 Model Setup

We embed feedback from stock prices to product decisions under market competition into an otherwise standard Cournot model. Consider $n \geq 2$ identical firms competing in production quantity, and each firm's equity is traded on a public stock exchange. Time is discrete and indexed by $t \in \{0, 1\}$. At $t = 0$, a group of speculators decide whether to acquire private information on the market prospects of the product and subsequently decide how to trade stocks.[6] Then, the manager of each firm makes a production decision, taking into account the production strategies of other firms and the trading on the stock exchange at $t = 0$. Finally, at $t = 1$, the cash flows for all firms are realized. The key departure from the Cournot model is that managers in our setting can learn and use information contained in stock prices for their production decisions.

The product market. Let q_i denote the output level of the ith firm, where $i \in \{1, \ldots, n\}$.[7] Denote the total supply of the product by $Q = \sum_{i=1}^{n} q_i = q_i + q_{-i}$, where $-i$ denotes all other firms. As in Xiong and Yang (2021), the market clearing price P is given by: $P = A - bQ$. Here, $b > 0$ indicates the sensitivity of demand to price and $A > 0$ captures the possible market prospect of the product. Depending on a relevant economic state $\omega \in \{H, L\}$, the

[6] We follow the literature by assuming that speculators only acquire information once (See, e.g., Gao and Liang, 2013; Goldstein et al., 2014; Dow et al., 2017; Xiong and Yang, 2021). The effects of introducing multiple rounds of trading will be discussed in Section 5.

[7] We focus on Cournot competition (i.e., quantity competition), rather than Bertrand price competition, for the following two reasons. First, in canonical Bertrand competition, the total welfare is independent of the total number of competing firms. Second, as shown in Kreps and Scheinkman (1983), the quantity (capacity) pre-commitment and the Bertrand price competition yield Cournot outcomes. In addition, we anticipate that Bertrand competition can weaken our result even with differentiated products. For example, Vives (1985) shows that prices and profits are generally higher and quantities are lower in Cournot competition than in Bertrand competition. Therefore, Bertrand competition can enhance the effect of market concentration, potentially reducing the relative significance of information feedback.

realization of the market prospect is given by $A(\omega) = A_\omega$, where $A_H > A_L > 0$. Both states are equally likely ex ante, i.e., $\Pr(\omega = H) = \Pr(\omega = L) = 1/2$. Given the production decisions $\{q_i\}_{1 \leq i \leq n}$, the ith firm receives an operating profit given by:

$$TP_i(q_i) = q_i(A - bQ - MC), \tag{1}$$

where MC is a constant marginal production cost. Without loss of generality, we assume that $A_H > A_L \geq MC$. To highlight the core mechanism, we leave out financing constraints.

All firms decide simultaneously on the production level q_i at time $t = 0$. Each firm manager maximizes the expected value of the firm after the stock prices are observed. In other words, conditional on the information observed, $\mathcal{F}_m$, at $t = 0$, the firm manager chooses the output level q_i to maximize:

$$V_i(q_i) = \mathbb{E}[TP_i(q_i) \mid \mathcal{F}_m]. \tag{2}$$

The stock market. All firms are publicly traded by three types of investors: (i) a continuum of risk-neutral speculators who can choose to acquire costly information; (ii) a group of liquidity traders for each firm $i \in \{1, \cdots, n\}$, who jointly submit an aggregate order $z_i \sim U([-1, 1])$, independently and uniformly distributed over $[-1, 1]$ across the identity of the firm i; and (iii) a set of risk neutral market makers. The free entry of market makers implies that each makes zero profit in equilibrium.

For each firm i, let $\alpha_i \in [0, 1]$ denote the size of speculators acquiring costly information at $t = 0$ as in Foucault and Frésard (2014). To endogenously determine the amount α_i of informed speculators, we assume that each speculator k must pay a cost $c > 0$ to become informed, i.e., receiving an informative signal $m_k^i \in \{H, L\}$.[8] With precision $\theta > \frac{1}{2}$, the signal structure is given by:

$$\Pr\left(m_k^i = H | \omega = H\right) = \Pr\left(m_k^i = L | \omega = L\right) = \theta. \tag{3}$$

Conditional on the realization of ω, m_k^i is independently and identically distributed across speculators (as in Goldstein et al., 2013; Dow et al., 2017). Upon observing the signal m_k^i, the kth informed speculator can choose to trade x_k^i shares of the ith firm, where $x_k^i \in [-1, 1]$ as in Dow et al. (2017). Thus, the aggregate demand for the ith stock from speculators is

[8]The superscript "i" in m_k^i is used to indicate that the kth speculator is trading the ith stock.

given by: $x_i = \int_0^{\alpha_i} x_k^i dk$. Recall that all liquidity traders submit an aggregate order z_i that is uniformly distributed. The total order flow f_i for the ith stock is: $f_i = z_i + x_i$.

As in Kyle (1985), the order flow f_i in each stock i is absorbed by market makers, and the stock price s_i reflects the expected value of the firm conditional on the total order flow:

$$s_i(f_i) = \mathbb{E}[V_i \mid f_i]. \tag{4}$$

Equilibrium definition. The equilibrium concept that we use is perfect Bayesian equilibrium, which consists of: (i) a production strategy for each manager that maximizes the expected firm value given the information conveyed in stock prices; (ii) an information production strategy and a trading strategy for speculators that maximize the expected trading profit given all others' strategies; (iii) a price-setting strategy for market makers that allows them to break even in expectation given all others' strategies; (iv) managers and market makers update their beliefs about the economic state according to the Bayes rule; and (v) each player's belief about other players' strategies is correct in equilibrium.

3 Equilibrium Characterization

We solve the model backward. We first derive the equilibrium strategy of firms, taking as a given the amount α_i of informed speculators for each firm i, and then we endogenize α_i. As shown later, an informed speculator k with a private signal m_k^i always buys one share of the stock of the ith firm when $m_k^i = H$, and sells one share when $m_k^i = L$. Given this observation, we can now investigate the production strategies of firms and the pricing rules for stocks in equilibrium.

Let us first consider the limit where the information acquisition cost c is sufficiently high that all speculators abstain from acquiring information. When this occurs, the stock price is uninformative and the market outcome reduces to the standard Cournot competition outcome with n identical firms. Therefore, each firm produces an identical output:

$$q_M = \frac{\bar{A} - MC}{(n+1)b}, \tag{5}$$

where $\bar{A} = \frac{1}{2}(A_H + A_L)$.

This can be compared with the market outcome when the actual market prospect $A(\omega)$

is publicly known to all market participants. Specifically, when $A(\omega) = A_H$, each firm produces a quantity of $q_H = \frac{A_H - MC}{(n+1)b}$, making a profit of $s_H = \frac{(A_H - MC)^2}{(n+1)^2 b}$. Similarly, when $A(\omega) = A_L$, each firm produces $q_L = \frac{A_L - MC}{(n+1)b}$, making a profit of $s_L = \frac{(A_L - MC)^2}{(n+1)^2 b}$. In contrast, in the absence of information produced by speculators, the equilibrium output q_M under uncertainty is just the expectation of outputs in both states, i.e., $q_M = \frac{1}{2}(q_H + q_L)$.

Next, we consider the case of informative stock trading. Intuitively, due to information-based speculative trading, stock prices contain useful information for managers to guide production decisions. Thus, to solve for the production strategy with informational feedback effects, we need to analyze stock pricing rules in equilibrium. Following Kyle (1985), market makers set stock prices based on the updated belief about the value of firms, given the total order flow observed. Given the information structure in Equation (3), by the law of large numbers (Dow et al., 2017), the aggregate order of informed speculators is $x_i = \alpha_i(2\theta - 1)$ when $\omega = H$, generating a total order flow of $f_i = \alpha_i(2\theta - 1) + z_i$. Similarly, if $\omega = L$, then: $f_i = -\alpha_i(2\theta - 1) + z_i$.

In summary, market makers condition the pricing on the observed total order flow, which aggregates the information from the trading activities of informed speculators. Therefore, the stock price contains valuable information for managers, which establishes an information feedback channel to the real economy. As shown in Lemma 1, the optimal production strategies of firms explicitly depend on stock prices.

Lemma 1. *Given the measures of informed speculators* $\{\alpha_i\}_{1 \le i \le n}$*, the equilibrium stock price for the* i*th firm is given by:*

$$s_i(f_i) = \begin{cases} s_H, & \text{if } f_i > \gamma_i \\ s_M^i, & \text{if } -\gamma_i \le f_i \le \gamma_i \\ s_L, & \text{if } f_i < -\gamma_i \end{cases}, \tag{6}$$

where $s_H = \frac{(A_H - MC)^2}{(n+1)^2 b}$, $s_M^i = \frac{1}{4(n+1)^2 b}\left\{2\left((A_H - MC)^2 + (A_L - MC)^2\right) - \beta_i (A_H - A_L)^2\right\}$, $s_L = \frac{(A_L - MC)^2}{(n+1)^2 b}$, $\gamma_i = 1 - \alpha_i(2\theta - 1)$*, and* $\beta_i = \prod_{j \ne i} \gamma_j$.

Furthermore, given all stock prices $\{s_i\}_{1 \le i \le n}$*, the* i*th firm produces an output of:*

$$q_i^* = \begin{cases} q_H, & \text{if } s_j = s_H \text{ for some } j \\ q_M, & \text{if } s_j = s_M^j \text{ for all } j \\ q_L, & \text{if } s_j = s_L \text{ for some } j \end{cases}, \tag{7}$$

where $q_H = \frac{A_H - MC}{(n+1)b}$, $q_L = \frac{A_L - MC}{(n+1)b}$, *and* q_M *is given by Equation (5).*

We make three comments on Lemma 1. First, the three conditions in Equation (6), as well as those in Equation (7), are mutually exclusive, which rules out the possibility of observing both $s_i = s_H$ and $s_j = s_L$ for some $i \neq j$.[9] Thus, the optimal production strategy q_i^* is well defined. Second, we can directly verify that $s_H > s_M^i > s_L$, which implies that the equilibrium stock price s_i increases weakly in the total order flow f_i. This result is consistent with those of the existing literature on feedback effects (Foucault and Frésard, 2014; Dow et al., 2017; Lin et al., 2019). Third, managers choose equilibrium output levels based on observed stock prices. Obviously, $q_H > q_M > q_L$, which implies that q_i^* generally tends to increase with stock prices.

We now proceed to analyze the optimal behavior of speculators in equilibrium. Specifically, we first derive the optimal trading strategy of an informed speculator and then calculate the resulting expected trading profits, which are summarized in Lemma 2 below.

Lemma 2. *For speculators that focus on the* i*th stock, the optimal trading strategy is to long one share (that is,* $x_k^i = +1$*) when* $m_k^i = H$ *and short one share (that is,* $x_k^i = -1$*) when* $m_k^i = L$*. The resulting expected trading profit is:*

$$\Pi_i(\boldsymbol{\alpha}) = \frac{\gamma_i(2\theta - 1)\left(2 + (n-1)\beta_i\right)}{2(n+1)^2 b}\left(\bar{A} - MC\right)\left(A_H - A_L\right).$$

Lemma 2 verifies the intuition that an informed speculator always follows his own signal, i.e., he longs the stock after receiving good news and shorts it after bad news. Also note that $\Pi_i(\boldsymbol{\alpha})$ depends on all $\{\alpha_i\}_{1 \leq i \leq n}$ through γ_i and β_i. Furthermore, the expected trading profit $\Pi_i(\boldsymbol{\alpha})$ strictly increases both in the average profitability, as measured by $\left(\bar{A} - MC\right)$, and in the uncertainty about the market prospects, as measured by $(A_H - A_L)$.

Finally, Lemma 2 is an important intermediate step in understanding the incentive for information production. Specifically, when acquiring costly information on market prospects, an uninformed speculator balances between the cost of information production $c > 0$ and the value of proprietary information $\Pi_i(\boldsymbol{\alpha})$. Since all firms are identical in the Cournot competition, we hereafter focus on the symmetric case $\alpha_i = \alpha$ $(\forall\ 1 \leq i \leq n)$ and define:

$$\Pi(\alpha) := \Pi_i(\boldsymbol{\alpha}) = \frac{\gamma(2\theta - 1)\left(2 + (n-1)\gamma^{n-1}\right)}{2(n+1)^2 b}\left(\bar{A} - MC\right)\left(A_H - A_L\right), \tag{8}$$

[9] To see this, given that $s_i = s_H$, the state consistent with the order flow of noise trading can only admit $\omega = H$, contradicting $s_j = s_L$ which fully reveals that $\omega = L$.

where $\gamma = 1 - \alpha(2\theta - 1)$.

Note that $\Pi(\alpha)$ in Equation (8) strictly decreases in α, i.e., $\frac{\partial \Pi(\alpha)}{\partial \alpha} < 0$. Thus, the value of private information decreases when more agents choose to do so, implying that information acquisition is a strategic substitute among speculators.

Intuitively, when the cost of information acquisition is large enough such that $\Pi(0) \leq c$, no speculator has an incentive to acquire education. However, when the cost parameter is sufficiently small such that $c \leq \Pi(1)$, all speculators choose to acquire information. Together, these two conditions establish two cut-off points, including an upper bound $\overline{c} = \Pi(0)$ and a lower bound $\underline{c} = \Pi(1)$. Specifically, we define:

$$\overline{c}_n = \frac{(2\theta - 1)}{2(n+1)b} \left(\bar{A} - MC\right) (A_H - A_L) \tag{9}$$

and

$$\underline{c}_n = \frac{(2\theta - 1)(1 - \theta) \left(2 + (n-1)(2 - 2\theta)^{n-1}\right)}{(n+1)^2 b} \left(\bar{A} - MC\right) (A_H - A_L) \tag{10}$$

Let $\widehat{\alpha}$ denote the optimal intensity of information acquisition.

Proposition 1 (Information Acquisition Intensity)**.**

(i) When $c \geq \overline{c}_n$, there is a unique symmetric equilibrium with no information production ($\widehat{\alpha} = 0$);

(ii) When $0 \leq c \leq \underline{c}_n$, then $\widehat{\alpha} = 1$ in the unique equilibrium; and

(iii) When $\underline{c}_n < c < \overline{c}_n$, there is a unique interior equilibrium with $\widehat{\alpha} \in (0, 1)$ such that $\Pi(\widehat{\alpha}) = c$.

Two comments are in order. When $\Pi'(\widehat{\alpha}) < 0$, an interior solution $\widehat{\alpha}$ is said to be locally stable because when we start with $\alpha < \widehat{\alpha}$, more speculators find it optimal to acquire information, increasing the intensity of information acquisition and vice versa. Moreover, the incentive to acquire and trade on private information is negatively associated with the cost of information production. Such an equilibrium on information acquisition is reminiscent of that in Grossman and Stiglitz (1980). A sufficiently large cost preempts the incentive to acquire information, and thus the informational feedback effect disappears. In general, the information content of stock prices depends on the amount of informed speculators in the stock market, which is pinned down uniquely by the information cost and other model parameters.

4 Competition and Efficiency Under Feedback Effects

We now establish that product market competition can decrease the incentive for speculators to produce information and then analyze the efficiency implications of firm competition with informational feedback from stock prices. Interestingly, reduced competition in the stock market can enhance informational efficiency, leading to allocative efficiency gains that significantly alter the efficiency implications of product market competition. When the feedback effect is sufficiently strong, Cournot competition may even produce negative welfare effects.

4.1 Information Production

We first analyze how information production, measured by the equilibrium size of informed speculators $\widehat{\alpha}_n := \widehat{\alpha}(n)$, varies with the number of firms n in the product market. For simplicity, we focus on the interior solution case; otherwise, we expect that $\partial\widehat{\alpha}_n/\partial n = 0$ under corner solutions. Then, we rewrite the equilibrium condition as:

$$\Pi(\widehat{\alpha}) = \Pi(n, \widehat{\alpha}_n) = c \tag{11}$$

A direct application of the implicit function theorem implies the following:

Proposition 2 (Competition and Information Production)**.** *When an interior solution $\widehat{\alpha}_n \in (0,1)$ exists $c \in (\underline{c}, \overline{c})$, $\widehat{\alpha}_n$ strictly decreases in n, that is, $\frac{\partial\widehat{\alpha}_n}{\partial n} < 0$.*

Proposition 2 verifies that the amount $\widehat{\alpha}_n$ of informed speculators increases as competition weakens driven by stronger incentives to acquire information. This result is consistent with empirical evidence in Farboodi et al. (2022) in which investors have relatively more data on large firms than on small ones because the incentive for speculators to produce information increases with reduced competition, which raises both firm profitability and size.

Furthermore, it is also worth examining how information production is affected by changes in other model parameters related to the product market, including the unit production cost MC, the price sensitivity of demand b and market prospect parameters A_H and A_L. Again, we can apply the implicit function theorem to the equilibrium condition (11) to derive:

Corollary 1. *When $c \in (\underline{c}_n, \overline{c}_n)$ so that an interior solution $\widehat{\alpha}_n \in (0,1)$ exists, the equilibrium features $\frac{\partial\widehat{\alpha}_n}{\partial MC} < 0$, $\frac{\partial\widehat{\alpha}_n}{\partial b} < 0$, $\frac{\partial\widehat{\alpha}_n}{\partial A_H} > 0$, and $\frac{\partial\widehat{\alpha}_n}{\partial A_L} < 0$.*

Information production, measured by the amount $\widehat{\alpha}_n$ of informed speculators, decreases with the production cost MC. This result can be understood by analyzing the expected trading profit $\Pi(\alpha)$, which is lower for a higher MC. Obviously, a lower expected trading profit will reduce the incentive for speculators to produce information, decreasing the equilibrium amount of information production. Similarly, when demand becomes relatively more sensitive to price (i.e., $b \uparrow$), the amount $\widehat{\alpha}_n$ of informed speculators will also decrease, since the expected trading profit Π is lower for a higher b. Furthermore, $\widehat{\alpha}_n$ increases in A_H and decreases in A_L. To understand these, note that the expected trading profit Π increases in the market uncertainty that is proportional to $(A_H - A_L)^2$. Therefore, a larger gap of $(A_H - A_L)$ increases the expected trading profit of informed speculators, inducing them to acquire more information.

4.2 Feedback Effects and Allocative Efficiency

The previous section shows that reduced competition in the product market enhances the information efficiency of the stock market. We now examine how this improvement in price informativeness affects allocative efficiency in the real economy. The central idea is that, through the feedback effect, managers' ability to learn from stock prices helps correct potential underestimation or overestimation of the market prospect $A(\omega)$, improving their production decisions and thereby increasing real efficiency via more effective information production.

We begin by introducing the probability of misallocation, which stems from managerial underestimation or overestimation of the market prospect. From Lemma 1,

$$\Pr(\forall i : q_i^* = q_M \mid \omega = H) = (\widehat{\gamma}_n)^n \quad \text{and} \quad \Pr(\forall i : q_i^* = q_H \mid \omega = H) = 1 - (\widehat{\gamma}_n)^n$$

Thus, with probability $1 - (\widehat{\gamma}_n)^n$, the true state $\{\omega = H\}$ is revealed through stock prices, allowing managers to correctly estimate the market prospect A_H. As a result, both the aggregate output and the price to align with those in Cournot competition under complete information; that is, $Q_H(n) = \frac{n(A_H - MC)}{b(n+1)}$ and $P_H(n) = \frac{A_H + nMC}{(n+1)}$. However, with complementary probability $(\widehat{\gamma}_n)^n$, stock prices remain uninformative, leading managers to underestimate the market prospect. This results in an inefficiently lower output $Q_M(n) = \frac{n(\bar{A} - MC)}{b(n+1)} < Q_H$ and a higher price $P_{MH}(n) = P_H(n) + \frac{n(A_H - A_L)}{2(n+1)} > P_H(n)$. Thus, $(\widehat{\gamma}_n)^n$ represents the probability of misallocation when the true state is $\omega = H$. Similarly, misallocation occurs with

probability $(\widehat{\gamma}_n)^n$ when the true state is $\omega = L$, where managers may overestimate the market prospect.

Next, we measure total welfare, $W(n;\omega)$, which includes both firm profits, $\Gamma_\omega(n) = \mathbb{E}\left[\sum_{i=1}^n TP_i \mid \omega\right]$ and consumer surplus, $CS_\omega(n) = \frac{1}{2}(A(\omega) - P)Q$. Formally, total welfare is given by:

$$W(n;\omega) = \frac{1}{2}(A(\omega) - P)Q + \mathbb{E}\left[\sum_{i=1}^{n} TP_i \mid \omega\right], \tag{12}$$

Since $A(\omega)$ is random, the expected total welfare and consumer welfare are given by $\overline{W} = \mathbb{E}_\omega[W(n;\omega)]$ and $\overline{CS} = \mathbb{E}_\omega[CS_\omega(n)]$, respectively.

Allocative efficiency gains. We now analyze how gains (or losses) in allocative efficiency arise through feedback effects. Figure 1 illustrates the source of these efficiency changes by comparing the total welfare between n firms and $(n-1)$ firms when the true state is $\omega = H$. Specifically, in the case of n firms, with probability $1-(\widehat{\gamma}_n)^n$, managers correctly estimate the market prospect A_H, resulting in an output of $Q_H(n)$ and corresponding welfare represented by the area Area$(ABNM)$. Conversely, with complementary probability $(\widehat{\gamma}_n)^n$, the output $Q_M(n)$ is lower due to managerial underestimation of the market prospect, and the welfare is represented by the area Area$(ABFE)$. By weighting these two areas by the probabilities of $(\widehat{\gamma}_n)^n$ and $1-(\widehat{\gamma}_n)^n$, we obtain the expected total welfare $W_H(n)$ given $\omega = H$, which corresponds to the blue trapezoid area, Area$(ABHG)$.

In contrast, when there are $(n-1)$ firms, with probability $1-(\widehat{\gamma}_{n-1})^{n-1}$, the output is $Q_H(n-1)$, and the corresponding welfare is represented by the area Area$(ABLK)$; with complementary probability $(\widehat{\gamma}_{n-1})^{n-1}$, managers underestimate the market prospect and the output is $Q_M(n-1)$, resulting in a lower welfare represented by the area Area$(ABDC)$. By weighting these two areas by the probabilities $(\widehat{\gamma}_{n-1})^{n-1}$ and $1-(\widehat{\gamma}_{n-1})^{n-1}$, we obtain the expected total welfare $W_H(n-1)$ given $\omega = H$, which corresponds to the blue trapezoid area Area$(ABJI)$.

The welfare gain due to reduced competition is then given by $W_H(n-1;\omega) - W_H(n;\omega)$, which is positive only when Area$(ABJI) >$ Area$(ABHG)$ holds. Indeed, this condition holds when the price impact from reduced competition is negative. To assess the price impact, note that $\overline{P}_H(n) = (\widehat{\gamma}_n)^n \times P_{HM} + (1-(\widehat{\gamma}_n)^n) \times P_H$ and $\overline{P}_H(n-1) = (\widehat{\gamma}_{n-1})^{n-1} \times P_{HM}(n-1) + (1-(\widehat{\gamma}_{n-1})^{n-1}) \times P_H(n-1)$. Thus, the price effect from reduced competition

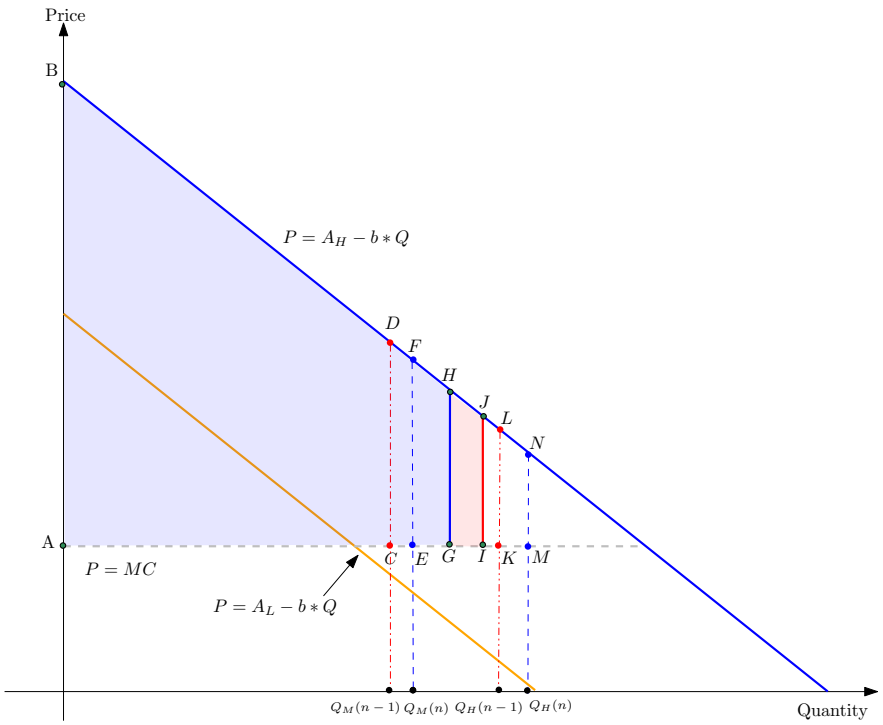


Figure 1: Allocative Efficiency Gain ($\omega = H$)

Notes: When there are n firms, with probability $1-(\widehat{\gamma}_n)^n$, the output is $Q_H(n)$ and the corresponding welfare is $Area(ABNM)$; with complementary probability $(\widehat{\gamma}_n)^n$, the output is $Q_M(n)$ and the welfare is $Area(ABFE)$. The expected welfare $W_H(n)$ then is the average of $Area(ABFE)$ and $Area(ABNM)$ weighted by $(\widehat{\gamma}_n)^n$ and $1-(\widehat{\gamma}_n)^n$, respectively. Similar discussion applies when there are $(n-1)$ firms, and the expected welfare $W_H(n-1)$ is the average of $Area(ABLK)$ and $Area(ABDC)$ weighted by $(\widehat{\gamma}_{n-1})^{n-1}$ and $1-(\widehat{\gamma}_{n-1})^{n-1}$, respectively. If $(\widehat{\gamma}_{n-1})^{n-1}$ is sufficiently small (compared with $(\widehat{\gamma}_n)^n$) such that condition (13) holds, an allocative efficiency gain will arise, i.e.,$W_H(n) = Area(ABHG) < W_H(n-1) = Area(ABJI)$.

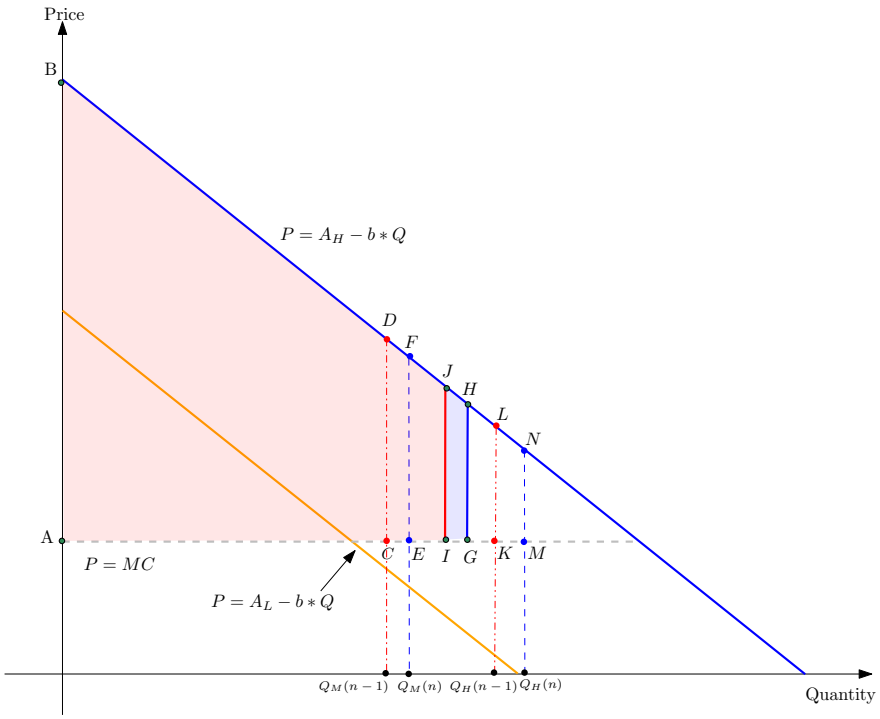


Figure 2: Allocative Efficiency Loss ($\omega = H$)

Notes: When there are n firms, with probability $1-(\widehat{\gamma}_n)^n$, the output is $Q_H(n)$ and the corresponding welfare is $Area(ABNM)$; with complementary probability $(\widehat{\gamma}_n)^n$, the output is $Q_M(n)$ and the welfare is $Area(ABFE)$. The expected welfare $W_H(n)$ then is the average of $Area(ABFE)$ and $Area(ABNM)$ weighted by $(\widehat{\gamma}_n)^n$ and $1-(\widehat{\gamma}_n)^n$, respectively. Similar discussion applies when there are $(n-1)$ firms, and the expected welfare $W_H(n-1)$ is the average of $Area(ABLK)$ and $Area(ABDC)$ weighted by $(\widehat{\gamma}_{n-1})^{n-1}$ and $1-(\widehat{\gamma}_{n-1})^{n-1}$, respectively. If $(\widehat{\gamma}_{n-1})^{n-1}$ is not sufficiently small (compared with $(\widehat{\gamma}_n)^n$) such that condition (13) does not hold, an allocative efficiency loss will arise, i.e., $W_H(n) = Area(ABHG) > W_H(n-1) = Area(ABJI)$.

in the state $\omega = H$ is:

$$\Delta\overline{P}_H(n) = \overline{P}_H(n-1) - \overline{P}_H(n) = \frac{A_H - MC}{n(n+1)} + \frac{A_H - A_L}{2n(n+1)}\left[\left(n^2-1\right)\left(\widehat{\gamma}_{n-1}\right)^{n-1} - n^2\left(\widehat{\gamma}_n\right)^n\right]$$

Interestingly, reduced competition can lead to a negative price impact (i.e., $\Delta\overline{P}_H(n) < 0$) when the reduction in misallocation probability is sufficiently significant, such that:

$$\left(\widehat{\gamma}_{n-1}\right)^{n-1} < \frac{n^2}{\left(n^2-1\right)}\left(\widehat{\gamma}_n\right)^n - \frac{2\left(A_H - MC\right)}{\left(n^2-1\right)\left(A_H - A_L\right)}. \tag{13}$$

Intuitively, this inequality holds if reduced competition significantly improves information production and lowers the value of $(\widehat{\gamma}_{n-1})^{n-1}$. This enables managers to better correct their underestimation of the market prospect A_H and reduce misallocation. This scenario corresponds to the allocative efficiency gain depicted in Figure 1. In contrast, Figure 2 illustrates allocative efficiency losses under weak feedback effects when equation (13) is violated.

Two clarifications are necessary regarding allocative efficiency gains (or losses). First, allocative efficiency gains cannot occur in the state $\omega = L$, as managers overestimate the market prospect A_L. Reduced competition ($n \downarrow$) decreases both $Q_L(n) = \frac{n(A_L - MC)}{b(n+1)}$ (when the state is revealed) and $Q_M(n) = \frac{n(\bar{A} - MC)}{b(n+1)}$ (when prices are uninformative). Furthermore, improved price informativeness under reduced competition corrects managers' upward biases, causing them to further reduce output and thus increase prices. Hence, reduced competition always results in higher prices in the low state. Second, the high state ($\omega = H$) has a greater impact on total welfare due to its larger market size. Since allocative efficiency gains from feedback effects arise mainly in the high state, these gains dominate welfare outcomes only when market uncertainty is sufficiently large, making welfare in the low state relatively less important.

4.3 Competition and Real Efficiency

We now formally analyze the efficiency implications of product market competition with feedback effects. Traditional wisdom claims that standard Cournot competition always improves economic efficiency and that imperfect/insufficient competition, such as oligopolies and monopolies, often leads to dead weight loss (Willner, 1989). However, existing studies on Cournot competition ignore the feedback effects of the financial market. Proposition 2 explains why the traditional argument may fail: product market competition lowers speculators' incentives to acquire information, leading to inefficient production decisions. The previous section also shows how feedback effects can create allocative efficiency gains, potentially reversing the link between product competition and welfare.

Specifically, the expected total welfare in the presence of feedback effects is given by:

$$\overline{W}\left(\widehat{\alpha}_n, n\right) = \frac{n(n+2)}{8b(n+1)^2}\left(4\left(\bar{A} - MC\right)^2 + \left(1 - \widehat{\gamma}_n{}^n\right)\left(A_H - A_L\right)^2\right), \tag{14}$$

where $\widehat{\gamma}_n = 1 - \widehat{\alpha}_n(2\theta - 1)$. Correspondingly, consumer welfare is given by:

$$\overline{CS}\left(\widehat{\alpha}_n, n\right) = \frac{n^2}{8b(n+1)^2}\left(4\left(\bar{A} - MC\right)^2 + \left(1 - \widehat{\gamma}_n{}^n\right)\left(A_H - A_L\right)^2\right). \tag{15}$$

Note that both $\overline{W}\left(\widehat{\alpha}_n, n\right)$ and $\overline{CS}\left(\widehat{\alpha}_n, n\right)$ strictly increase with average profitability $(\overline{A} - MC)$ and market uncertainty $(A_H - A_L)$. Notably, $\overline{W}\left(\widehat{\alpha}_n, n\right)$ becomes more sensitive to $(A_H - A_L)$ as the number of informed speculators increases (i.e., $\widehat{\alpha}_n \uparrow$), reducing the

probability of misallocation $(\widehat{\gamma}_n)^n$. This effect arises only due to informational feedback.

Next, we examine the relationship between total welfare and firm competition in the presence of feedback effects and investigate whether total welfare $\overline{W}(\widehat{\alpha}_n, n)$ can be negatively associated with the competition parameter n. To this end, we compute the total derivative of total welfare $\overline{W}(\widehat{\alpha}_n, n)$ with respect to n, the number of firms, as follows:

$$\frac{d\overline{W}(\widehat{\alpha}_n, n)}{dn} = \underbrace{\frac{\partial \overline{W}(\widehat{\alpha}_n, n)}{\partial n}}_{\text{Competition Effects}} + \underbrace{\frac{\partial \overline{W}(\widehat{\alpha}_n, n)}{\partial \widehat{\alpha}_n}\frac{\partial \widehat{\alpha}_n}{\partial n}}_{\text{Feedback Effects}}. \tag{16}$$

Equation (16) decomposes the total welfare effect into direct competition effects and feedback effects. Obviously, one can verify that $\frac{\partial \overline{W}(\widehat{\alpha}_n, n)}{\partial n} > 0$, which is consistent with the conventional wisdom that product market competition tends to increase total welfare (see, e.g., Willner, 1989). Meanwhile, since Proposition 2 establishes that $\frac{\partial \widehat{\alpha}_n}{\partial n} < 0$ (i.e., fierce product competition discourages information production), it might be possible for $\frac{d\overline{W}(\widehat{\alpha}_n, n)}{dn}$ to be negative when $\frac{\partial \overline{W}(\widehat{\alpha}_n, n)}{\partial \widehat{\alpha}_n}$ is positive and sufficiently large. Note that $\frac{\partial \overline{W}(\widehat{\alpha}_n, n)}{\partial \widehat{\alpha}_n}$ measures the sensitivity of total welfare to the amount of information produced by speculators $\widehat{\alpha}_n$ in the stock market. Intuitively, as $\widehat{\alpha}_n$ increases, a higher level of informativeness of the stock market improves real efficiency in production, and thus a positive value of $\frac{\partial \overline{W}(\widehat{\alpha}_n, n)}{\partial \widehat{\alpha}_n}$ follows.[10]

Lemma 3 (Competition and Real Efficiency)**.**

Define $G_1(A_H, A_L, MC) = 2 + 8\left(\bar{A} - MC\right)^2/(A_H - A_L)^2$, $\gamma = 1 - \alpha(2\theta - 1)$ *and*

$$g_1(\alpha, n) = 2\gamma^n + \frac{n(n+2)\gamma^n}{2 + n(n-1)\gamma^{n-1}}\left(4n + n(n-3)\gamma^{n-1} - 2(n+1)\ln\frac{1}{\gamma}\right)$$
$$g_2(\alpha, n) = 2\gamma^n + \frac{n\gamma^n}{2 + n(n-1)\gamma^{n-1}}\left(4n + n(n-3)\gamma^{n-1} - 2(n+1)\ln\frac{1}{\gamma}\right)$$

Then: (i) when $g_1(\widehat{\alpha}_n, n) > G_1(A_H, A_L, MC)$ *holds,* $\frac{d\overline{W}(\widehat{\alpha}_n, n)}{dn} < 0$*, that is, product market competition decreases total welfare; and*

(ii) when $g_2(\widehat{\alpha}_n, n) > G_1(A_H, A_L, MC)$ *holds,* $\frac{d\overline{CS}(\widehat{\alpha}_n, n)}{dn} < 0$*, that is, product market competition decreases consumer welfare.*

Lemma 3 characterizes when competition decreases real efficiency. First, note that the condition in Lemma 3 is non-empty. For example, this occurs when the price sensitivity b of demand is sufficiently high such that the probability of misallocation is large.[11]

[10] Using Equation (14), we can directly compute: $\frac{\partial \overline{W}(\widehat{\alpha}_n, n)}{\partial \widehat{\alpha}_n} = \frac{n^2(n+2)(2\theta-1)\widehat{\gamma}_n{}^{n-1}}{8b(n+1)^2}(A_H - A_L)^2 > 0$.

[11] Note that $\lim_{b\to\infty}\widehat{\alpha}_n = 0$. Then, we get the approximation $g_1(\widehat{\alpha}_n, n) = \frac{n^2(n+1)(n+2)}{n(n-1)+2} + 2 + O(n\widehat{\alpha}_n)$,

Second, Lemma 3 examines the role of market uncertainty $(A_H - A_L)$ and average profitability $(\bar{A} - MC)$ in shaping the efficiency effects of product market competition through feedback. Specifically, $G_1(A_H, A_L, MC)$ increases with average profitability and decreases with market uncertainty. Thus, when market uncertainty is high and average profitability low, the condition in Lemma 4.2(i) is more likely to hold, leading to a negative welfare effect from product market competition.

Third, the potential negative welfare effect depends on the probability of misallocation $(\widehat{\gamma}_n)^n$ through $g_1(\widehat{\alpha}_n, n)$. When the probability of misallocation is maximized $(\widehat{\gamma}_n = 1)$, we estimate $g_1 = 2 + \frac{n^2(n+1)(n+2)}{2+n(n-1)}$. As $\widehat{\gamma}_n$ approaches zero, g_1 tends to zero. Thus, g_1 increases with the probability of misallocation or decreases with information production, although it is not strictly monotonic in either variable. This suggests that the negative welfare effect of competition $(g_1(\widehat{\alpha}_n, n) > G_1(A_H, A_L, MC))$ is more likely when the probability of misallocation is not too low, allowing feedback effects to generate enough gains in allocative efficiency when competition decreases. However, Section 4.2 points out that feedback effects may instead cause a loss in allocative efficiency. Such losses would reduce total welfare, consistent with the nonmonotonicity of $g_1(\widehat{\alpha}_n, n)$.

Since Lemma 3 involves the endogenous variable of information acquisition, we now provide a more direct result through constructive derivations.

Proposition 3 (Welfare-destructive Overcompetition)**.**

Consider a pair of positive integers (m, n) *satisfying* $\Phi(m) \geq 1$ *and* $n > N(m)$*, where*[12]

$$\Phi(m) = \left(1 + \frac{(A_H - A_L)^2(1 - (2-2\theta)^m)}{4(\overline{A} - MC)^2}\right) \times \frac{m(m+2)}{(m+1)^2}$$
$$N(m) = \frac{(m+1)^2}{(2-2\theta)\left(2 + (m-1)(2-2\theta)^{m-1}\right)} \geq m+1$$

Then: $\overline{W}\left(\widehat{\alpha}_m, m\right) > \overline{W}\left(\widehat{\alpha}_n, n\right)$ *holds for any* $c \in [\bar{c}_n, \underline{c}_m)$ *with* $\bar{c}_n < \underline{c}_m$*.*

Denote $m_0 := \inf\{m \in \mathbb{N} : \Phi(m) \geq 1\} < \infty$. Proposition 3 shows that when the number of firms exceeds $N(m_0)$, the total welfare is strictly less than with m_0 firms.

Theorem 1 below directly follows from Proposition 3.

---

where $O(\cdot)$ means "big O". Now suppose that $g(0, n) > G_1$, or equivalently, $\frac{(\bar{A}-MC)^2}{(A_H-A_L)^2} < \frac{n^2(n+1)(n+2)}{8(n(n-1)+2)}$. By continuity, for any $\widehat{\alpha}_n > 0$ sufficiently small, $g_1\left(\widehat{\alpha}_n, n\right) > G_1(A_H, A_L, MC)$ holds.

[12]Note that $\Phi(m) \geq 1$ is non-empty because $\lim_{m\to\infty} \Phi(m) = 1 + \frac{(A_H-A_L)^2}{4(\overline{A}-MC)^2} > 1$. Furthermore, since $\Phi(m)$ strictly increases in m, $\Phi(m_1) > \Phi(m_2)$ if $m_1 > m_2$.

Theorem 1. *Competition can reduce total welfare through informational feedback effects.*

Theorem 1 underscores the welfare-reducing effect of competition through information feedback. Specifically, when information production $\widehat{\alpha}$ is fixed, Equation (14) shows that increasing the number of firms always raises total welfare. Thus, Theorem 1 reveals that competition reduces welfare solely through the information production channel. Furthermore, for any positive integer m that satisfies $\Phi(m) \geq 1$, there exists a range of cost parameters c for which excessive competition lowers the total welfare when $n \geq N(m)$.

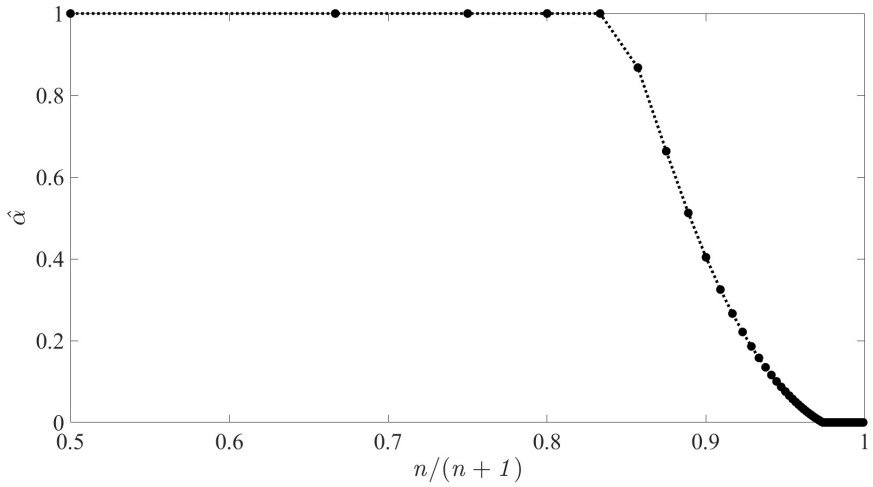


Figure 3: Product Competition and Information Production

Our main insight is illustrated in Figures 3 and 4.[13] First, Figure 3 shows how intensified competition affects information production incentives (Proposition 2). As competition increases ($n \uparrow$), information production transitions from full information ($\widehat{\alpha} = 1$), to partial information ($0 < \widehat{\alpha} < 1$), and ultimately to none ($\widehat{\alpha} = 0$). Second, Figure 4 illustrates the non-monotonic welfare effects of competition, with total welfare maximized at $n = 6$. Specifically: (i) for n small, the welfare increases as the market power declines; (ii) for n intermediate, the welfare decreases as the feedback effect dominates; and (iii) for n large, the welfare increases again as information production ceases, making the market power concentration channel dominant.

Interestingly, the interplay between Figure 3 and Figure 4 reveals two notable patterns that warrant closer examination. First, the decline in information production precedes the reduction in total welfare. Second, the observed non-monotonicity is primarily attributable to an interior solution in information production, rather than corner solutions. In addition,

[13]Baseline parameters are $\theta = 0.75$, $b = 1.5$, $A_H = 30$, $A_L = 10$, $c = 1.5$, and $MC = 3$, used throughout unless stated otherwise. See online Appendix B.3 for analogous results using US market data.

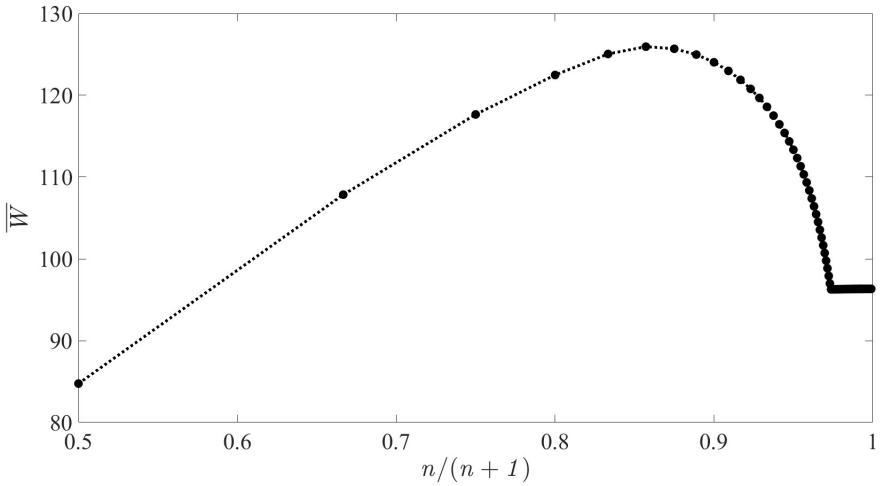


Figure 4: Product Competition and Total Welfare

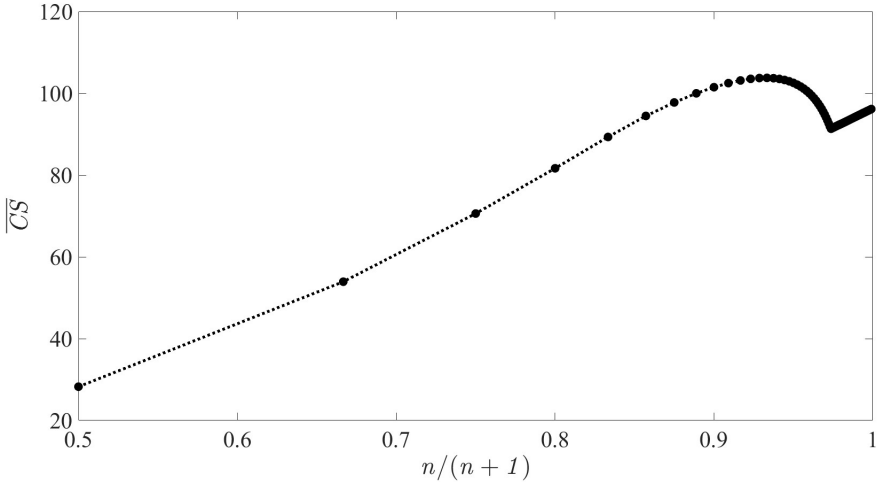


Figure 5: Product Competition and Consumer Surplus

Figure 5 illustrates a similar non-monotonic pattern in consumer surplus when we vary the number of firms n.[14]

Remark 1. *Under extreme parameter values, where low market uncertainty reduces the informational value of managerial learning, the stock market feedback effect may not overturn the positive link between competition and total welfare. Nonetheless, it can significantly shape the efficiency implications of firm competition, making it a crucial factor in regulating horizontal mergers. See online Appendix B.1 for a detailed discussion.*

[14]Specifically, in this numerical example, the consumer surplus increases first for $n \leq 14$, then decreases for $14 \leq n \leq 37$, and finally increases again for $n \geq 37$. Note that the consumer surplus is maximized at $n = 14$, rather than at $n = 6$.

4.4 Optimal Market Structure and Comparative Statics

This section examines the optimal market structure and performs comparative statics. Without feedback effects, the maximum total welfare is achieved as $n \to \infty$. However, with feedback effects, competition may reduce efficiency, and the maximum welfare may occur at a finite n^*, which we define as the optimal market structure.

Proposition 4 (Optimal Market Structure)**.** *The optimal market structure, n^*, can be non-monotonic in the information production cost c and the price sensitivity b.*

The non-monotonicity in Proposition 4 is driven by feedback effects and allocative efficiency gains. The negative relationship between competition and welfare results from the sensitivity - rather than the absolute level - of information production to changes in competition. When information costs are very high, no speculators acquire information, eliminating feedback effects. Conversely, when costs are very low, all speculators acquire information, making information production insensitive to competition. Thus, competition reduces welfare only for intermediate information costs where an interior equilibrium emerges.

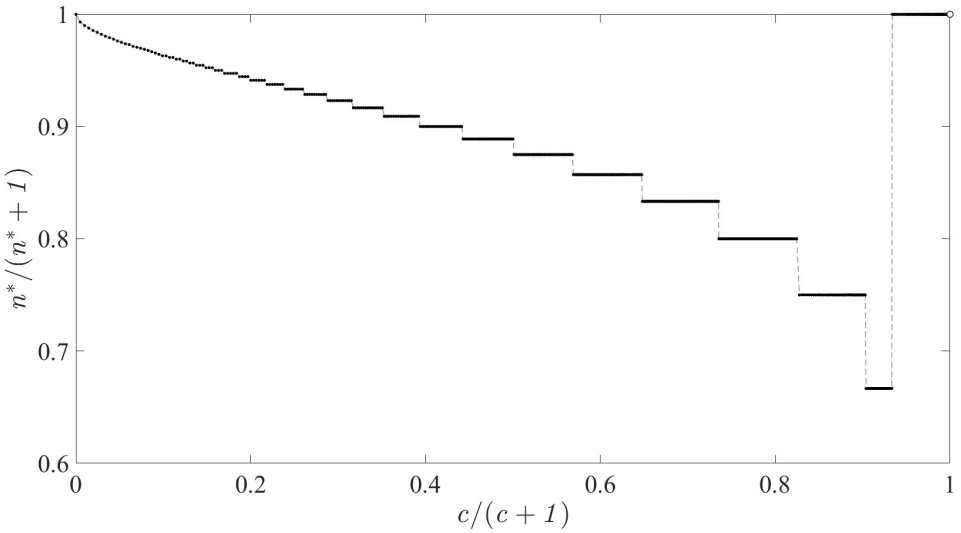


Figure 6: Optimal Market Structure n^*

Figure 6 illustrates the non-monotonic dependence of the optimal market structure n^* on information production cost c. As c decreases, n^* initially moves from perfect competition to a duopoly and then expands to three or more firms. In intermediate ranges of c, partial information production occurs, and fewer firms may dominate more firms in terms of welfare. For sufficiently low costs, most speculators become informed, making information production

insensitive to changes in n and leading welfare to rise with increased competition. A similar pattern emerges for price sensitivity b (see online Appendix B.2).

Average profitability and market uncertainty. To better illustrate their economic intuition and implications, we discuss the role of average profitability and market uncertainty in shaping the link between competition and total welfare when $n^* < \infty$. Specifically, we use numerical methods to address the complexity of the auxiliary function $g_1(\alpha, n)$, complementing our earlier analytical results. Theoretical insights, including Lemma 3 and the following discussions in Section 4.3, provide guidance for the numerical analysis. We anticipate that a negative relationship between competition and total welfare is more likely to occur with high market uncertainty $(A_H - A_L)$ and low average profitability $\left(\bar{A} - MC\right)$. Meanwhile, by Equation (8) and Equation (11), these two factors also contribute to information production $\widehat{\alpha}$ in equilibrium. Define:

$$\Delta W_n := \overline{W}\left(\widehat{\alpha}_n, n\right) - \overline{W}\left(\widehat{\alpha}_{n-1}, n-1\right).$$

Obviously, a negative relationship between product market competition and total welfare ensues when $\Delta W_n < 0$ holds. We also focus on interior solutions of $\widehat{\alpha}_n$. Sensitivity analyses performed on a wide range of model parameter values have shown a similar pattern.

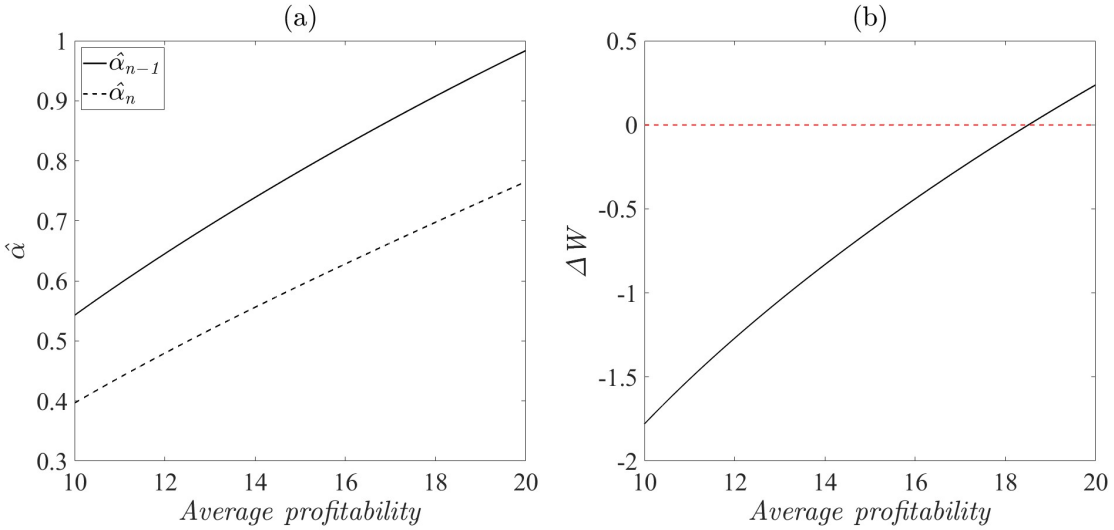


Figure 7: Average Profitability, Information Quality and Welfare.

Then we analyze the impact of average profitability $(\bar{A} - MC)$ on equilibrium information production $\widehat{\alpha}_n$ and total welfare ΔW_n. For this exercise, we fix the value of $(A_H - A_L)$ and

other parameters. The results are plotted in Figure 7. We make three observations: First, Figure 7a shows that $\widehat{\alpha}_n$ is always lower than $\widehat{\alpha}_{n-1}$, which is consistent with the prediction of Proposition 2 that product market competition dampens the incentive for speculators to produce information. Second, both $\widehat{\alpha}_n$ and $\widehat{\alpha}_{n-1}$ increase strictly in average profitability, implying that higher profitability improves information acquisition. Third, Figure 7b shows that the welfare gain ΔW_n is smaller for a lower level of average profitability. In particular, when the average profitability is sufficiently low, ΔW_n can be negative, indicating that intensified competition decreases the total welfare. Note that this result coincides with our discussion following Lemma 3.

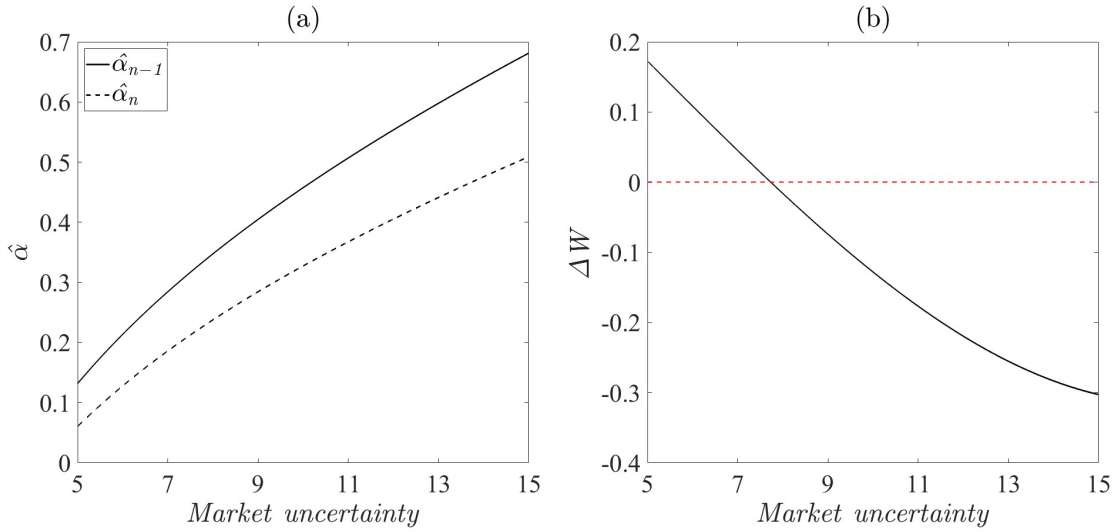


Figure 8: Market Uncertainty, Information Quality and Welfare.

Next, we investigate the effects of market uncertainty on $\widehat{\alpha}_n$ and ΔW_n by varying $(A_H - A_L)$ while keeping the average profitability $(\bar{A} - MC)$ and other parameters unchanged. These results are depicted in Figure 8. We make two observations: First, Figure 8a shows that both $\widehat{\alpha}_n$ and $\widehat{\alpha}_{n-1}$ increase as $(A_H - A_L)$ increases, which implies that increasing market uncertainty improves information production. Second, as shown in Figure 8b, competition can decrease total welfare when market uncertainty is high, despite the high incentive of information production (i.e., $\widehat{\alpha}$ is high).

This illustrates a sharp difference between average profitability and market uncertainty. Although both exhibit similar effects on information production, the welfare implications of competition diverge. Specifically, a negative relationship between competition and total welfare is more likely to occur when: (i) the average profitability is low; or (ii) the market uncertainty is high. To understand this divergence, we highlight two observations: First, an

increase in average profitability directly increases total welfare, which reduces the relative impact of information production, while an increase in market uncertainty amplifies that of information production (see Equation (14)). Second, the negative link between competition and welfare depends on the relative gap, rather than the absolute intensity, in information production when the level of competition varies.

4.5 Implications for Horizontal Mergers

To better illustrate the empirical implications for horizontal mergers, we first compare a monopoly (i.e., $n = 1$) and a duopoly (i.e., $n = 2$) in perfectly symmetric Cournot competition.By Equation (14), the total welfare for a monopolist seller is given by:

$$\overline{W}\left(\widehat{\alpha}_1, 1\right) = \frac{3}{32b}\left(4\left(\bar{A} - MC\right)^2 + \left(1 - \widehat{\gamma}_1\right)\left(A_H - A_L\right)^2\right) \tag{17}$$

and that for two duopoly sellers are given by

$$\overline{W}\left(\widehat{\alpha}_2, 2\right) = \frac{1}{9b}\left(4\left(\bar{A} - MC\right)^2 + \left(1 - (\widehat{\gamma}_2)^2\right)\left(A_H - A_L\right)^2\right) \tag{18}$$

Obviously, if we fix the size of informed traders $\widehat{\alpha}_1 = \widehat{\alpha}_2$ (or equivalently $\widehat{\gamma}_1 = \widehat{\gamma}_2$) to shut down the information production channel, a duopoly market always outperforms a monopoly in total welfare. In other words, any regulatory action based on market concentration measures is well-founded. However, if we allow for endogenous information production, the above insight might not hold, as illustrated by Lemma 4 below.

Lemma 4 (Monopoly VS. Duopoly)**.**

Assume that $A_H > A_L = MC$. *Denote* $\kappa = (2\theta - 1)(A_H - A_L)^2/b$.

(i) When $\frac{\kappa}{12} \leq c < \frac{11}{108}\kappa$, *then* $\overline{W}\left(\widehat{\alpha}_1, 1\right) > \overline{W}\left(\widehat{\alpha}_2, 2\right)$; *and*

(ii) when $c \geq \frac{11}{108}\kappa$ *or* $c < \frac{(1-\theta)(2-\theta)\kappa}{9}$, *then* $\overline{W}\left(\widehat{\alpha}_1, 1\right) \leq \overline{W}\left(\widehat{\alpha}_2, 2\right)$.

We briefly comment on Lemma 4. First, a monopoly dominates a duopoly for an intermediate level of information production cost c. In Statement (i), a lower bound $c \geq \frac{\kappa}{12}$ is imposed to completely remove information production in a duopoly market (i.e., $\widehat{\alpha}_2 = 0$), while an upper bound $c < \frac{11\kappa}{108}$ ensues that the incentive to produce information is strong enough in a monopoly market (i.e., $\widehat{\alpha}_1 \uparrow$). Second, when information production is too cheap or too costly, the relative gap in information production is small, and thus a duopoly market is more efficient due to lowered market concentration.

Obviously, our theory differs sharply from the existing literature on merger analysis, which largely ignores the information efficiency of the stock market and often features a monotonic relationship between competition and total welfare in perfectly symmetric Cournot competition when all firms are equally efficient (see, e.g., Farrell and Shapiro, 1990). In contrast, even in the simplest case here, merging two competing and equally efficient firms into a monopolist can improve social welfare for an intermediate level of information production cost when market concentration significantly increases information production. This naturally arises when managerial learning from the stock market benefits production decisions in a feedback loop. Our theory highlights the importance of considering the interaction between the product market and the financial market in M&As regulations from an informational perspective.[15]

Remark 2 (Beyond Monopoly & Duopoly)**.** *We can extend the analysis beyond two firms. Theorem 1 offers a framework for this analysis. Define* $m_0 := \inf\{m \in \mathbb{N} : \Phi(m) \geq 1\}$*. For* $n \geq N(m_0)$*, over-competition emerges in terms of total welfare within an intermediate range of information production costs, as it is strictly dominated by a market structure with* $n = m_0$*. Thus, reducing the number of firms to* $n < N(m_0)$ *can enhance total welfare, though the optimal number* n^* *requires numerical determination.*[16]

Furthermore, our treatment of M&As closely follow the spirit of Cournot competition in the long-run sense, differing from that of Nocke and Whinston (2022), where the post-merger HHI merely aggregates pre-merger market shares. Our analysis complements existing M&A frameworks by emphasizing the interplay between financial and product markets, alongside well-documented factors such as production efficiency asymmetries (Farrell and Shapiro, 1990), synergies (see, e.g., Maksimovic and Phillips, 2001), disclosure (Xiong and Yang, 2021), investment (Mermelstein et al., 2020; Motta and Tarantino, 2021), and innovation (Yi, 1999; Aghion et al., 2005; Segal and Whinston, 2007; Spulber, 2013).

A "calibrated" illustration. We present a numerical example to illustrate the welfare effects of a horizontal merger under the feedback effect. Although this is not intended as a formal calibration directly comparable to the US economy, it offers qualitative insights into

[15]While this non-monotonic relationship between competition and total welfare also appears in other studies on, the non-monotonicity there stems from some presumptions of anticompetitive effects such as cost synergies (see, e.g., Nocke and Whinston, 2022). We abstract away from those considerations to focus on the impact of informational feedback.

[16]The dominated structures $n \geq N(m)$ can also be chosen conditional on the information cost c.

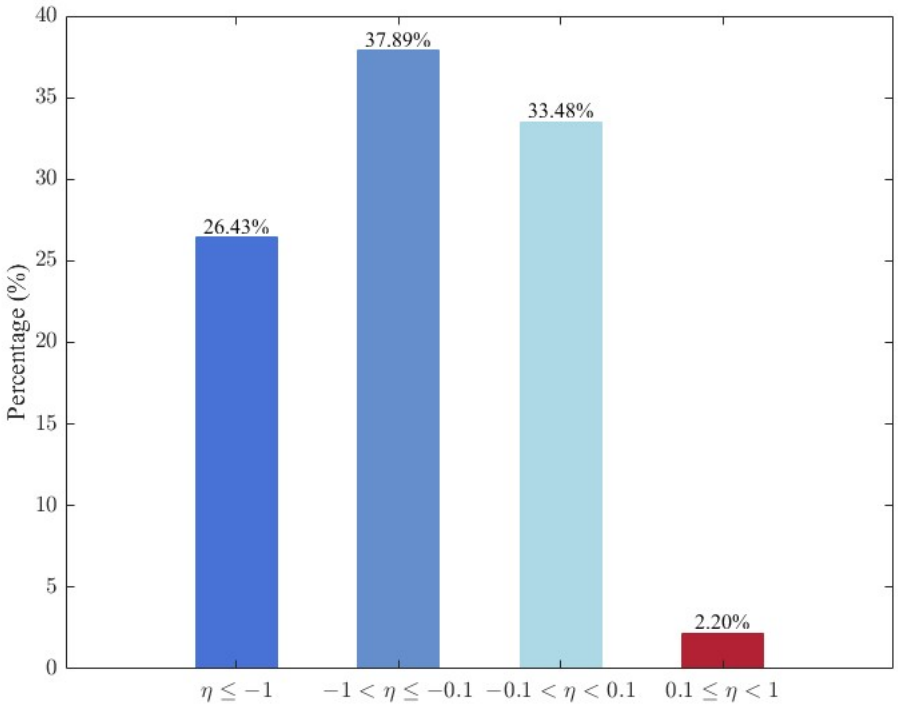


Figure 9: Estimation of η by industries

Notes: This histogram summarizes the estimation of η across industries, which are classified following Gu (2016) and Hou and Robinson (2006). The estimation is based on model parameters calibrated with US market data over 2000–2010. A negative value of η indicates that the welfare effect of a horizontal merger will be overestimated if the feedback effect is ignored. A positive value of η then suggests that the feedback effect augments the welfare effect of a horizontal merger.

the significance of feedback effects in assessing the economic implications of mergers.

Specifically, the welfare effect of a horizontal merger, both with and without feedback effects, can be expressed as $\overline{W}\left(\widehat{\alpha}_n, n\right) - \overline{W}\left(\widehat{\alpha}_{n-1}, n-1\right)$ and $\overline{W}(0,n) - \overline{W}(0,n-1)$. We then define the impact of informational feedback from the stock market on the welfare of horizontal mergers as:

$$\eta = \frac{\overline{W}\left(\widehat{\alpha}_n, n\right) - \overline{W}\left(\widehat{\alpha}_{n-1}, n-1\right)}{\overline{W}(0,n) - \overline{W}(0,n-1)} - 1. \tag{19}$$

Using US market data and the calibration method detailed in online Appendix B.3, we estimate model parameters and compute the corresponding values of η in all industries after excluding firms in the financial and utility industries, as well as industries with negative gross margins.

Figure 9 illustrates the industry-level distribution of η values. The key findings are as follows. On the one hand, in 64.32% of all industries, including the first two bars in Figure 9,

the feedback effects of the stock market significantly weaken the welfare effect of horizontal mergers by more than 10%. Furthermore, in 26.43% of all industries, the impact of stock market feedback exceeds 100%, which implies that it completely reverses the welfare effects. On the other hand, in 2.20% of all industries, feedback effects amplify the welfare effect of mergers (referred to as the augmentation effect).

Overall, these results highlight that feedback effects from the stock market constitute a critical factor in analyzing the welfare impact of horizontal mergers and the efficiency of market competition. Ignoring these effects can lead to misinterpretations of merger outcomes.

5 Further Discussions

5.1 Cross-Asset Trading

Although standard in the literature (see, e.g., Foucault and Frésard, 2014, 2019), bounded asset positions ($x_k^i \in [-1,1]$) in our baseline model may not be as harmless as in other settings: If the total product market size is stable, with an increase in the number of firms, the size and, consequently, the equity value of each firm decrease. Therefore, the dollar value of the maximum trade size could decrease in n, and thus the incentive to acquire information might mechanically decrease. To address this concern and show robustness, we now allow cross-asset trading, in which a fraction of speculators can trade all stocks. All baseline findings continue to hold.

Specifically, we consider an economy with $n \geq 2$ identical firms competing in quantities and a stock exchange, which is populated with four types of investors, including: (i) a mass $\lambda \in [0,1]$ of risk-neutral L-traders $k \in [0,\lambda]$, who choose whether to acquire a costly signal m_k at a cost $c_L > 0$, and trade all stock shares $y_k^i \in [-1,1]$ for all i; (ii) a mass $1-\lambda$ of risk-netural S-traders $k \in [0,1-\lambda]$ for each stock i, who choose whether to acquire a costly signal m_k^i at a cost $c_S > 0$ and only trade shares $x_k^i \in [-1,1]$ for the ith stock. (iii) liquidity traders with aggregate demand z_i, uniformly distributed over $[-1,1]$, for each firm i, and (iv) risk-neutral market makers who set prices to clear each stock.

Let $y_i = \int_0^{\alpha_L} y_k^i dk$ and $x_i = \int_0^{\alpha_{i,S}} x_k^i dk$ denote the aggregate demand for stock i by L- and S-traders. Recall that the aggregate order submitted by liquidity traders is z_i. Thus, the total order flow f_i for the ith stock is then given by: $f_i = x_i + y_i + z_i$. As in Goldstein et al. (2014), we assume that $c_L \leq c_S$, i.e., an L-trader has a relatively lower cost of information

production.[17] For ease of reference, let α_L and $\alpha_{i,S}$ denote the measure of informed L-traders and that of informed S-traders for the ith firm. Define $\boldsymbol{\alpha} := (\alpha_L, \alpha_{1,S}, \cdots, \alpha_{n,S})$. All other features of the model are the same. Note that when $\lambda = 0$, it reduces to the baseline setup.

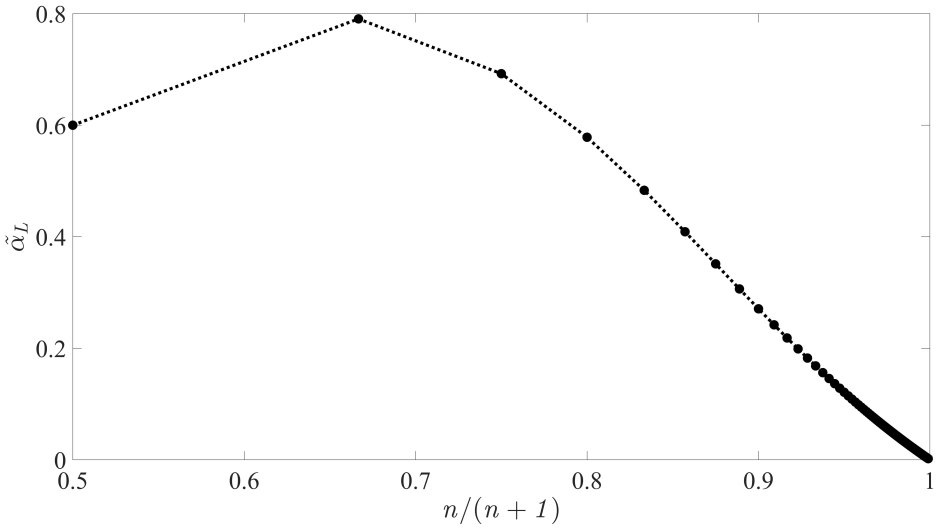


Figure 10: Trading Opportunities & (Non-monotonic) Information Production

We briefly summarize the key insights, while the equilibrium analysis can be found in online Appendix B.4. First, L-traders have a stronger incentive to acquire information than S-traders, given that $c_L \leq c_S$. Actually, the incentive for L-traders to acquire information can even increase in the number of firms n, which differs sharply from S-traders for whom the incentive for information acquisition is always maximized in a monopoly. This complexity is illustrated in Figure 10.[18] In particular, when we move from a monopoly ($n = 1$) to a duopoly ($n = 2$), the size of the informed L-traders $\widetilde{\alpha}_L$ first increases and then decreases.[19]

Second, our baseline result remains valid in the presence of L-traders, because the incentive for information production for L-traders will drop quickly after achieving its maximum level, and thus a negative relationship between competition and total welfare ensues.

5.2 Cross-Asset Learning

In the baseline model, we assume that the market maker of the ith firm does not observe the order flow of the other firms. Therefore, there may be arbitrage opportunities between

[17] To be precise, Goldstein et al. (2014) sets $c_S > c_L = 0$, i.e., an L-trader costlessly observes a signal.

[18] Parameters used for the extended model with cross-asset trading are: $\lambda = 0.8$, $\theta = 0.75$, $b = 3.5$, $A_H = 20$, $A_L = 10$, $MC = 9$, and $c_L = c_S = 1.5$.

[19] Vives (1985) shows that the profit of competing firms vanishes at a speed order of $1/n$. When multiplied by the number of firms n, the trading profits for L-traders can be non-monotonicity in n. We term this the "trading opportunity effect" in cross-asset trading.

competing firms. This section removes this restriction and considers cross-asset learning, which refers to the possibility that market makers observe the order flow in all stocks before setting the price (see, e.g., Pasquariello and Vega, 2015; Foucault and Frésard, 2019). Specifically, we modify the more general setup in Section 5.1 by allowing for cross-asset learning, i.e., the information set for market makers is $\Omega = \{f_1, \cdots, f_n\}$. Again, as in Kyle (1985), risk-neutral market makers absorb excess order flow and break even only in expectation. Thus, the stock price of the ith firm is given by $s_i(\Omega) = \mathbb{E}[V_i|\Omega]$.

Here, we briefly discuss the main results with cross-asset learning, and delegate the formal analysis to online Appendix B.5. First, the baseline result holds in the presence of cross-asset learning when there are only S-traders. Intuitively, cross-asset learning empowers market makers, reducing trading profits for speculators, except for the special case with a monopoly. This in turn makes the trading profits more sensitive to the change in the number of competing firms when it is small. Thus, the information feedback channel is strengthened.

Second, the non-monotonicity can also appear when the cost of information production is small such that all L-traders choose to acquire information. Note that L-traders have a stronger incentive to acquire information compared to S-traders. Cross-asset trading makes S-traders more prone to competition compared to L-traders, and thus L-traders may crowd out S-traders due to their trading opportunities.

Third, total welfare can strictly increase with the number of firms n in the presence of cross-asset learning when S-traders are totally absent. In practice, however, markets are unlikely to consist solely of L-traders, as segmentation due to various frictions is common; see, e.g., Goldstein et al. (2014) for real-world examples of market segmentation. Moreover, even in markets with only L-traders, the efficiency implications of firm competition can still be significantly influenced by informational feedback from the stock market (though not to the extent of creating a non-monotonic relationship between competition and welfare). This feedback effect often exacerbates allocative efficiency losses as product market competition weakens, amplifying the welfare losses associated with market power concentration. Thus, the feedback effect remains a critical factor to consider in regulating horizontal mergers, even in the absence of S-traders. For a detailed discussion on the divergent impacts of cross-asset learning on L-traders and S-traders, see online Appendix B.5.

5.3 Investor Welfare

Investor welfare, especially that of liquidity traders, is largely missing from the total welfare defined in Equation (14), which essentially captures the welfare of the product market, including both the consumer surplus and the producers' surplus. We now show that our theoretical insights still hold when we include investor welfare in the calculation of total welfare. Recall that: (1) market makers always break even in expectation; (2) informed speculators incur acquisition costs but earn positive trading profits; (3) liquidity traders incur trading losses but enjoy liquidity benefits; and (4) informed speculators' trading profits equal liquidity traders' trading losses. Although liquidity benefits are conceptually endogenous, most papers treat them and liquidity trading as completely exogenous. The total cost of information acquisition varies with the size of informed speculators α, and given that we focus on the benefits of information, the cost of information acquisition should not be overlooked.

Specifically, let $B(n)$ denote the aggregate benefit of liquidity trading. Thus, total welfare $\overline{W}_{PF}$, including both product market welfare and investor welfare, can be measured as:

$$\overline{W}_{PF} = \overline{W} - n * \widehat{\alpha}_n * c + B(n) \tag{20}$$

where $\overline{W}(\widehat{\alpha}_n, n)$ is given by Equation (14).

When the aggregate benefits of liquidity trading are exogenously fixed (i.e., $B(n) = B_0$ for some non-negative constant B_0), a non-monotonic relationship between product competition and total welfare can arise, and the optimal market structure features a finite number of firms. Such non-monotonicities may manifest under other specifications if the aggregate benefits of liquidity trading are proportional to the number of stocks, although the optimal market structure might approach perfect competition when the benefits of liquidity trading become dominant. Online Appendix B.6 contains a formal analysis.

5.4 Discount Rates

In our primary analysis, we have not accounted for the effects of discounting. However, as Cochrane (2011) highlights, discount rates, rather than cash flows, may drive movements in stock prices, at least at the aggregate level. Given that variations in industrial competition can influence discount rates (Dou et al., 2021), incorporating discounting into the evaluation of firm value and stock prices could potentially alter our findings. To address this, we extend

our baseline model to explore the implications of discounting.

Let $r_n \geq 0$ denote the discount rate when n symmetric firms compete in the industry. Then, the expected firm value given in Equation (2) can be rewritten as:

$$V_i(q_i) = \frac{1}{1+r_n}\mathbb{E}\left[TP_i(q_i) \mid \mathrm{F}_m\right]$$

Note that the profit function $TP_i(q_i)$ is linear in the parameters A, b and MC, as shown in Equation (1). Thus, introducing discounting into the model is equivalent to replacing the original parameters (A, b, MC) with a set of new parameters (A', b', MC'), where

$$A'_\omega = \frac{A_\omega}{1+r_n}, \quad b' = \frac{b}{1+r_n}, \quad \text{and} \quad MC' = \frac{MC}{1+r_n}$$

Furthermore, the linearity implies that the baseline results in Section 3 can be obtained using (A', b', MC'). We now discuss the relationship between competition and discount rates and how it affects our results in Section 4. First, we assume that the discount rate r_n strictly increases in n (that is, $\frac{\partial r_n}{\partial n} > 0$) because increased competition can erode profitability and increase risk. This assumption is consistent with the existing literature that documents a positive correlation between competition and discount rates (Dou et al., 2021). We can use the chain rule of differentiation to get: $\frac{\partial \Pi'(n,\alpha)}{\partial n} = \frac{1}{1+r_n}\frac{\partial \Pi(n,\alpha)}{\partial n} - \frac{1}{(1+r_n)^2}\frac{\partial r_n}{\partial n} < \frac{1}{1+r_n}\frac{\partial \Pi(n,\alpha)}{\partial n}$ and $\frac{\partial \Pi'(n,\alpha)}{\partial \alpha} = \frac{1}{1+r_n}\frac{\partial \Pi(n,\alpha)}{\partial \alpha}$, which further implies:

$$\frac{\partial \widehat{\alpha}'_n}{\partial n} < \frac{\partial \widehat{\alpha}_n}{\partial n} < 0$$

Thus, when discounting is considered, increased competition discourages speculators from acquiring information. More importantly, discounting can exacerbate this negative impact of competition on information production. Consequently, we can reasonably anticipate that our main result will not only remain valid but may also be strengthened by the compounding effects of discounting. Specifically, reduced information production in the stock market, driven by intensified competition, could significantly decrease the allocative efficiency of the real economy, potentially leading to a negative relationship between competition and real efficiency due to feedback effects.

5.5 Dynamic Trading

Most existing studies focus on a static framework when modeling Cournot competition and feedback effects, as incorporating dynamic trading and competition can rapidly render the model intractable (Edmans et al., 2015; Goldstein and Yang, 2019; Lin et al., 2019). Consequently, we only provide an informal exploration of how our main results might be affected in a dynamic setting.

In general, introducing multiple rounds of trading creates opportunities for market manipulation, as the feedback effect from the stock market incentivizes speculators to influence stock prices (Edmans et al., 2015; Goldstein and Guembel, 2008). Specifically, uninformed traders may profit from selling the stock when feedback effects are present, partly because their trading distorts the information content of stock prices and misleads the firm's investment decisions. Consequently, we may expect that market manipulation, stemming from dynamic trading opportunities, could influence our main results by altering the informativeness of stock prices.

However, we argue that manipulation is more likely to occur in the stock trading of small firms rather than large firms. For instance, stocks characterized by high illiquidity and significant information asymmetry are more susceptible to manipulation (Comerton-Forde and Putniņš, 2014), and small-cap stocks typically exhibit low liquidity and limited transparency (Banz, 1981; Acharya and Pedersen, 2005). The reasoning is as follows. First, intensified competition reduces the size of firms, which in turn increases the potential for market manipulation. Second, information distortion caused by market manipulation can lead to a loss in real efficiency through feedback effects. As a result, our main findings should remain valid, and dynamic trading opportunities can further amplify the negative impact of competition on welfare by further suppressing the informativeness of stock prices when competition intensifies.

5.6 Additional Robustness Analysis

We discuss the robustness of our main results in three additional extensions, including the cost of information acquisition, the risk attitude of speculators, and firm heterogeneity.

First, our results depend on how product market competition affects speculators' costs of information acquisition. If increased competition raises these costs, reducing competition (e.g., via horizontal mergers) would encourage greater information production in the stock

market and generate gains in allocative efficiency through the feedback mechanism. In contrast, if intensified competition reduces these information costs, horizontal mergers would suppress information production and harm allocative efficiency. However, empirical findings by Farboodi et al. (2022) indicate that information production is more active among larger firms. Since intensified competition shrinks the size of firms, information acquisition costs likely increase with competition. Hence, horizontal mergers—by decreasing competition and increasing firm size—should lower these costs, increase information production, and strengthen the feedback effect. Consequently, horizontal mergers are more likely to produce positive welfare effects when this feedback mechanism is present.

Second, traders and speculators are often risk-averse in practice, but our main findings remain valid. Increased competition raises firms' risks, discouraging risk-averse speculators from entering the market. This further reduces information production compared to the risk-neutral scenario, significantly harming real efficiency.

Third, we focus on symmetric Cournot competition and abstract from firm heterogeneity and synergies typically emphasized in merger analyses, where welfare effects depend on balancing market concentration (higher prices due to reduced competition) against operating efficiencies (cost reductions from synergies). For example, merging firms with complementary strengths, such as lower production costs and superior distribution, can create synergies that improve efficiency. Similarly, synergies can also be achieved through shared technologies or improved management practices. However, the main insights should extend to scenarios with firm heterogeneity and synergies. After a horizontal merger, reduced competition improves information production in the stock market. Since managers often misestimate market conditions, more informative stock prices help them correct biases and improve decisions. Therefore, stock market feedback provides an additional important channel that affects the welfare implications of horizontal mergers.

6 Conclusion

By incorporating information production and learning into a standard Cournot game, we analyze the interaction between product market competition and informational feedback in financial markets. Although intensified competition can reduce the concentration of market power and enhance the economic efficiency in production, it also reduces the incentives for speculators to acquire proprietary information on firms' market prospects. Consequently,

a novel trade-off between economic efficiency and informational efficiency emerges endogenously when production decisions depend on the information conveyed in stock prices. Intensified product market competition can discourage information production in the stock market and generate losses in allocative efficiency through feedback effects, thus impacting the positive welfare effects of competition on real efficiency. When the feedback effect of stock prices is sufficiently strong, a negative relationship between product market competition and total welfare can even arise. Our model provides new insights for antitrust regulations in horizontal mergers, and a guidance for future studies exploring the intersection of financial market efficiency and product market competition.

References

Acharya, Viral V and Lasse Heje Pedersen, "Asset pricing with liquidity risk," *Journal of financial Economics*, 2005, *77* (2), 375–410.

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt, "Competition and innovation: An inverted-U relationship," *The Quarterly Journal of Economics*, 2005, *120* (2), 701–728.

Angeletos, George-Marios, Guido Lorenzoni, and Alessandro Pavan, "Wall street and silicon valley: a delicate interaction," *The Review of Economic Studies*, 2023, *90* (3), 1041–1083.

Ashenfelter, Orley C, Daniel S Hosken, and Matthew C Weinberg, "Efficiencies brewed: pricing and consolidation in the US beer industry," *The RAND Journal of Economics*, 2015, *46* (2), 328–361.

Asker, John and Volker Nocke, "Collusion, mergers, and related antitrust issues," in "Handbook of industrial organization," Vol. 5, Elsevier, 2021, pp. 177–279.

Bai, Jennie, Thomas Philippon, and Alexi Savov, "Have financial markets become more informative?," *Journal of Financial Economics*, 2016, *122* (3), 625–654.

Banz, Rolf W, "The relationship between return and market value of common stocks," *Journal of financial economics*, 1981, *9* (1), 3–18.

Björnerstedt, Jonas and Frank Verboven, "Does merger simulation work? Evidence from the Swedish analgesics market," *American Economic Journal: Applied Economics*, 2016, *8* (3), 125–164.

Boleslavsky, Raphael, David L Kelly, and Curtis R Taylor, "Selloffs, bailouts, and feedback: Can asset markets inform policy?," *Journal of Economic Theory*, 2017, *169*, 294–343.

Bond, Philip, Alex Edmans, and Itay Goldstein, "The real effects of financial markets," *Annual Review of Finance and Economics*, 2012, *4* (1), 339–360.

Braguinsky, Serguey, Atsushi Ohyama, Tetsuji Okazaki, and Chad Syverson, "Acquisitions, productivity, and profitability: evidence from the Japanese cotton spinning industry," *American Economic Review*, 2015, *105* (7), 2086–2119.

Chen, Yangyang, Jeffrey Ng, and Xin Yang, "Talk less, learn more: Strategic disclosure in response to managerial learning from the options market," *Journal of Accounting Research*, 2021, *59* (5), 1609–1649.

Cochrane, John H, "Presidential address: Discount rates," *The Journal of finance*, 2011, *66* (4), 1047–1108.

Comerton-Forde, Carole and Tālis J Putniņš, "Stock price manipulation: Prevalence and determinants," *Review of Finance*, 2014, *18* (1), 23–66.

Compte, Olivier, Frederic Jenny, and Patrick Rey, "Capacity constraints, mergers and collusion," *European Economic Review*, 2002, *46* (1), 1–29.

Cournot, Augustine, "Of the competition of producers," *Chapter 7 in Researches into the Mathematical Principles of the Theory of Wealth*, 1838.

Covarrubias, Matias, Germán Gutiérrez, and Thomas Philippon, "From good to bad concentration? US industries over the past 30 years," *NBER Macroeconomics Annual*, 2020, *34* (1), 1–46.

Dou, Winston Wei, Yan Ji, and Wei Wu, "Competition, profitability, and discount rates," *Journal of Financial Economics*, 2021, *140* (2), 582–620.

Dow, James and Gary Gorton, "Stock market efficiency and economic efficiency: Is there a connection?," *The Journal of Finance*, 1997, *52* (3), 1087–1129.

_ , Itay Goldstein, and Alexander Guembel, "Incentives for information production in markets where prices affect real investment," *Journal of the European Economic Association*, 2017, *15* (4), 877–909.

Easley, David, Nicholas M Kiefer, Maureen O'hara, and Joseph B Paperman, "Liquidity, information, and infrequently traded stocks," *The Journal of Finance*, 1996, *51* (4), 1405–1436.

Edmans, Alex, Itay Goldstein, and Wei Jiang, "The real effects of financial markets: The impact of prices on takeovers," *The Journal of Finance*, 2012, *67* (3), 933–971.

_ , _ , and _ , "Feedback effects, asymmetric trading, and the limits to arbitrage," *American Economic Review*, 2015, *105* (12), 3766–3797.

_ , Sudarshan Jayaraman, and Jan Schneemeier, "The source of information in prices and investment-price sensitivity," *Journal of Financial Economics*, 2017, *126* (1), 74–96.

Farboodi, Maryam, Adrien Matray, Laura Veldkamp, and Venky Venkateswaran, "Where has all the data gone?," *The Review of Financial Studies*, 2022, *35* (7), 3101–3138.

Farrell, Joseph and Carl Shapiro, "Horizontal mergers: an equilibrium analysis," *The American Economic Review*, 1990, pp. 107–126.

Fishman, Michael J and Kathleen M Hagerty, "Disclosure decisions by firms and the competition for price efficiency," *The Journal of Finance*, 1989, *44* (3), 633–646.

Foucault, Thierry and Laurent Frésard, "Learning from peers' stock prices and corporate investment," *Journal of Financial Economics*, 2014, *111* (3), 554–577.

_ and _ , "Corporate strategy, conformism, and the stock market," *The Review of Financial Studies*, 2019, *32* (3), 905–950.

Gao, Pingyang and Pierre Jinghong Liang, "Informational feedback, adverse selection, and optimal disclosure policy," *Journal of Accounting Research*, 2013, *51* (5), 1133–1158.

Geurts, Karen and Johannes Van Biesebroeck, "Employment growth following takeovers," *The RAND Journal of Economics*, 2019, *50* (4), 916–950.

Glosten, Lawrence R and Paul R Milgrom, "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders," *Journal of financial economics*, 1985, *14* (1), 71–100.

Goldstein, Itay, "Information in financial markets and its real effects," *Review of Finance*, 2023, *27* (1), 1–32.

_ and Alexander Guembel, "Manipulation and the allocational role of prices," *The Review of Economic Studies*, 2008, *75* (1), 133–164.

_ **and Liyan Yang**, "Good disclosure, bad disclosure," *Journal of Financial Economics*, 2019, *131* (1), 118–138.

_ **, Emre Ozdenoren, and Kathy Yuan**, "Trading frenzies and their impact on real investment," *Journal of Financial Economics*, 2013, *109* (2), 566–582.

_ **, Yan Li, and Liyan Yang**, "Speculation and hedging in segmented markets," *The Review of Financial Studies*, 2014, *27* (3), 881–922.

Grossman, Sanford J and Joseph E Stiglitz, "On the impossibility of informationally efficient markets," *The American economic review*, 1980, *70* (3), 393–408.

Gu, Lifeng, "Product market competition, R&D investment, and stock returns," *Journal of Financial Economics*, 2016, *119* (2), 441–455.

Guesnerie, Roger and Oliver Hart, "Welfare losses due to imperfect competition: asymptotic results for Cournot Nash equilibria with and without free entry," *International Economic Review*, 1985, pp. 525–545.

Han, Bing and Liyan Yang, "Social networks, information acquisition, and asset prices," *Management Science*, 2013, *59* (6), 1444–1457.

Hellwig, Martin F, "On the aggregation of information in competitive markets," *Journal of economic theory*, 1980, *22* (3), 477–498.

Hemphill, C Scott and Nancy L Rose, "Mergers that harm sellers," *Yale Law Journal*, 2017, *127*, 2078.

Hou, Kewei and David T Robinson, "Industry concentration and average stock returns," *The journal of finance*, 2006, *61* (4), 1927–1956.

Huang, Chong and Xiaoqi Xu, "Informed Trading and Product Market Competition," *Available at SSRN 4451871*, 2023.

Jayaraman, Sudarshan and Joanna Shuang Wu, "Is silence golden? Real effects of mandatory disclosure," *The Review of Financial Studies*, 2019, *32* (6), 2225–2259.

Kreps, David M and Jose A Scheinkman, "Quantity precommitment and Bertrand competition yield Cournot outcomes," *The Bell Journal of Economics*, 1983, pp. 326–337.

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

Landes, William M and Richard A Posner, "Market power in antitrust cases," *J. Reprints Antitrust L. & Econ.*, 1997, *27*, 493.

Leland, Hayne E, "Insider trading: Should it be prohibited?," *Journal of political economy*, 1992, *100* (4), 859–887.

Lin, Tse-Chun, Qi Liu, and Bo Sun, "Contractual managerial incentives with stock price feedback," *American Economic Review*, 2019, *109* (7), 2446–2468.

Luo, Yuanzhi, "Do insiders learn from outsiders? Evidence from mergers and acquisitions," *The Journal of Finance*, 2005, *60* (4), 1951–1982.

Maksimovic, Vojislav and Gordon Phillips, "The market for corporate assets: Who engages in mergers and asset sales and are there efficiency gains?," *The Journal of Finance*, 2001, *56* (6), 2019–2065.

Mermelstein, Ben, Volker Nocke, Mark A Satterthwaite, and Michael D Whinston, "Internal versus external growth in industries with scale economies: A computational model of optimal merger policy," *Journal of Political Economy*, 2020, *128* (1), 301–341.

Miller, Nathan H and Matthew C Weinberg, "Understanding the price effects of the MillerCoors joint venture," *Econometrica*, 2017, *85* (6), 1763–1791.

Motta, Massimo and Emanuele Tarantino, "The effect of horizontal mergers, when firms compete in prices and investments," *International Journal of Industrial Organization*, 2021, *78*, 102774.

Nevo, Aviv, "Mergers with differentiated products: The case of the ready-to-eat cereal industry," *The RAND Journal of Economics*, 2000, pp. 395–421.

Nocke, Volker and Michael D Whinston, "Dynamic merger review," *Journal of Political Economy*, 2010, *118* (6), 1200–1251.

_ **and** _ , "Concentration thresholds for horizontal mergers," *American Economic Review*, 2022, *112* (6), 1915–1948.

Pasquariello, Paolo and Clara Vega, "Strategic cross-trading in the US stock market," *Review of Finance*, 2015, *19* (1), 229–282.

Peress, Joel, "Product market competition, insider trading, and stock market efficiency," *The Journal of Finance*, 2010, *65* (1), 1–43.

Polk, Christopher and Paola Sapienza, "The stock market and corporate investment: A test of catering theory," *The Review of Financial Studies*, 2008, *22* (1), 187–217.

Porter, Robert H, "Mergers and coordinated effects," *International Journal of Industrial Organization*, 2020, *73*, 102583.

Röller, Lars-Hendrik, Johan Stennek, and Frank Verboven, "Efficiency gains from mergers," *European Economic Review*, 2001, *5*, 31–127.

Segal, Ilya and Michael D Whinston, "Antitrust in innovative industries," *American Economic Review*, 2007, *97* (5), 1703–1730.

Smith, Adam, "An inquiry into the nature and causes of the wealth of nations: Volume One," in "in," London: printed for W. Strahan; and T. Cadell, 1776., 1776.

Spulber, Daniel F, "How do competitive pressures affect incentives to innovate when there is a market for inventions?," *Journal of Political Economy*, 2013, *121* (6), 1007–1054.

Subrahmanyam, Avanidhar and Sheridan Titman, "The going-public decision and the development of financial markets," *The Journal of Finance*, 1999, *54* (3), 1045–1082.

Vives, Xavier, "On the efficiency of Bertrand and Cournot equilibria with product differentation," *Journal of Economic Theory*, 1985, *36* (1), 166–175.

Weinberg, Matthew C, "More evidence on the performance of merger simulations," *American Economic Review*, 2011, *101* (3), 51–55.

Werden, Gregory J and Luke M Froeb, "The effects of mergers in differentiated products industries: Logit demand and merger policy," *The Journal of Law, Economics, and Organization*, 1994, *10* (2), 407–426.

Williamson, Oliver E, "Economies as an anti-trust defense: The welfare tradeoffs," *American Economic Review*, 1968, *58* (1), 18–36.

Willner, Johan, "Price leadership and welfare losses in US manufacturing: Comment," *The American Economic Review*, 1989, *79* (3), 604–609.

Xiong, Yan and Liyan Yang, "Disclosure, competition, and learning from asset prices," *Journal of Economic Theory*, 2021, *197*, 105331.

Yi, Sang-Seung, "Market structure and incentives to innovate: the case of Cournot oligopoly," *Economics Letters*, 1999, *65* (3), 379–388.

Appendix

A Proofs of Lemmas and Propositions

A.1 Proof of Lemma 1

Proof. We first compute the beliefs of the market makers. Recall that the total order flow for the ith stock is $f_i = \alpha_i(2\theta - 1) * (\mathbb{1}(\{\omega = H\}) - \mathbb{1}(\{\omega = L\})) + z_i$.[20] Denote $\gamma_i = 1 - \alpha_i(2\theta - 1)$. Note that condition $f_i > \gamma_i$ contradicts the event that $\omega = L$ because: (1) $f_i = z_i + x_i$ by definition; (2) $x_i = -\alpha_i(2\theta - 1)$ if $\omega = L$ by the law of large numbers; and (3) $z_i \leq 1$. Conversely, when $z_i > \gamma_i - \alpha_i(2\theta - 1)$ and $\omega = H$, then $f_i > \gamma_i$. Therefore, the aggregate order flow f_i is a sufficient statistic to update the beliefs of the market makers. In summary, if the aggregate order flow satisfies $f_i > \gamma_i$, it can be inferred that $\omega = H$. Similarly, if the aggregate order flow of stock i is $f_i < -\gamma_i$, the market makers will infer that $\omega = L$. Furthermore, when the aggregate order flow satisfies $f_i \in (-\gamma_i, \gamma_i)$, an application of the Bayes rule implies that

$$\Pr\left(\omega = H \mid f_i \in (-\gamma_i, \gamma_i)\right) = \frac{\Pr(\omega = H)\Pr\left(f_i \in (-\gamma_i, \gamma_i) \mid \omega = H\right)}{\Pr\left(f_i \in (-\gamma_i, \gamma_i)\right)} = \frac{1}{2}$$

because $\Pr\left(f_i \in (-\gamma_i, \gamma_i) \mid \omega = H\right) = \Pr\left(-\gamma_i - \alpha_i(2\theta - 1) \leq z_i \leq \gamma_i - \alpha_i(2\theta - 1)\right) = \gamma_i$ and $\Pr\left(f_i \in (-\gamma_i, \gamma_i)\right) = \Pr\left(f_i \in (-\gamma_i, \gamma_i), \omega = H\right) + \Pr\left(f_i \in (-\gamma_i, \gamma_i), \omega = L\right) = \gamma_i$. This also means that an order flow such that $f_i \in [-\gamma_i, \gamma_i]$ is uninformative.

Second, we analyze the belief updating rule for the ith manager, given the equilibrium prices $\{s_i(f_i)\}_{1 \leq i \leq n}$. Specifically, when $s_i(f_i) = s_H$ is observed, the manager i infers that $f_i > \gamma_i$ and thus $\omega = H$, which is exactly the reason for the market makers. Similarly, when $s_i(f_i) = s_L$ is observed, it can be inferred that $f_i < -\gamma_i$ and thus $\omega = L$. Finally, when $s_i(f_i) = s_M^i$, it must be the case that $f_i \in (-\gamma_i, \gamma_i)$, implying that the ith firm stock price is not informative about the market prospects. The ith manager depends on all other firms' stock prices to infer about the state, and there are three cases, including: (i) there exists some $j \neq i$ such that $s_j = s_H$, then again $f_j > \gamma_j$ and thus $\omega = H$; (ii) if there exists some $j \neq i$ such that $s_j = s_L$, then $f_j < -\gamma_j$ and thus $\omega = L$; (iii) if for all $j \neq i$ such that $s_j = s_M^j$, then it can be inferred that all stock prices are uninformative.

Next, we analyze the ith firm's production strategy, given the manager's posterior belief on the state ω after observing stock prices. Let θ_m be the posterior probability of $\omega = H$. Then, the ith manager's problem is to choose the quantity q_i to maximize:

$$V_i(q_i) = \mathbb{E}\left[TP_i(q_i) \mid \theta_m\right] = q_i\left(A_m - b\left(q_i + q_{-i}\right) - MC\right) \tag{A.1}$$

where $A_m = \mathbb{E}[\tilde{A}|\mathcal{F}_m] = \theta_m A_H + (1 - \theta_m) A_L$ is the expected value of the market prospect A conditional on posterior belief. From Equation (A.1), we know that $V_i(q_i)$ is concave in q_i, and thus $q_i^*(q_{-i}) = \frac{1}{2b}(A_m - bq_{-i} - MC)$. Given a common posterior belief updating rule, we can invoke $q_i = q_j$ for any $i \neq j$. Therefore, $q_i^* = \frac{A_m - MC}{(n+1)b}$.

[20] $\mathbb{1}(\{x \in A\})$ is an indicator function that equals one only when $x \in A$ holds, and equals zero otherwise.

Denote $q_H = \frac{A_H - MC}{(n+1)b}, q_L = \frac{A_L - MC}{(n+1)b}$, and $\beta_i = \prod_{j \neq i} \gamma_j$. Then, combining the belief updating rule of the common posterior, we conclude: (1) if $s_j = s_H$ for some j, then $\theta_m = 1$, $A_m = A_H$ and $q_i^* = q_H$; (2) if $s_j = s_L$ for some j, then $\theta_m = 0$, $A_m = A_L$ and $q_i^* = q_L$; and (3) if $s_j = s_M^j$ for all $1 \leq j \leq n$, then $\theta_m = \frac{1}{2}$, $A_m = \bar{A}$ and $q_i^* = q_M$.

We now check that the stock price rule $s_i(f_i)$ in Equation (6) satisfies condition (4). First, when the total order flow of the ith stock satisfies $f_i > \gamma_i$, then $\omega = H$, and thus $q_i^* = q_H$. By Equations (1) and (2), $\mathbb{E}\left[V_i\left(q_i^*\right) \mid f_i\right] = \frac{(A_H - MC)^2}{(n+1)^2 b}$, which is equal to s_H. Thus, condition (4) is satisfied when $f_i > \gamma_i$. Second, when the total order flow satisfies $f_i < -\gamma_i$, the net demand for the ith stock reveals that $\omega = L$, and thus $q_i^* = q_L$. Hence, $\mathbb{E}\left[V_i\left(q_i^*\right) \mid f_i\right] = \frac{(A_L - MC)^2}{(n+1)^2 b}$ for $f_i < -\gamma_i$, which is equal to s_L. Thus, for $f_i < -\gamma_i$, condition (4) is satisfied.

Third, when $f_i \in (-\gamma_i, \gamma_i)$, the investor demand for the ith stock is not informative about the state, i.e., $\Pr\left(\omega = H \mid f_i \in (-\gamma_i, \gamma_i)\right) = \frac{1}{2}$. Furthermore, by the argument of common posterior belief above, the manager i will produce q_H if $s_j = s_H$ for some $j \neq i$, produce q_L if $s_j = s_L$ for some $j \neq i$, and produce q_M if $s_j = s_M^j$ for all $j \neq i$. Thus, given that $f_i \in (-\gamma_i, \gamma_i)$ and $\exists j \neq i : s_j = s_H$, the ith firm's total profit at time $t = 1$ from producing q_H is

$$TP_H = \frac{(A_H - MC)^2}{(n+1)^2 b}$$

When $f_i \in (-\gamma_i, \gamma_i)$ and $\exists j \neq i : s_j = s_L$, firm i 's total profit from producing q_L is

$$TP_L = \frac{(A_L - MC)^2}{(n+1)^2 b}.$$

When $f_i \in (-\gamma_i, \gamma_i)$ and $s_j = s_M^j$ for $\forall j \neq i$, we deduce that: (1) if $\omega = H$, firm i 's total profit in $t = 1$ from producing q_M is

$$TP_{MH} = \frac{(n+1)\left(\bar{A} - MC\right)(A_H - MC) - n\left(\bar{A} - MC\right)^2}{(n+1)^2 b};$$

and (2) if $\omega = L$, firm i 's total profit in $t = 1$ from producing q_M is

$$TP_{ML} = \frac{(n+1)\left(\bar{A} - MC\right)(A_L - MC) - n\left(\bar{A} - MC\right)^2}{(n+1)^2 b}.$$

Furthermore, by Equation (2), we obtain the following.

$$\begin{aligned}
\mathbb{E}\left[V_i\left(q_i^*\right) \mid f_i \in (-\gamma_i, \gamma_i)\right] &= \Pr\left(\exists j \neq i : s_j = s_H \mid f_i \in (-\gamma_i, \gamma_i)\right) \times TP_H \\
&+ \Pr\left(\exists j \neq i : s_j = s_L \mid f_i \in (-\gamma_i, \gamma_i)\right) \times TP_L \\
&+ \Pr\left(\forall j \neq i : s_j = s_M^j, \omega = H \mid f_i \in (-\gamma_i, \gamma_i)\right) \times TP_{MH} \\
&+ \Pr\left(\forall j \neq i : s_j = s_M^j, \omega = L \mid f_i \in (-\gamma_i, \gamma_i)\right) \times TP_{ML}.
\end{aligned}$$

To compute $\mathbb{E}\left[V_i\left(q_i^*\right) | f_i \in (-\gamma_i, \gamma_i)\right]$, we first calculate the conditional probabilities. Applying

the Bayes rule, we get:

$$\Pr\left(\exists j \neq i : s_j = s_H \mid f_i \in (-\gamma_i, \gamma_i)\right) = \frac{\Pr\left(\exists j \neq i : s_j = s_H, f_i \in (-\gamma_i, \gamma_i)\right)}{\Pr\left(f_i \in (-\gamma_i, \gamma_i)\right)}. \tag{A.2}$$

Using the law of total probability, we have

$$\begin{aligned}\Pr\left(\exists j \neq i : s_j = s_H, f_i \in (-\gamma_i, \gamma_i)\right) &= \Pr\left(\exists j \neq i : s_j = s_H, f_i \in (-\gamma_i, \gamma_i), \omega = H\right) \\ &+ \Pr\left(\exists j \neq i : s_j = s_H, f_i \in (-\gamma_i, \gamma_i), \omega = L\right)\end{aligned}$$

Note that $\Pr(\exists j \neq i : s_j = s_H, f_i \in (-\gamma_i, \gamma_i), \omega = L) = 0$ and that

$$\begin{aligned}\Pr\left(\exists j \neq i : s_j = s_H, f_i \in (-\gamma_i, \gamma_i), \omega = H\right) &= \Pr(\omega = H) \times \Pr\left(f_i \in (-\gamma_i, \gamma_i) \mid \omega = H\right) \\ &\times \Pr\left(\exists j \neq i : s_j = s_H \mid f_i \in (-\gamma_i, \gamma_i), \omega = H\right) = \frac{1}{2}\left(1 - \beta_i\right)\gamma_i\end{aligned}$$

Thus, $\Pr\left(\exists j \neq i : s_j = s_H, f_i \in (-\gamma_i, \gamma_i)\right) = \frac{1}{2}(1 - \beta_i)\gamma_i$.
Plugging this into Equation (A.2), we obtain: $\Pr\left(\exists j \neq i : s_j = s_H \mid f_i \in (-\gamma_i, \gamma_i)\right) = \frac{1}{2}\left(1 - \beta_i\right)$.
Analogously, we can show: $\Pr\left(\exists j \neq i : s_j = s^L \mid f_i \in (-\gamma_i, \gamma_i)\right) = \frac{1}{2}\left(1 - \beta_i\right)$ and

$$\begin{aligned}&\Pr(\forall j \neq i : s_j = s_M^j, \omega = H | f_i \in (-\gamma_i, \gamma_i)) \\ &\qquad = \Pr\left(\forall j \neq i : s_j = s_M^j, \omega = L \mid f_i \in (-\gamma_i, \gamma_i)\right) = \frac{1}{2}\beta_i\end{aligned}$$

Finally, plugging in these conditional probabilities, we have:

$$\mathbb{E}\left[V_i\left(q_i^*\right) \mid f_i \in (-\gamma_i, \gamma_i)\right] = \frac{2\left(\left(A_H - MC\right)^2 + \left(A_L - MC\right)^2\right) - \beta_i\left(A_H - A_L\right)^2}{4(n+1)^2 b}$$

which is equal to s_M^i. Therefore, condition (4) is satisfied for $f_i \in [-\gamma_i, \gamma_i]$. The proof concludes. □

A.2 Proof of Lemma 2

Proof. Let $\Pi_i\left(x_k^i, m_k^i\right)$ be the expected profit of the speculator k who trades $x_k^i \in [-1, 1]$ shares of the ith firm when his signal is m_k^i, and let V_2^i be the market value of the ith firm at $t = 1$. Since each speculator is risk neutral and a price taker in the stock market, speculators will trade the maximum size possible if they acquire information, i.e., $x_k^i = \pm 1$.

First, consider an informed speculator who observes $m_k^i = H$. If he buys the asset, his expected profit is $\Pi_k^i(+1, H) = \mathbb{E}\left[V_2^i - s_i\left(f_i\right) \mid m_k^i = H, x_k^i = 1\right]$.

From the proof of Lemma 1, firm i 's value at $t = 1$ is

$$V_2^i = \begin{cases} TP_H & \text{if} \quad \exists j \in \{1, \ldots, n\} \text{ such that } s_j = s_H; \\ TP_{MH} & \text{if} \quad \omega = H \;\&\; s_j = s_M^j, \forall j \in \{1, \ldots, n\}; \\ TP_L & \text{if} \quad \exists j \in \{1, \ldots, n\} \text{ such that } s_j = s_L; \\ TP_{ML} & \text{if} \quad \omega = L \;\&\; s_j = s_M^j, \forall j \in \{1, \ldots, n\}. \end{cases} \tag{A.3}$$

Thus, using Equation (A.3), we can calculate $\Pi_i(+1, H)$ as follows:

$$\begin{aligned}\Pi_i(+1,H) = &\Pr\left(\omega = H, f_i > \gamma_i \mid m_k^i = H\right) \times (TP_H - s_H) \\ &+ \Pr\left(\omega = L, f_i < -\gamma_i \mid m_k^i = H\right) \times (TP_L - s_L) \\ &+ \Pr\left(\omega = H, f_i \in (-\gamma_i, \gamma_i), \exists j \neq i : s_j = s_H \mid m_k^i = H\right) \times \left(TP_H - s_M^i\right) \\ &+ \Pr\left(\omega = H, f_i \in (-\gamma_i, \gamma_i), \forall j \neq i : s_j = s_M^j \mid m_k^i = H\right) \times \left(TP_{MH} - s_M^i\right) \\ &+ \Pr\left(\omega = L, f_i \in (-\gamma_i, \gamma_i), \exists j \neq i : s_j = s_L \mid m_k^i = H\right) \times \left(TP_L - s_M^i\right) \\ &+ \Pr(\omega = L, f_i \in (-\gamma_i, \gamma_i), \forall j \neq i : s_j = s_M^j \mid m_k^i = H\Big) \times \left(TP_{ML} - s_M^i\right).\end{aligned}$$

Since $s^H = TP_H$ and $s^L = TP_L$, we can rewrite the expression of $\Pi_i(+1, H)$ as:

$$\begin{aligned}\Pi_i(+1,H) &= \Pr\left(\omega = H, f_i \in (-\gamma_i, \gamma_i), \exists j \neq i : s_j = s_H \mid m_k^i = H\right) \times \left(TP_H - s_M^i\right) \\ &+ \Pr\left(\omega = H, f_i \in (-\gamma_i, \gamma_i), \forall j \neq i : s_j = s_M^j \mid m_k^i = H\right) \times \left(TP_{MH} - s_M^i\right) \\ &+ \Pr\left(\omega = L, f_i \in (-\gamma_i, \gamma_i), \exists j \neq i : s_j = s_L \mid m_k^i = H\right) \times \left(TP_L - s_M^i\right) \\ &+ \Pr(\omega = L, f_i \in (-\gamma_i, \gamma_i), \forall j \neq i : s_j = s_M^j \mid m_k^i = H\Big) \times \left(TP_{ML} - s_M^i\right).\end{aligned}$$

Now, we use the Bayes rule to calculate $\Pr\left(\omega = H, f_i \in (-\gamma_i, \gamma_i), \exists j \neq i : s_j = s_H \mid m_k^i = H\right)$.

$$\begin{aligned}\Pr(\omega = H, f_i \in (-\gamma_i, \gamma_i), \exists j \neq i : s_j = s_H \mid m_k^i = H) &= \frac{1}{\Pr\left(m_k^i = H\right)} \times \Pr(\omega = H) \\ \times \Pr\left(f_i \in (-\gamma_i, \gamma_i) \mid \omega = H\right) &\times \Pr\left(\exists j \neq i : s_j = s_H \mid \omega = H, f_i \in (-\gamma_i, \gamma_i)\right) \\ \times \Pr\left(m_k^i = H \mid \omega = H, f_i \in (-\gamma_i, \gamma_i), \exists j \neq i : s_j = s_H\right)\Big) &= \theta\gamma_i\left(1 - \beta_i\right)\end{aligned}$$

We have used the following facts in the last equation, including:

$$\begin{aligned}&\Pr\left(\exists j \neq i : s_j = s_H \mid \omega = H, f_i \in (-\gamma_i, \gamma_i)\right) = \Pr\left(\exists j \neq i : s_j = s_H \mid \omega = H\right) = 1 - \beta_i; \\ &\Pr(m_k^i = H \mid \omega = H, f_i \in (-\gamma_i, \gamma_i), \exists j \neq i : s_j = s_H = \Pr(m_k^i = H \mid \omega = H) = \theta; \\ &\Pr(m_k^i = H) = \sum_{\omega \in \{H,L\}} \Pr(\omega) \Pr(m_k^i = H \mid \omega) = \frac{1}{2}.\end{aligned}$$

Similarly, we have:

$$\begin{aligned}&\Pr\left(\omega = H, f_i \in (-\gamma_i, \gamma_i), \forall j \neq i : s_j = s_M^j \mid m_k^i = H\right) = \theta\gamma_i\beta_i; \\ &\Pr\left(\omega = L, f_i \in (-\gamma_i, \gamma_i), \exists j \neq i : s_j = s_L \mid m_k^i = H\right) = \gamma_i(1-\theta)\left(1 - \beta_i\right); \\ &\Pr\left(\omega = L, f_i \in (-\gamma_i, \gamma_i), \forall j \neq i : s_j = s_M^j \mid m_k^i = H\right) = \gamma_i\beta_i(1-\theta).\end{aligned}$$

Plugging these conditional probabilities back into the formula of $\Pi_i(+1, H)$, we have:

$$\Pi_i(+1,H) = \frac{(2\theta - 1)\gamma_i(2 + \beta_i(n-1))\left((A_H - MC)^2 - (A_L - MC)^2\right)}{4(n+1)^2 b} > 0$$

If instead the speculator sells, his expected profit is

$$\Pi_i(-1,H) = -\frac{(2\theta-1)\gamma_i(2+\beta_i(n-1))\left((A_H-MC)^2-(A_L-MC)^2\right)}{4(n+1)^2 b} < 0$$

Thus, the optimal trading strategy is to buy $x_k^i = +1$ when $m_k^i = H$.

Symmetric reasoning shows that the speculator's optimal trading strategy is to sell $x_k^i = +1$ when $m_k^i = L$. And in this case, his trading profit satisfies $\Pi_i(-1,L) = \Pi_i(+1,H)$. Furthermore, since $(A_H - MC)^2 - (A_L - MC)^2 = 2\left(\bar{A} - MC\right)(A_H - A_L)$, we conclude that

$$\Pi_i = \frac{(2\theta-1)\gamma_i\left(2+(n-1)\beta_i\right)\left(\bar{A}-MC\right)(A_H-A_L)}{2(n+1)^2 b}.$$

The proof concludes. □

A.3 Proof of Proposition 1

Proof. By Equation (8), $\frac{\partial \Pi(\alpha)}{\partial \alpha} < 0$. Thus, $\Pi(0) > \Pi(\alpha) > \Pi(1)$ for all $\alpha \in (0,1)$. Furthermore, by definition, we have: (i) when $c \geq \Pi(0) =: \bar{c}$, $\Pi(\alpha) < 0$ for any $\alpha > 0$, and thus $\widehat{\alpha} = 0$; (ii) when $c \leq \Pi(1) =: \underline{c}$, $\Pi(\alpha) < 0$ for any $\alpha < 1$, and thus $\widehat{\alpha} = 1$; and (iii) when $c \in (\underline{c}, \bar{c})$, by the intermediate value theorem and $\Pi(0) - c > 0 > \Pi(1) - c$, there exists a solution $\widehat{\alpha}$ such that $\Pi(\widehat{\alpha}) = c$, which is also unique since $\Pi'(\alpha) < 0$. □

A.4 Proof of Proposition 2

Proof. First, we can use Equation (8) to calculate the partial derivatives:

$$\frac{\partial \Pi\left(n,\widehat{\alpha}_n\right)}{\partial \widehat{\alpha}_n} = -\frac{(2\theta-1)^2\left(2+n(n-1)\widehat{\gamma}^{n-1}\right)\left(\bar{A}-MC\right)(A_H-A_L)}{2b(n+1)^2}$$

$$\frac{\partial \Pi\left(n,\widehat{\alpha}_n\right)}{\partial n} = -\frac{\widehat{\gamma}_n(2\theta-1)\left(\bar{A}-MC\right)(A_H-A_L)}{2b(n+1)^3}\left\{4+\widehat{\gamma}_n^{n-1}\left(n-3+\left(n^2-1\right)\ln\frac{1}{\widehat{\gamma}_n}\right)\right\}$$

where $\widehat{\gamma}_n = 1 - \widehat{\alpha}_n(2\theta-1)$.

By the implicit function theorem, we further have:

$$\begin{aligned}\frac{\partial \widehat{\alpha}_n}{\partial n} &= -\left(\frac{\partial \Pi\left(n,\widehat{\alpha}_n\right)}{\partial n}\right)\Bigg/\left(\frac{\partial \Pi\left(n,\widehat{\alpha}_n\right)}{\partial \widehat{\alpha}_n}\right) \\ &= -\frac{{\widehat{\gamma}_n}^n}{(2\theta-1)(n+1)\left(2+n(n-1){\widehat{\gamma}_n}^{n-1}\right)}\left(4\widehat{\gamma}_n^{1-n}+n-3+(n+1)(n-1)\ln\frac{1}{\widehat{\gamma}_n}\right) \qquad \text{(A.4)}\end{aligned}$$

Obviously, when $n \geq 3$, it is easy to verify that $\frac{\partial \widehat{\alpha}_n}{\partial n} < 0$. Furthermore, we next show that $\frac{\partial \widehat{\alpha}_n}{\partial n} < 0$ holds when $n = 2$. Plugging in $n = 2$, it yields:

$$\frac{\partial \widehat{\alpha}_n}{\partial n}\Bigg| n = 2 = -\frac{\widehat{\gamma}_2^2}{6(2\theta-1)\left(1+\widehat{\gamma}_2\right)}\left(4\widehat{\gamma}_2^{-1}+3\ln\frac{1}{\widehat{\gamma}_2}-1\right)$$

Since $0 \leq \widehat{\gamma}_n = 1 - \widehat{\alpha}_n(2\theta-1) \leq 1$, the result follows. The proof concludes. □

A.5 Proof of Corollary 1

Proof. We first show that $\frac{\partial \widehat{\alpha}_n}{\partial A_H} > 0$. Applying the implicit function theorem implies:

$$\frac{\partial \widehat{\alpha}_n}{\partial A_H} = -\left(\frac{\partial \Pi\left(\widehat{\alpha}_n\right)}{\partial A_H}\right) \Big/ \left(\frac{\partial \Pi\left(\widehat{\alpha}_n\right)}{\partial \widehat{\alpha}_n}\right)$$

We have already shown in the proof of Proposition 2 that $\frac{\partial \Pi(\widehat{\alpha}_n)}{\partial \widehat{\alpha}_n} < 0$. Hence, it suffices to show that $\frac{\partial \Pi(\widehat{\alpha}_n)}{\partial A_H} > 0$. Again, Using Equation (8), we obtain:

$$\frac{\partial \Pi\left(\widehat{\alpha}_n\right)}{\partial A_H} = \frac{2\widehat{\gamma}_n(2\theta-1)\left(A_H - MC\right)\left(2+(n-1)\widehat{\gamma}_n{}^{n-1}\right)}{4b(n+1)^2} > 0$$

Similarly, we can show that:

$$\begin{aligned}
\frac{\partial \Pi\left(\widehat{\alpha}_n\right)}{\partial A_L} &= -\frac{2\widehat{\gamma}_n(2\theta-1)\left(A_L - MC\right)\left(2+(n-1)\widehat{\gamma}_n{}^{n-1}\right)}{4b(n+1)^2} < 0, \\
\frac{\partial \Pi\left(\widehat{\alpha}_n\right)}{\partial MC} &= -\frac{\widehat{\gamma}_n(2\theta-1)\left(A_H - A_L\right)\left(2+(n-1)\widehat{\gamma}_n{}^{n-1}\right)}{2b(n+1)^2} < 0, \\
\frac{\partial \Pi\left(\widehat{\alpha}_n\right)}{\partial b} &= -\frac{\widehat{\gamma}_n(2\theta-1)\left(\bar{A} - MC\right)\left(A_H - A_L\right)\left(2+(n-1)\widehat{\gamma}_n{}^{n-1}\right)}{2b^2(n+1)^2} < 0.
\end{aligned}$$

Hence, $\frac{\partial \widehat{\alpha}_n}{\partial A_L} < 0$, $\frac{\partial \widehat{\alpha}_n}{\partial MC} < 0$, and $\frac{\partial \widehat{\alpha}_n}{\partial b} < 0$. The proof concludes. □

A.6 Derivation of Equation (14) and (15)

From Lemma 1 and Equation (12), we can calculate total welfare at $t = 1$ as

$$W = \begin{cases} W_H & \text{if } s_i = s_H \text{ for some } i \in \{1, \ldots, n\}; \\ W_{MH} & \text{if } \omega = H \ \&\ s_i = s_M^i \ \forall i \in \{1, \ldots, n\}; \\ W_{ML} & \text{if } \omega = L \ \&\ s_i = s_M^i \ \forall i \in \{1, \ldots, n\}; \text{and} \\ W_L & \text{if } s_i = s_L \text{ for some } i \in \{1, \ldots, n\}. \end{cases}$$

where $W_H = \frac{n(n+2)(A_H - MC)^2}{2b(n+1)^2}$, $W_{MH} = \frac{n\left(\bar{A}-MC\right)\left((2n+4)(A_H - MC) + n(A_H - A_L)\right)}{4b(n+1)^2}$, $W_L = \frac{n(n+2)(A_L - MC)^2}{2b(n+1)^2}$, and $W_{ML} = \frac{n\left(\bar{A}-MC\right)\left((2n+4)(A_L - MC) + n(A_L - A_H)\right)}{4b(n+1)^2}$.

Then, the expected total welfare is given by:

$$\begin{aligned}
\overline{W} = &\Pr\left(\exists i : s_i = s_H\right) \times W_H + \Pr\left(\forall i : s_i = s_M^i, \omega = H\right) \times W_{MH} \\
&+ \Pr\left(\exists i : s_i = s_L\right) \times W_L + \Pr\left(\forall i : s_i = s_M^i, \omega = L\right) \times W_{ML}
\end{aligned}$$

From the proof of Lemma 1, we already know that $f_i > \widehat{\gamma}_n$ (i.e., $s_i = s_H$) is impossible when $\omega = L$ and $f_i < \widehat{\gamma}_n$ (i.e., $s_i = s_L$) is impossible when $\omega = H$. Hence, we have:

$$\begin{aligned}
\overline{W} = &\Pr\left(\exists i : s_i = s_H, \omega = H\right) \times W_H + \Pr\left(\forall i : s_i = s_M^i, \omega = H\right) \times W_{MH} \\
&+ \Pr\left(\exists i : s_i = s_L, \omega = L\right) \times W_L + \Pr\left(\forall i : s_i = s_M^i, \omega = L\right) \times W_{ML}
\end{aligned}$$

To compute $\overline{W}$, we use the Bayes rule to calculate $\Pr(\exists i : s_i = s_H, \omega = H)$.

$$\Pr\left(\exists i : s_i = s_H, \omega = H\right) = \Pr(\omega = H)\Pr\left(\exists i : s_i = s_H \mid \omega = H\right)$$

Using the expression of $s_i(f_i)$ in Equation (6), we know:

$$\begin{aligned}\Pr\left(s_i = s_M^i \mid \omega = H\right) &= \Pr\left(-\widehat{\gamma}_n \leq f_i \leq \widehat{\gamma}_n \mid \omega = H\right) = \widehat{\gamma}_n \\ \Pr\left(s_i = s_H \mid \omega = H\right) &= \Pr\left(f_i > \widehat{\gamma}_n \mid \omega = H\right) = 1 - \widehat{\gamma}_n\end{aligned}$$

and thus: $\Pr\left(\exists i : s_i = s_H \mid \omega = H\right) = 1 - \Pr\left(\forall i : s_i = s_M^i \mid \omega = H\right) = 1 - (\widehat{\gamma}_n)^n$.

Since $\Pr(\omega = H) = 1/2$, we further have:

$$\Pr\left(\exists i : s_i = s_H, \omega = H\right) = \frac{1 - (\widehat{\gamma}_n)^n}{2}$$

Similarly, we have

$$\begin{aligned}&\Pr\left(\exists i : s_i = s_L, \omega = L\right) = \frac{1 - (\widehat{\gamma}_n)^n}{2}, \\ &\Pr\left(\forall i : s_i = s_M^i, \omega = H\right) = \Pr\left(\forall i : s_i = s_M^i, \omega = L\right) = \frac{(\widehat{\gamma}_n)^n}{2}\end{aligned}$$

Therefore, $\overline{W}$ can be written as

$$\overline{W}(\widehat{\alpha}_n, n) = \frac{n(n+2)}{8(n+1)^2 b}\left(4\left(\bar{A} - MC\right)^2 + \left(1 - (\widehat{\gamma}_n)^n\right)\left(A_H - A_L\right)^2\right)$$

Obviously, $\overline{W}$ depends on n and $\widehat{\alpha}_n$, which implicitly depends on n, and we can explicitly write: $\overline{W}(\widehat{\alpha}_n, n)$. Given the monotone relationship between $\widehat{\alpha}_n$ and n, we know that the expected total welfare is uniquely determined for any fixed n.

Last, note that we can show for the formula of $\overline{CS}(\widehat{\alpha}_n, n)$ in a similar way. Again, from Lemma 1 and Equation (12), we can calculate consumer surplus at $t = 1$ as

$$CS = \begin{cases} CS_H & \text{if } s_i = s_H \text{ for some } i \in \{1, \ldots, n\}; \\ CS_{MH} & \text{if } \omega = H \;\&\; s_i = s_M^i \;\forall i \in \{1, \ldots, n\}; \\ CS_{ML} & \text{if } \omega = L \;\&\; s_i = s_M^i \;\forall i \in \{1, \ldots, n\}; \text{and} \\ CS_L & \text{if } s_i = s_L \text{ for some } i \in \{1, \ldots, n\}. \end{cases}$$

where $CS_H = \frac{n^2(A_H - MC)^2}{2b(n+1)^2}$, $CS_L = \frac{n^2(A_L - MC)^2}{2b(n+1)^2}$, and $CS_{MH} = CS_{ML} = \frac{n^2\left(\bar{A} - MC\right)^2}{2b(n+1)^2}$

Furthermore, similar to $\bar{W}$, we have:

$$\begin{aligned}\overline{CS} = &\Pr\left(\exists i : s_i = s_H, \omega = H\right) \times CS_H + \Pr\left(\forall i : s_i = s_M^i, \omega = H\right) \times CS_{MH} \\ &+ \Pr\left(\exists i : s_i = s_L, \omega = L\right) \times CS_L + \Pr\left(\forall i : s_i = s_M^i, \omega = L\right) \times CS_{ML}\end{aligned}$$

Thus, $\overline{CS}$ can be calculated as

$$\overline{CS} = \frac{1 - (\widehat{\gamma}_n)^n}{2} \times (CS_H + CS_L) + \frac{(\widehat{\gamma}_n)^n}{2} \times (CS_{MH} + CS_{ML})$$

From the expression of the consumer surplus at $t = 1$, we further have:

$$\overline{CS}\left(\widehat{\alpha}_n, n\right) = \frac{n^2}{8b(n+1)^2}\left(4\left(\bar{A} - MC\right)^2 + \left(1 - \left(\widehat{\gamma}_n\right)^n\right)\left(A_H - A_L\right)^2\right).$$

The derivation concludes.

A.7 Proof of Lemma 3

Proof. **(i) Total welfare.** Based on the expression for $\overline{W}(\widehat{\alpha}_n, n)$ in Equation (14), we know that

$$\frac{d\overline{W}\left(\widehat{\alpha}_n, n\right)}{dn} = \frac{\partial\overline{W}\left(\widehat{\alpha}_n, n\right)}{\partial n} + \frac{\partial\overline{W}\left(\widehat{\alpha}_n, n\right)}{\partial\widehat{\alpha}_n}\frac{\partial\widehat{\alpha}_n}{\partial n}$$

First, the partial derivative of $\overline{W}\left(\widehat{\alpha}_n, n\right)$ with respect to n can be calculated as

$$\begin{aligned}\frac{\partial\overline{W}\left(\widehat{\alpha}_n, n\right)}{\partial n} &= \frac{n(n+2)\left(A_H - A_L\right)^2\left(\widehat{\gamma}_n\right)^n \ln(1/\widehat{\gamma}_n)}{8b(n+1)^2} \\ &+ \frac{1}{4b(n+1)^3}\left(4\left(\bar{A} - MC\right)^2 + \left(1 - \left(\widehat{\gamma}_n\right)^n\right)\left(A_H - A_L\right)^2\right)\end{aligned}$$

Second, we calculate the partial derivative of $\overline{W}\left(\widehat{\alpha}_n, n\right)$ with respect to $\widehat{\alpha}_n$ as follows:

$$\frac{\partial\overline{W}\left(\widehat{\alpha}_n, n\right)}{\partial\widehat{\alpha}_n} = \frac{\left(\widehat{\gamma}_n\right)^{n-1}n^2(n+2)(2\theta-1)\left(A_H - A_L\right)^2}{8b(n+1)^2}.$$

Using Equations (A.4) and the two partial derivatives above, we get:

$$\frac{d\overline{W}\left(\widehat{\alpha}_n, n\right)}{dn} = \frac{\left(A_H - A_L\right)^2}{8b(n+1)^3}\left\{\frac{2\left(4\left(\bar{A} - MC\right)^2 + \left(A_H - A_L\right)^2\right)}{\left(A_H - A_L\right)^2} - g_1\left(\widehat{\alpha}_n, n\right)\right\}$$

Therefore, $\frac{d\overline{W}(\widehat{\alpha}_n, n)}{dn} < 0$ holds if and only if: $g_1\left(\widehat{\alpha}_n, n\right) > \frac{8\left(\bar{A} - MC\right)^2}{\left(A_H - A_L\right)^2} + 2$.

(ii) Consumer surplus. Obviously, $\overline{CS}\left(\widehat{\alpha}_n, n\right) = \frac{n}{n+2}\overline{W}\left(\widehat{\alpha}_n, n\right)$. Thus, the total derivative of $\overline{CS}\left(\widehat{\alpha}_n, n\right)$ with respect to n can be written as follows:

$$\frac{d\overline{CS}\left(\widehat{\alpha}_n, n\right)}{dn} = \frac{n}{n+2} \times \frac{d\overline{W}\left(\widehat{\alpha}_n, n\right)}{dn} + \frac{2}{(n+2)^2} \times \overline{W}\left(\widehat{\alpha}_n, n\right)$$

Recall that $\overline{W}\left(\widehat{\alpha}_n, n\right) = \frac{n(n+2)}{8b(n+1)^2}\left\{4\left(\bar{A} - MC\right)^2 + \left(1 - \left(\widehat{\gamma}_n\right)^n\right)\left(A_H - A_L\right)^2\right\}$ and $\frac{d\overline{W}(\widehat{\alpha}_n, n)}{dn} = \frac{\left(A_H - A_L\right)^2}{8b(n+1)^3}\left(G_1 - g_1\left(\widehat{\alpha}_n, n\right)\right)$. Then, we can calculate $d\overline{CS}\left(\widehat{\alpha}_n, n\right)/dn$ as follows:

$$\frac{d\overline{CS}\left(\widehat{\alpha}_n, n\right)}{dn} = \frac{n\left(A_H - A_L\right)^2}{8b(n+1)^3}\left(G_1 - g_2\left(\widehat{\alpha}_n, n\right)\right)$$

Thus, $\frac{d\overline{CS}(\widehat{\alpha}_n, n)}{dn} < 0$ holds if and only if $g_2\left(\widehat{\alpha}_n, n\right) > G_1$ is true. The proof concludes. □

A.8 Proof of Proposition 3

Proof. The idea is to construct a set U of the information production cost such that for any $c \in U$, we have: (i) $\hat{\alpha}_m = 1$, $\hat{\alpha}_n = 0$; (ii) $n > m$; and (iii) $\overline{W}(\widehat{\alpha}_m, m) > \overline{W}(\widehat{\alpha}_n, n)$. It suffices to show that competition can decrease total welfare through informational feedback when $U \neq \varnothing$, because whenever information production is fixed, an increase in the number of firms always improves total welfare in Cournot competition.

Now, we come to construct U. First, given condition (i),

$$\frac{\overline{W}(\widehat{\alpha}_m, m)}{\overline{W}(\widehat{\alpha}_n, n)} = \frac{\left(1 - \frac{1}{(m+1)^2}\right) * (1 + \mu * (1 - (2-2\theta)^m))}{\left(1 - \frac{1}{(n+1)^2}\right)}$$

Thus, $\overline{W}(\widehat{\alpha}_m, m) > \overline{W}(\widehat{\alpha}_n, n)$ holds whenever $\Phi(m) \geq 1$ is true, since the denominator is always smaller than m for any $n \in \mathbb{N}$.

Second, since $\Phi(m)$ is continuous and strictly increasing in m and that $\lim_{l \to \infty} \Phi(m) = (1+\mu) > 1$, there exists some m_0 sufficiently large such that $\Phi(m) \geq 1$ for all $m \geq m_0$. Fix any m such that $\Phi(m) \geq 1$, and we can define $\underline{c}_m$ by Equation (10).

Third, we can use the floor function $[x] = \{z \in \mathbb{Z} : z \leq x\}$ to define:

$$N(m) = \frac{(m+1)^2}{(2-2\theta)\left(2 + (m-1)(2-2\theta)^{m-1}\right)}$$

By construction, we have $\underline{c}_m > \bar{c}_N$. Therefore, we can define $U = [\bar{c}_n, \underline{c}_m)$ for any $n \geq N$ because $\bar{c}_n$ is strictly decreasing in n. By construction, $U = [\bar{c}_n, \underline{c}_m)$ is the desired set that satifies conditions (i)-(iii). The proof concludes. □

A.9 Proof of Proposition 4

Proof. We prove this result for all parameters one by one.

Case (i): Information production cost c. First, when $c = 0$, $\hat{\alpha}_n = 1$ for all $n \in \mathbb{N}$. Therefore, $n^* \to \infty$. Second, when $c > \bar{c}_1$, then $\hat{\alpha}_n = 0$, and thus $n^* \to \infty$. Then, the non-monotonicity of $n^*(c)$ follows from Corollary A.1 below.

Corollary A.1. *Consider n_1 such that $\Phi(n_1) \geq 1$ and $n_2 \geq N(n_1)$. Then:*
(1) When $c < \underline{c}_{n_2}$ or $c > \bar{c}_{n_1}$, $\overline{W}(\widehat{\alpha}_{n_2}, n_2) > \overline{W}(\widehat{\alpha}_{n_1}, n_1)$; and
(2) When $\bar{c}_{n_2} < c < \underline{c}_{n_1}$, $\overline{W}(\widehat{\alpha}_{n_2}, n_2) < \overline{W}(\widehat{\alpha}_{n_1}, n_1)$.

Note that Corollary A.1 follows directly from Proposition 3.

Case (ii): Price sensitivity b. First, when $b \to \infty$, we have $\Pi(\alpha) \to 0$, which implies that $\hat{\alpha}_n = 0$ for all $n \in \mathbb{N}$ and thus $n^* \to \infty$. Second, when $b \to 0$, then $\hat{\alpha}_n = 1$, and thus $n^* \to \infty$. Then, the non-monotonicity of $n^*(b)$ follows from Corollary A.1. To see it, select positive integers n_1 and n_2 such that: $\Phi(n_1) \geq 1$ and $n_2 \geq N(n_1)$. By Corollary A.1, $n^* < n_2$ when $\bar{c}_{n_2} < c < \underline{c}_{n_1}$, which translates into:

$$\frac{(2\theta-1)(A_H - A_L)(\bar{A} - MC)}{2(n_2+1)c} < b < \frac{(2\theta-1)(1-\theta)(2 + (n_1-1)(2-2\theta)^{n_1-1})(A_H - A_L)(\bar{A} - MC)}{(n_1+1)^2 c}$$

Therefore, n^* is non-monotonic in b.

Case (iii): Market prospect in good state A_H. First, when $A_H \to \infty$, we have $\Pi(\alpha) \to \infty$, which implies that $\hat{\alpha}_n = 1$ for all $n \in \mathbb{N}$ and thus $n^* \to \infty$. Second, when $(A_H - A_L) \to 0$, then $\hat{\alpha}_n = 0$, and thus $n^* \to \infty$. Then, the non-monotonicity of n^* follows from Corollary A.1. To see it, select positive integers n_1 and n_2 such that: $\Phi(n_1) \geq 1$ and $n_2 \geq N(n_1)$. By Corollary A.1, $n^* < n_2$ when $\bar{c}_{n_2} < c < \underline{c}_{n_1}$, which translates into:

$$A_L + \frac{2(n_2+1)bc}{(2\theta-1)(\bar{A}-MC)} > A_H > A_L + \frac{(n_1+1)^2 bc}{(2\theta-1)(1-\theta)(2+(n_1-1)(2-2\theta)^{n_1-1})(\bar{A}-MC)}$$

Thus, $n^* < \infty$ can be finite. Therefore, n^* is non-monotonic in $(A_H - A_L)$. The proof concludes. □

A.10 Proof of Lemma 4

Proof. First, note that by the assumed condition $A_L = MC$, $4(\bar{A} - MC)^2 = (A_H - A_L)^2$. Thus, $\overline{W}(\widehat{\alpha}_1, 1) > \overline{W}(\widehat{\alpha}_2, 2)$ reduces to:

$$\frac{3}{32}(2 - \widehat{\gamma}_1) > \frac{1}{9}(2 - (\widehat{\gamma}_2)^2)$$

Second, when $c \geq \frac{(2\theta-1)(A_H-A_L)^2}{12b}$, by Equation (9), we have: $\widehat{\alpha}_2 = 0$ and thus $\widehat{\gamma}_2 = 1$. This further implies that $\overline{W}(\widehat{\alpha}_1, 1) > \overline{W}(\widehat{\alpha}_2, 2)$ if and only if $\widehat{\gamma}_1 < \frac{22}{27}$.

Finally, note that $\widehat{\gamma}_1$ is governed by Equation (8). Simple algebra yields the bound $c \leq \frac{11}{108}\kappa$. The other condition $c < \frac{(1-\theta)(2-\theta)\kappa}{9}$ follows from the definition of $\underline{c}$ for $n = 1$ and $n = 2$. Indeed, if $c < \min\{\underline{c}_1, \underline{c}_2\}$, then $\widehat{\gamma}_1 = \widehat{\gamma}_2 = 1$, and thus $\overline{W}(\widehat{\alpha}_1, 1) \leq \overline{W}(\widehat{\alpha}_2, 2)$. The proof concludes. □

Online Appendix

B Extended Discussions

B.1 Impacts of feedback effects from stock market: an alterative scenario

Under extreme parameter values, where low market uncertainty reduces the informational value of managerial learning, the stock market feedback effect may not overturn the positive relationship between competition and total welfare. Nonetheless, it can significantly shape the efficiency implications of firm competition, making it a crucial factor in regulating horizontal mergers.

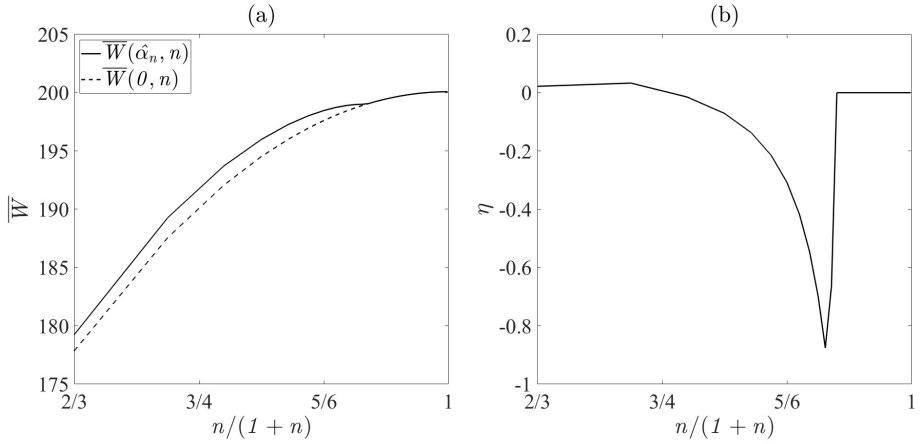


Figure 11: Small Market Uncertainty ($A_L = 25$)

Notes: This figure estimates the total welfare with and without feedback effects, as well as $\eta = \frac{\overline{W}(\widehat{\alpha}_n,n)-\overline{W}(\widehat{\alpha}_{n-1},n-1)}{\overline{W}(0,n)-\overline{W}(0,n-1)} - 1$. A negative value of η indicates that the welfare effect of a horizontal merger will be overestimated if the feedback effect is ignored. A positive value of η then suggests that the feedback effect augments the welfare effect of a horizontal merger.

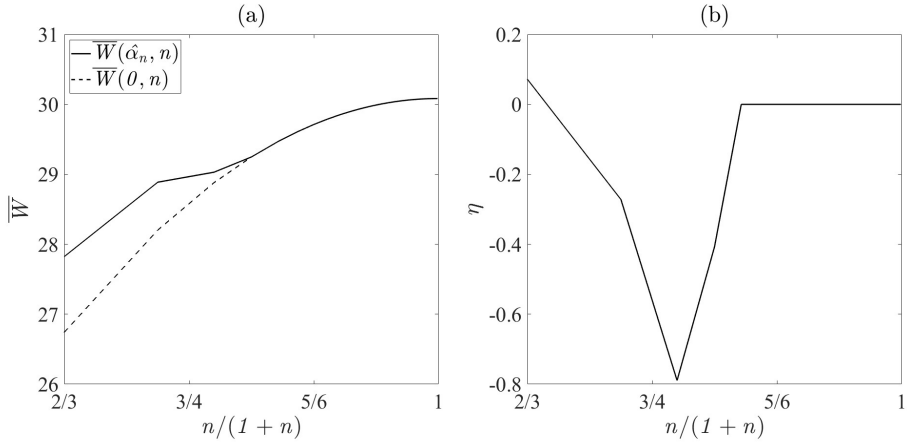
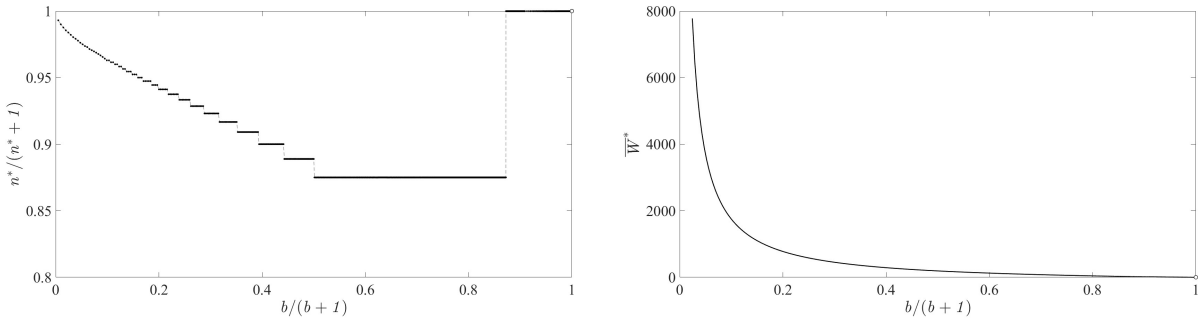


Figure 12: Small Market Uncertainty ($A_H = 15$)

Compared to the baseline model, Figures 11 and 12 adjust the parameter values of A_L from 10 to 25 and A_H from 30 to 15, respectively, while keeping all other parameters unchanged. These modifications are quite extreme, reducing the ratio $\frac{A_H - A_L}{MC}$ by 75%, from $\frac{20}{3}$ to $\frac{5}{3}$. Under these two sets of parameter configurations, the feedback effects are insufficient to reverse the positive relationship between firm competition and total welfare. Nevertheless, the feedback effect continues to exert a significant influence on the efficiency implications of competition. Specifically, when the intensity of firm competition varies, the welfare change without considering feedback effects can be substantially smaller — by as much as 80%.

B.2 An Extended Discussion for Section 4.4

Price sensitivity b. Figure 13 depicts the optimal market structure $n^*/(n^*+1)$ and the corresponding total welfare $W(n^*)$ under the optimal market structure n^*. When b is high, the market price is very sensitive to the quantity of production, reducing profits for the firms and thus discouraging the production of information. Therefore, the information production gap disappears when we vary n, leading to a dominant role of market power concentration. Similarly, when b is low, the market price is insensitive, increasing profits for all firms and thus enhancing information production. Again, the information production gap disappears when we vary n, and the market concentration channel becomes dominant. For an intermediate level of price sensitivity b, the information production gap can be relatively large when changing the number of firms in the market, and the information production channel can dominate that of market concentration. This pattern is illustrated in Figure 13a. However, note that a decrease in b always improves total welfare, because it directly increases firms' profits and consumer welfare and indirectly improves total welfare by enhancing information production.



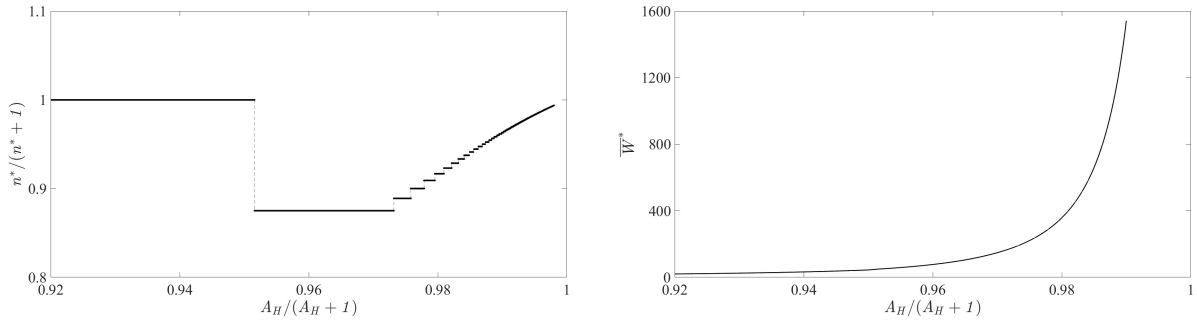
(a) Optimal Market Structure n^* (b) Total Welfare $W(n^*)$

Figure 13: Price Sensitivity b

Parameters: $\theta = 0.75$, $c = 1.5$, $MC = 3$, $A_H = 30$, $A_L = 10$.

Market prospect parameters A_H. Figure 14 depicts n^* and $W(n^*)$ when we vary the market prospect A_H in the good state $\omega = H$. Specifically, when A_H increases from zero to ∞, the optimal market structure n^* first decreases and then increases. Similar to other parameters, the total welfare under the optimal market structure always increases in A_H. Unlike other parameters, A_H affects the equilibrium through two forces, including market uncertainty $(A_H - A_L)$ and average

profitability. These two forces can both increase information production (see, e.g., Equation (8)). However, their impacts on the optimal market structure can diverge, as illustrated in the discussion below, i.e., the negative relationship between competition and total welfare is more likely to occur when average profitability is relatively small (but not too tiny, otherwise the information production gap disappears) or the uncertainty is relatively large (but not too large). In other words, an increase in average profitability weakens, while an increase in market uncertainty reinforces the importance of the information production channel in the negative relationship between competition and total welfare.



(a) Optimal Market Structure n^* (b) Total Welfare $W(n^*)$

Figure 14: Market Prospect Paramter A_H

Parameters: $\theta = 0.75$, $b = 1.5$, $c = 1.5$, $MC = 3$, $A_L = 10$.

B.3 Calibration Based on US market data

This section provides a detailed explanation of the process used to estimate model parameters based on US market data. Before introducing the specific estimation procedure, we first clarify the parameters required to compute the impact of feedback effects, denoted as η.

Substituting the expression for $\overline{W}\left(\widehat{\alpha}_n, n\right)$ into Equation (19) and simplifying, we obtain:

$$\eta = \frac{T_W\left(\widehat{\alpha}_n, n\right) - T_W\left(\widehat{\alpha}_{n-1}, n-1\right)}{T_W(0,n) - T_W(0,n-1)} - 1 \tag{B.1}$$

where

$$T_W\left(\widehat{\alpha}_n, n\right) = \frac{n(n+2)}{8(n+1)^2}\left(\left(\frac{A_H}{MC} + \frac{A_L}{MC} - 2\right)^2 + \left(1 - \hat{\gamma}_n^n\right)\left(\frac{A_H}{MC} - \frac{A_L}{MC}\right)^2\right).$$

Furthermore, using the equilibrium condition $\Pi\left(\widehat{\alpha}_n\right) = c$, we derive:

$$\frac{\gamma_n(2\theta-1)\left(2 + (n-1)\gamma_n^{n-1}\right)\left(\frac{A_H}{MC} + \frac{A_L}{MC} - 2\right)\left(\frac{A_H}{MC} - \frac{A_L}{MC}\right)}{4(n+1)^2} = \frac{b * c}{MC^2}. \tag{B.2}$$

From Equations (B.1) and (B.2), we need to estimate the parameters n and θ, as well as the

three ratios $\frac{A_H}{MC}$, $\frac{A_L}{MC}$, and $\frac{bc}{MC^2}$, to compute η. Without loss of generality, we assume $b = MC = 1$.[21] Additionally, since the information precision parameter θ is difficult to estimate from real-world data, we rely on the restriction $\theta \in (0.5, 1)$ and a reasonable compromise is to set $\theta = 0.75$.

Next, we proceed with estimating the remaining four parameters: n, A_H, A_L, and c. Specifically, we used US industry data to illustrate the parameter estimation process, which is similar for industry-specific estimations. The required data includes firm financial data from *Compustat* (1950–2023), analyst forecasts from *Zacks Investment Research Database* (2000–2023), and PIN data from *Stephen Brown's website* (1993–2010). The sample period for parameter estimation is 2000–2010. Following Gu (2016) and Hou and Robinson (2006), industries are classified using three-digit SIC codes from *CRSP*. Financial and utility firms, as well as industries with negative gross margins, are excluded to align with the Cournot model. Continuous variables are winsorized at the 1st and 99th percentiles to reduce extreme value effects.

First, we estimate competition intensity n using the **Herfindahl-Hirschman Index (HHI)**. Following Gu (2016), we can define:

$$HHI_{jt} = \sum_{i=1}^{N_j} s_{ijt}^2,$$

where s_{ij} is firm i's market share in industry j in year t, and N_j is the number of firms. Market share is computed as *net sales* (Compustat *SALE*) divided by total industry sales. The sample mean of US industry HHI is 0.361. In the Cournot model, with n homogeneous firms, $HHI = \sum_{i=1}^{n} \frac{1}{n^2} = \frac{1}{n}$. Thus, we estimate: $n = \frac{1}{0.361} \approx 3$.

Second, we will estimate A_H and A_L. Since these parameters are not directly convenient to estimate, we instead estimate the average profitability $\bar{A} - MC$ and market uncertainty $A_H - A_L$. First, we use the gross margin GM_{it} to estimate the average profitability $\bar{A} - MC$. The gross margin GM_{it} for each firm i in year t is calculated as one minus the cost of goods sold scaled by sales. From this, the sample mean of the gross margin for U.S. firms is calculated to be 0.236. In the Cournot model, the average gross margin (GM) can be expressed as:

$$GM = \frac{\bar{P} - MC}{\bar{P}} = \frac{\bar{A} - bnq_M - MC}{\bar{A} - bnq_M} = \frac{\bar{A} - MC}{\bar{A} + nMC}.$$

Using this, along with $MC = 1$ and $n = 3$, we can estimate $\bar{A} - MC = 1.236$.

Third, we estimate market uncertainty $A_H - A_L$ using analyst forecast errors, as they reflect both public market information and managerial insights, with higher uncertainty leading to larger errors. The mean absolute percentage error (MAPE) is calculated as:

$$\text{MAPE} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left| \frac{\text{Sales}_{\text{Fit}} - \text{Sales}_{\text{Ait}}}{\text{Sales}_{\text{Ait}}} \right| \times 100\%,$$

where i is the firm index, t is the year index, N is the number of firms, T is the number of years, $\text{Sales}_{\text{Ait}}$ is actual sales in year t, and $\text{Sales}_{\text{Fit}}$ is the median analyst forecast for year t in year

[21] Note that in Equation (B.2), the ratio $\frac{b*c}{MC^2}$, rather than b alone, enters the equilibrium condition and is related to the probability of misallocation in equilibrium. In calibration, we directly estimate the size of informed speculators α and the probability of misallocation γ.

$t-1$ (Polk and Sapienza, 2008). The MAPE is 0.292. Since MAPE measures relative market uncertainty, we compare it to the Coefficient of Variation (CV) of A:

$$CV = \frac{\sqrt{\Pr(\omega = H) \times \left(A_H - \bar{A}\right)^2 + \Pr(\omega = L) \times \left(A_L - \bar{A}\right)^2}}{\bar{A}} = \frac{A_H - A_L}{2\bar{A}}.$$

Given $\bar{A} - MC = 1.236$, we estimate $A_H - A_L = 1.306$, yielding $A_H = 2.889$ and $A_L = 1.583$.

Fourth, we estimate the information cost c using sample data of PIN (Probability of Informed Trading, see Easley et al. (1996)). Since PIN directly estimates the probability of informed trading (Easley et al., 1996), its sample mean provides a reasonable estimate of $\widehat{\alpha}$ at equilibrium, allowing us to estimate c. With a full-sample mean of PIN equal to 0.233, we substitute $\widehat{\alpha} = 0.233$ and the other estimated parameters into equation (B.2), yielding $c = 0.079$. A similar approach allows for parameter estimation across industries.

In addition, we use parameters calibrated from US market data to redraw Figures 3-8.

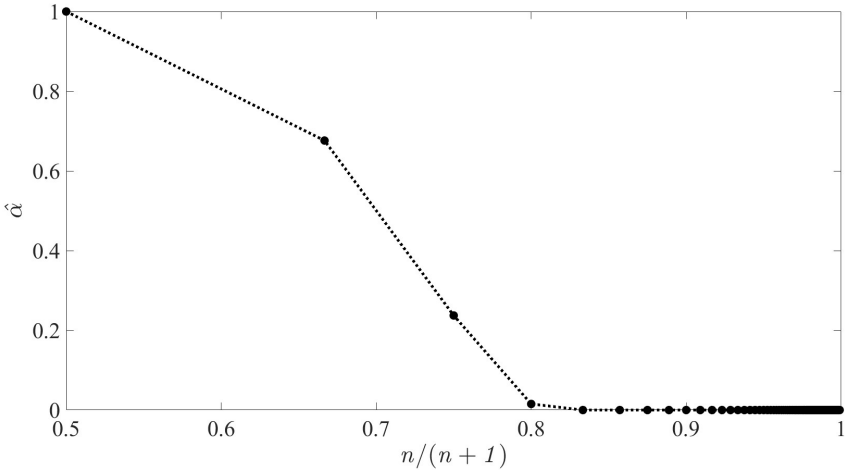


Figure 15: Production Competition and Information Production (Calibrated Data)

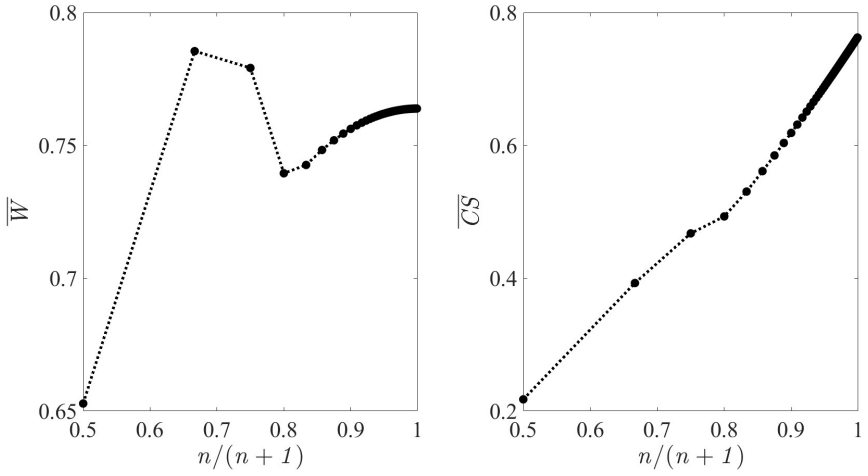


Figure 16: Competition, Total Welfare and Consumer Welfare (Calibrated Data)

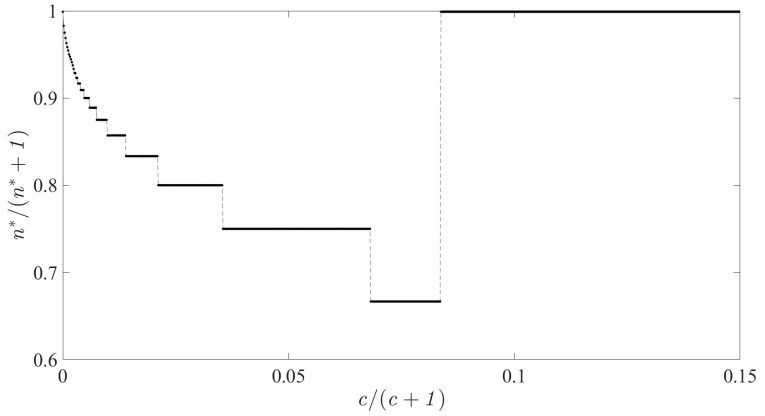


Figure 17: Optimal Market Structure (Calibrated Data)

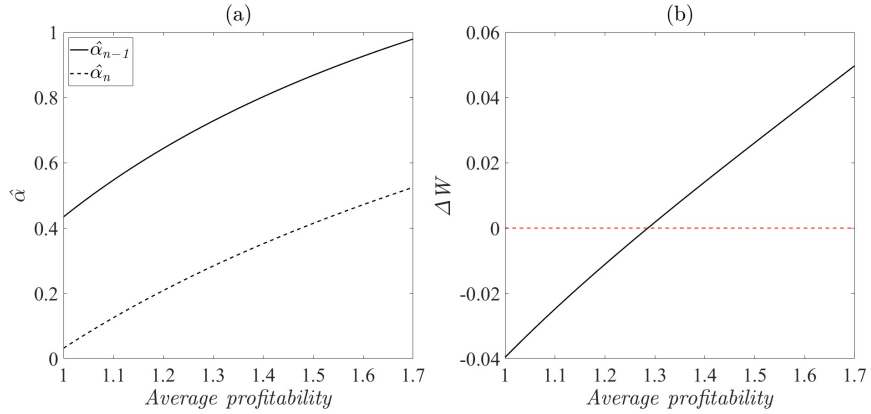


Figure 18: Average Profitability (Calibrated Data)

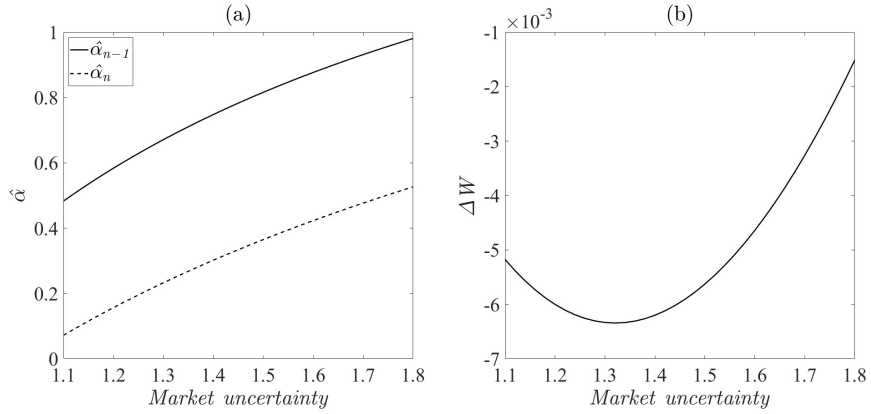


Figure 19: Market Uncertainty (Calibrated Data)

B.4 Equilibrium Analysis in Section 5.1

This section analyzes the equilibrium for the cross-asset trading setup in Section 5.1. We first solve the equilibrium, taking as given the measures of informed speculators $\boldsymbol{\alpha}$, which is then determined by investigating the incentive for information acquisition. Analogous to Lemma 1, given $\boldsymbol{\alpha}$, the stock price $s_i(f_i)$ is determined as:

$$s_i\left(f_i\right)=\begin{cases} s_H & \text{if } \ f_i\in(\gamma_i^{LS},\infty);\\ s_M^i & \text{if } \ f_i\in[-\gamma_i^{LS},\gamma_i^{LS}];\\ s_L & \text{if } \ f_i\in(-\infty,-\gamma_i^{LS}).\end{cases}$$

where $s_H=\frac{(A_H-MC)^2}{(n+1)^2b}$, $s_M^i=\frac{1}{4(n+1)^2b}\left(2\left(A_H-MC\right)^2+2\left(A_L-MC\right)^2-\beta_i^{LS}\left(A_H-A_L\right)^2\right)$, $s_L=\frac{(A_L-MC)^2}{(n+1)^2b}$, $\gamma_i^{LS}=1-(2\theta-1)(\alpha_L+\alpha_{i,S})$ and $\beta_i^{LS}=\prod_{j\neq i}\gamma_i^{LS}$.

Furthermore, the ith firm's optimal production strategy, conditional on the stock prices observed, is given by:

$$q_i^*(\boldsymbol{s})=\begin{cases} q_H & \text{if } \ \exists j\in\{1,\ldots,n\}:s_j=s_H;\\ q_M & \text{if } \ \forall j\in\{1,\ldots,n\}:s_j=s_M^j;\\ q_L & \text{if } \ \exists j\in\{1,\ldots,n\}:s_j=s_L.\end{cases}$$

where $q_H=\frac{A_H-MC}{(n+1)b},\bar{A}=\frac{1}{2}\left(A_H+A_L\right),q_M=\frac{\bar{A}-MC}{(n+1)b}$, and $q_L=\frac{A_L-MC}{(n+1)b}$.

Next, we endogenize the measure of informed traders $\boldsymbol{\alpha}$. Specifically, for an informed L-trader k with a private signal m_k, the optimal trading strategy is to hold $y_k^j=+1$ ($y_k^j=-1$) share of each firm $j\in\{1,\ldots,n\}$ when $m_k=H$ ($m_k=L$), leading to an expected trading profit given by:

$$\Pi_L(\boldsymbol{\alpha})=\frac{\left(\bar{A}-MC\right)\left(A_H-A_L\right)(2\theta-1)\sum_{j=1}^n\gamma_j^{LS}\left(2+(n-1)\beta_j^{LS}\right)}{2b(n+1)^2}$$

Similarly, for an informed S-trader k with a private signal m_k^i, the optimal trading strategy is to buy $x_k^i=+1$ shares of the ith stock when $m_k^i=H$, and sell $x_k^i=-1$ shares of the ith stock when $m_k^i=L$. This leads to an expected trading profit:

$$\Pi_S^i(\boldsymbol{\alpha})=\frac{\left(\bar{A}-MC\right)\left(A_H-A_L\right)(2\theta-1)\gamma_i^{LS}\left(2+(n-1)\beta_i^{LS}\right)}{2b(n+1)^2}$$

Since all firms in the Cournot competition are identical, we can focus on the symmetric equilibrium in which $\alpha_{i,S}=\alpha_S$. Then, with information acquisition, the expected profits for the L- and S-traders can be further written as: $\Pi_L(\boldsymbol{\alpha})=n\Pi_S(\boldsymbol{\alpha})$ and

$$\Pi_S(\boldsymbol{\alpha})=\Pi_S(\alpha_L,\alpha_S)=\frac{\left(\bar{A}-MC\right)\left(A_H-A_L\right)(2\theta-1)\gamma^{LS}\left(2+(n-1)(\gamma^{LS})^{n-1}\right)}{2b(n+1)^2}\tag{B.3}$$

where $\gamma^{LS}=1-(2\theta-1)(\alpha_L+\alpha_S)$.

By comparing $\Pi_L(\boldsymbol{\alpha})$ and $\Pi_S(\boldsymbol{\alpha})$, we can observe that L-traders have a stronger incentive to acquire information than S-traders, given that $c_L\leq c_S$.This further implies: (1) if $\alpha_S>0$, then

$\alpha_L = \lambda$; and (2) if $\alpha_L < \lambda$, then $\alpha_S = 0$. Using this property, we can derive the optimal strategies for information production as follows.

Lemma B.1 (Information Production)**.** *The equilibrium intensity of information production* $(\widetilde{\alpha}_L, \widetilde{\alpha}_S)$ *satisfies the following:*

(i) when $c_L \geq \Pi_L(0,0)$*, then* $\widetilde{\alpha}_L = \widetilde{\alpha}_S = 0$*;*

(ii) when $\Pi_L(\lambda,0) < c_L < \Pi_L(0,0)$*, then* $\widetilde{\alpha}_S = 0$ *and* $\widetilde{\alpha}_L \in (0,\lambda)$*, where* $\Pi_L(\widetilde{\alpha}_L, 0) = c_L$*;*

(iii) when $c_L < \Pi_L(\lambda,0)$ *and* $c_S \geq \Pi_S(\lambda,0)$*, then* $\widetilde{\alpha}_L = \lambda$ *and* $\widetilde{\alpha}_S = 0$*;*

(iv) when $c_L < \Pi_L(\lambda,0)$ *and* $\Pi_S(\lambda, 1-\lambda) < c_S < \Pi_S(\lambda,0)$*, then* $\widetilde{\alpha}_L = \lambda$ *and* $\widetilde{\alpha}_S \in (0, 1-\lambda)$*, where* $\Pi_S(\lambda, \widetilde{\alpha}_S) = c_S$*; and*

(v) when $c_L < \Pi_L(\lambda,0)$ *and* $c_S \leq \Pi_S(\lambda, 1-\lambda)$*, then* $\widetilde{\alpha}_L = \lambda$ *and* $\widetilde{\alpha}_S = 1-\lambda$*.*

Define $\widetilde{\boldsymbol{\alpha}}_n := \widetilde{\boldsymbol{\alpha}}(n)$. Finally, following the derivation of Equation (14), we can compute the expected total welfare $\widetilde{W}\left(\widetilde{\boldsymbol{\alpha}}_n, n\right)$ as follows:

$$\widetilde{W}\left(\widetilde{\boldsymbol{\alpha}}_n, n\right) = \frac{n(n+2)}{8b(n+1)^2}\left(4\left(\bar{A} - MC\right)^2 + \left(1 - (\widetilde{\gamma}^{LS})^n\right)\left(A_H - A_L\right)^2\right) \tag{B.4}$$

where $\widetilde{\gamma}^{LS} = 1 - (\widetilde{\alpha}_L + \widetilde{\alpha}_S) \times (2\theta - 1)$.

Furthermore, define $\gamma_S = 1 - (2\theta-1)\left(\lambda + \widetilde{\alpha}_S\right)$, $\gamma_L = 1 - \widetilde{\alpha}_L(2\theta - 1)$,

$$g_S\left(\widetilde{\alpha}_S, n\right) = 2\gamma_S^n + \frac{n(n+2)\gamma_S^n}{2 + n(n-1)\gamma_S^{n-1}}\left(4n + n(n-3)\gamma_S^{n-1} - 2(n+1)\ln\frac{1}{\gamma_S}\right)$$

and

$$g_L\left(\widetilde{\alpha}_L, n\right) = \frac{(\gamma_L)^n \times \left(2n(n-1)(n+2) + 4 - 3n^2(n+1)\gamma_L^{n-1} - 2n(n+1)(n+2)\ln\frac{1}{\gamma_L}\right)}{2 + n(n-1)\gamma_L^{n-1}}$$

With the aid of Equation (B.4), we can check the relationship between competition and total welfare when an interior solution arises for information production.

Lemma B.2 (Competition and Welfare with Cross-Asset Trading)**.** *Product competition decreases total welfare* $\widetilde{W}(\widetilde{\alpha}_L, \widetilde{\alpha}_S, n)$*, i.e.,* $\frac{d\widetilde{W}(\widetilde{\alpha}_L, \widetilde{\alpha}_S, n)}{dn} < 0$*, when:*

(i) $g_S\left(\widetilde{\alpha}_S, n\right) > G_1(A_H, A_L, MC)$ *in Case 1 such that* $\widetilde{\alpha}_L = \lambda$*; and*

(ii) $g_L\left(\widetilde{\alpha}_L, n\right) > G_1(A_H, A_L, MC)$ *in Case 2 so that* $\widetilde{\alpha}_S = 0$*.*

We make two comments. First, Lemma B.2 verifies the validity of our key result on the non-monotonic relationship between competition and total welfare in the presence of L-traders. The numerical insights are similar and are shown in Appendix B.4.

Second, the incentive for information production can increase with the number of firms for L-traders (i.e., $\frac{d\widetilde{\alpha}_L}{dn} > 0$ for a certain range of n when $\widetilde{\alpha}_S = 0$), which differs significantly from the case for S-traders when $\lambda = 0$ (i.e., $\frac{d\widetilde{\alpha}_S}{dn} < 0$ by Proposition 2). This complexity is illustrated in Figure 10. In particular, when we move from a monopoly ($n = 1$) to a duopoly ($n = 2$), the size of the informed L-traders $\widetilde{\alpha}_L$ first increases and then decreases when n increases. To understand this

non-monotonicity, we plug in $\widetilde{\alpha}_S = 0$ and use Equation (B.3) to obtain:

$$\Pi_L(\boldsymbol{\alpha}) = n\Pi_S(\alpha_L, \alpha_S) = \frac{n\widetilde{\gamma}\left(\bar{A} - MC\right)(A_H - A_L)(2\theta - 1)\left(2 + (n-1)\widetilde{\gamma}^{n-1}\right)}{2b(n+1)^2}$$

where $\widetilde{\gamma} = 1 - (2\theta - 1)\widetilde{\alpha}_L$. We can further compute:

$$\frac{\partial \Pi_L}{\partial n} = \frac{(2\theta - 1)(A_H - A_L)\left(\bar{A} - MC\right)}{2b(n+1)^3} \times \left\{\widetilde{\gamma}^n(3n-1) - 2\widetilde{\gamma}(n-1) - \left(\log\frac{1}{\widetilde{\gamma}}\right)\widetilde{\gamma}^n n(n-1)(n+1)\right\}$$

Therefore, it is possible that $\frac{\partial \Pi_L}{\partial n} > 0$. For example, when α_L is sufficiently small,

$$\frac{\partial \Pi_L}{\partial n} = \frac{(2\theta - 1)(A_H - A_L)\left(\bar{A} - MC\right)}{2b(n+1)^2} + \frac{n(n-1)\widetilde{\alpha}_L}{(n+1)^2} \times O(1) > 0$$

Note that $\frac{\partial \Pi_L}{\partial n} > 0$ implies that increased competition in the product market can strengthen the incentive for L-traders to acquire and trade on private information. Intuitively, as shown in Vives (1985), the profit of firms converges to zero at a speed of $1/n$. When multiplied by the number of firms n, the trading profits for L-traders can be non-monotonicity in n. We term this the "trading opportunity effect" in cross-asset trading.

Numerical analysis. Here, we use numerical methods to verify that the basic insights still hold when there are both L-traders and S-traders in the stock market. Again, let $\Delta\widetilde{W}_n$ denote the incremental change in total welfare when the number of firms increases from $(n-1)$ to n, i.e., $\Delta\widetilde{W}_n = \widetilde{W}(\widetilde{\boldsymbol{\alpha}}_n, n) - \widetilde{W}(\widetilde{\boldsymbol{\alpha}}_{n-1}, n-1)$.

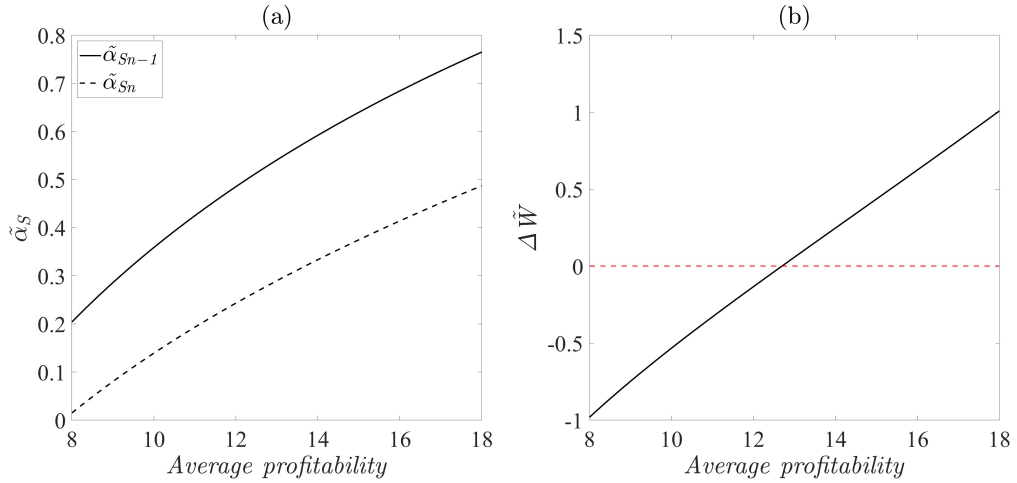


Figure 20: Average Profitability, Information Quality and Welfare.

Parameters: $A_H - A_L = 10, b = 1.5, \theta = 0.75, n = 5, MC = 3, c_L = c_S = 1.5, \lambda = 0.2$.

Remark: (Case 1) the intensity of information production for L-traders satisfies: $\widetilde{\alpha}_L = \lambda$.

First, Figure 20 illustrates how average profitability $\left(\bar{A} - MC\right)$ affects information production $\widetilde{\alpha}_S$ and total welfare $\Delta\widetilde{W}_n$ when all L-traders choose to acquire information. Specifically, similar

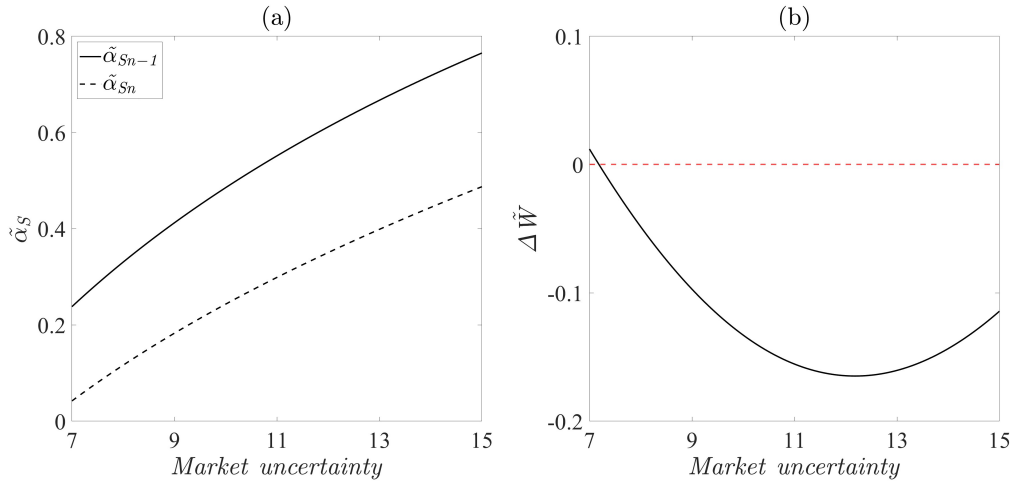


Figure 21: Uncertainty, Information Quality and Welfare.

Parameters: $\bar{A} = 15, b = 1.5, \theta = 0.75, n = 5, MC = 3, c_L = c_S = 1.5, \lambda = 0.2$.

Remark: (Case 1) the intensity of information production for L-traders satisfies: $\widetilde{\alpha}_L = \lambda$.

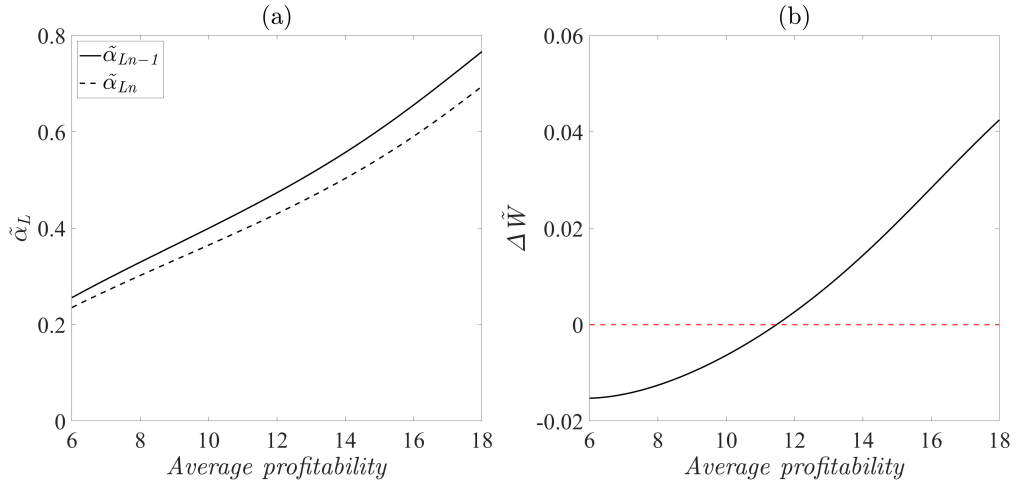


Figure 22: Average Profitability, Information Quality and Welfare.

Parameters: $A_H - A_L = 10, b = 2.5, \theta = 0.75, n = 14, MC = 6.5, c_L = c_S = 1.5, \lambda = 0.8$.

Remark: (Case 2) the intensity of information production for S-traders satisfies: $\widetilde{\alpha}_S = 0$.

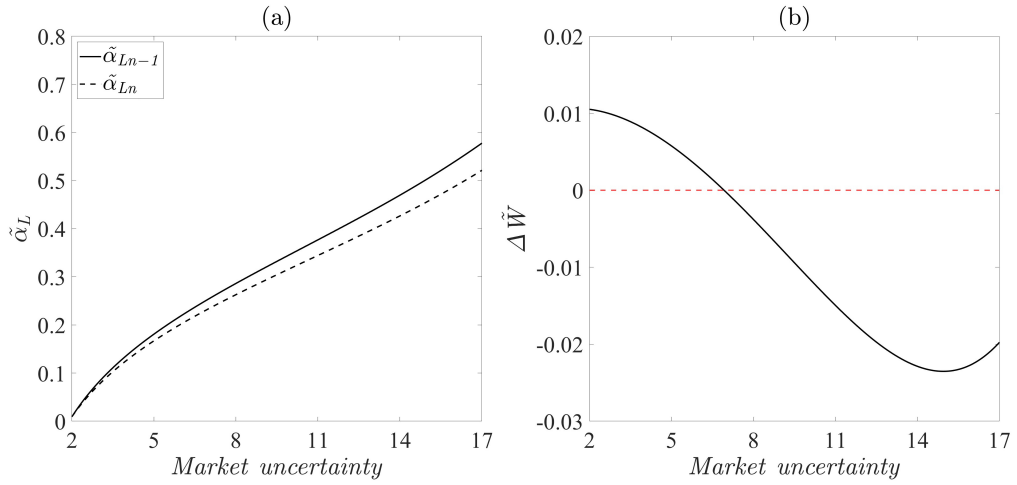


Figure 23: Uncertainty, Information Quality and Welfare.

Parameters: $A_H = 20, A_L = 10, b = 2.5, \theta = 0.75, n = 14, MC = 6.5, c_L = c_S = 1.5, \lambda = 0.8$.

Remark: (Case 2) the intensity of information production for S-traders satisfies: $\widetilde{\alpha}_S = 0$.

to Figure 7, it delivers three messages, including: (1) the intensity of information production $\widetilde{\boldsymbol{\alpha}}_n$ decreases in the number of firms n; (2) both $\widetilde{\boldsymbol{\alpha}}_n$ and $\widetilde{\boldsymbol{\alpha}}_{n-1}$ increase the average profitability $\left(\bar{A} - MC\right)$; and (3) the welfare gain $\Delta\widetilde{W}_n$ is smaller for a lower average profitability, which can even be negative when the average profitability is sufficiently low.

Furthermore, Figure 21 shows the impact of uncertainty, measured by $(A_H - A_L)$, on information production and total welfare. Specifically, it delivers three messages, including: (1) the intensity of information production $\widetilde{\boldsymbol{\alpha}}_n$ decreases in the number of firms n; (2) both $\widetilde{\boldsymbol{\alpha}}_n$ and $\widetilde{\boldsymbol{\alpha}}_{n-1}$ increase in market uncertainty $(A_H - A_L)$; and (3) the incremental welfare change can be negative when market uncertainty $(A_H - A_L)$ is high. Finally, a similar pattern ensues when all S-traders abstain from acquiring information and only a fraction of L-traders choose to produce information.

B.5 Equilibrium Analysis in Section 5.2

Equilibrium analysis. Recall that we let α_L and $\alpha_{i,S}$ denote the measure of informed L-traders and that of informed S-traders for the ith firm, and the size of L-traders is $\lambda = 0$. We first solve the equilibrium for a fixed $\boldsymbol{\alpha}$. Specifically:

$$s_i(\Omega) = \begin{cases} s_H & \text{if } \exists j : f_j \in (\gamma_j^{LS}, \infty); \\ s_M & \text{if } \forall j : f_j \in [-\gamma_j^{LS}, \gamma_j^{LS}]; \\ s_L & \text{if } \exists j : f_j \in (-\infty, -\gamma_j^{LS}). \end{cases} \tag{B.5}$$

where $s_H = \frac{(A_H - MC)^2}{(n+1)^2 b}$, $s_M = \frac{\left(\bar{A} - MC\right)^2}{(n+1)^2 b}$, $s_L = \frac{(A_L - MC)^2}{(n+1)^2 b}$, and $\gamma_i^{LS} = 1 - (2\theta - 1)(\alpha_L + \alpha_{i,S})$.

Furthermore, the ith firm optimally chooses production based on observed stock prices:

$$q_i^*(\boldsymbol{s}) = \begin{cases} q_H & \text{if } \exists j : s_j = s_H; \\ q_M & \text{if } \forall j : s_j = s_M; \\ q_L & \text{if } \exists j : s_j = s_L. \end{cases}$$

where $q_H = \frac{A_H - MC}{(n+1)b}$, $q_M = \frac{\bar{A} - MC}{(n+1)b}$ and $q_L = \frac{A_L - MC}{(n+1)b}$.

Again, for an informed L-trader k with a private signal m_k, the optimal trading strategy is to buy $y_k^j = +1$ ($y_k^j = -1$) share of each firm j when $m_k = H$ ($m_k = L$), leading to an expected trading profit given by:

$$\Pi_{L,C}(\boldsymbol{\alpha}) = \frac{n(2\theta-1)\left(\bar{A} - MC\right)(A_H - A_L)\left(\prod_{j=1}^n \gamma_j^{LS}\right)}{2b(n+1)}$$

Similarly, for an informed S-trader k with a private signal m_k^i, the optimal trading strategy is to buy $x_k^i = +1$ shares of the ith stock when $m_k^i = H$, and sell $x_k^i = -1$ shares of the ith stock when $m_k^i = L$, leading to an expected trading profit of:

$$\Pi_{S,C}(\boldsymbol{\alpha}) = \frac{(2\theta-1)\left(\bar{A} - MC\right)(A_H - A_L)\left(\prod_{j=1}^n \gamma_j^{LS}\right)}{2b(n+1)}$$

Here, the symbol "C" in the subscript means "cross-asset learning".

By focusing on the symmetric equilibrium (i.e., $\alpha_{i,S} = \alpha_S$), the expected profits for the L- and S-traders can be further written as: $\Pi_L(\boldsymbol{\alpha}) = n\Pi_S(\boldsymbol{\alpha})$ and

$$\Pi_{S,C}(\boldsymbol{\alpha}) = \frac{(2\theta-1)\left(\bar{A} - MC\right)(A_H - A_L)(\gamma^{LS})^n}{2b(n+1)} \tag{B.6}$$

where $\gamma^{LS} = 1 - (2\theta - 1) \times (\alpha_L + \alpha_S)$.

Now, we turn to equilibrium information production. Define

$$\nu = \frac{1}{(2\theta-1)} - \frac{1}{(2\theta-1)}\left(\frac{2bc_L(n+1)}{n(2\theta-1)\left(\bar{A} - MC\right)(A_H - A_L)}\right)^{1/n}, \quad \text{and}$$

$$\xi = \frac{1}{(2\theta-1)} - \frac{1}{(2\theta-1)}\left(\frac{2bc_S(n+1)}{n(2\theta-1)\left(\bar{A} - MC\right)(A_H - A_L)}\right)^{1/n} - \lambda$$

Lemma B.3 (Information Production)**.** *The equilibrium intensity of information production* $(\widetilde{\alpha}_{L,C}, \widetilde{\alpha}_{S,C})$ *satisfies the following:*

(i) when $c_L \geq \Pi_{L,C}(0,0)$*, then* $\widetilde{\alpha}_{L,C} = \widetilde{\alpha}_{S,C} = 0$*;*

(ii) when $\Pi_{L,C}(\lambda,0) < c_L < \Pi_{L,C}(0,0)$*, then* $\widetilde{\alpha}_{S,C} = 0$ *and* $\widetilde{\alpha}_{L,C} = \nu \in (0,\lambda)$*;*

(iii) when $c_L < \Pi_{L,C}(\lambda,0)$ *and* $c_S \geq \Pi_{S,C}(\lambda,0)$*, then* $\widetilde{\alpha}_{L,C} = \lambda$ *and* $\widetilde{\alpha}_{S,C} = 0$*;*

(iv) when $c_L < \Pi_{L,C}(\lambda,0)$ *and* $\Pi_{S,C}(\lambda,1-\lambda) < c_S < \Pi_{S,C}(\lambda,0)$*, then* $\widetilde{\alpha}_{L,C} = \lambda$ *and* $\widetilde{\alpha}_{S,C} = \xi \in (0,1-\lambda)$*; and*

(v) when $c_L < \Pi_{L,C}(\lambda,0)$ *and* $c_S \leq \Pi_{S,C}(\lambda,1-\lambda)$*, then* $\widetilde{\alpha}_{L,C} = \lambda$ *and* $\widetilde{\alpha}_{S,C} = 1-\lambda$*.*

Define $\widetilde{\boldsymbol{\alpha}}_n := \widetilde{\boldsymbol{\alpha}}(n)$. Finally, following the derivation of Equation (14), we can compute the expected total welfare $\widetilde{W}_{LS}(\widetilde{\boldsymbol{\alpha}}_n, n)$ as follows:

$$\widetilde{W}_{LS}(\widetilde{\boldsymbol{\alpha}}_n, n) = \frac{n(n+2)}{8b(n+1)^2}\left(4\left(\bar{A} - MC\right)^2 + \left(1 - (\widetilde{\gamma}^{LS})^n\right)(A_H - A_L)^2\right) \tag{B.7}$$

where $\widetilde{\gamma}^{LS} = 1 - (2\theta - 1) \times (\widetilde{\alpha}_L + \widetilde{\alpha}_S)$.

Recall that $\gamma_S = 1 - (2\theta - 1)(\lambda + \widetilde{\alpha}_S)$, $\gamma_L = 1 - \widetilde{\alpha}_L(2\theta - 1)$. Define

$$g_{S,C}(\gamma_S, n) = (\gamma_S)^n(2 + n(n+1)(n+2)).$$

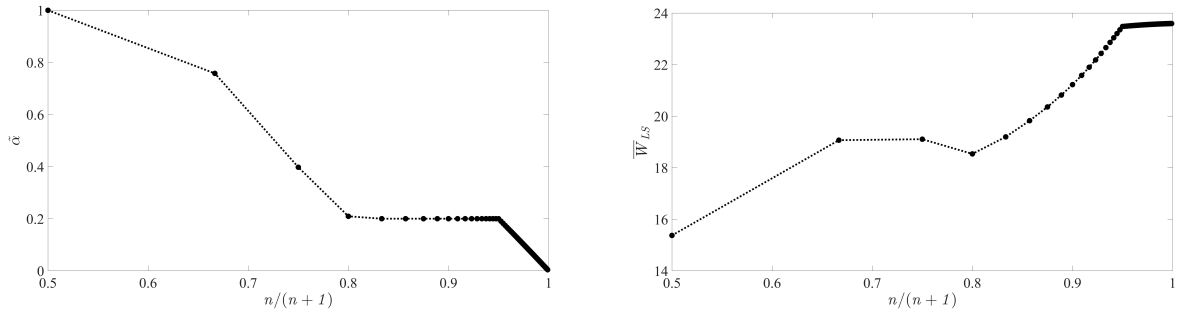
Lemma B.4 (Competition and Welfare with Cross-Asset Learning)**.**

(i) Case 1: $\widetilde{\alpha}_{L,C} = \lambda$. Then, the total welfare decreases in the number of firms n (i.e., $\frac{d\widetilde{W}_{LS}(\widetilde{\boldsymbol{\alpha}}_n, n)}{dn} < 0$) if and only if $g_{S,C}(\gamma_S, n) > G_1(A_H, A_L, MC)$; and

(ii) Case 2: $\widetilde{\alpha}_{S,C} = 0$. Then, the total welfare increases strictly in the number of firms n, i.e., $\frac{d\widetilde{W}_{LS}(\widetilde{\boldsymbol{\alpha}}_n, n)}{dn} > 0$.

Lemma B.4 requires several additional clarifications, given that market makers can observe the flow of orders in all stocks. First, when there are only S-traders in the stock market (i.e., $\lambda = 0$ and thus $\widetilde{\alpha}_{L,C} = 0 = \lambda$ always holds), the nonmonotonic relationship between competition and total welfare still holds. Second, the non-monotonicity also holds when the cost of information production is small such that $\widetilde{\alpha}_{L,C} = \lambda$. Note that L-traders have a stronger incentive to acquire information, compared to S-traders. Third, when there are only L-traders (i.e., $\lambda = 1$ and thus $\widetilde{\alpha}_{S,C} = 0$ always holds), the total welfare increases strictly in the number of firms n. In other words, the non-monotonic relationship between competition and total welfare holds when we allow cross-asset trading by L-traders or cross-asset learning by market makers, but not both. Intuitively, there are two economic forces behind this. On the one hand, as discussed in Section 5.1, intensified competition can improve trading profits for L-traders by granting them more trading opportunities. On the other hand, cross-asset learning provides market makers with more information, decreasing speculators' trading profits, and information production in equilibrium. In summary, both the trading opportunity effect and the cross-asset learning effect reduce the impact of the information production channel. A more detailed discussion about the divergent impact of cross-asset learning on L-traders and S-traders can be found in online Appendix B.5.

We first illustrate how competition shapes information production and total welfare when market makers can observe the order flow of all stocks.



(a) Information Production (b) Total Welfare

Figure 24: Competition, Information Production and Total Welfare

Parameters: $\lambda = 0.2$, $\theta = 0.75$, $b = 1.5$, $A_H = 20$, $A_L = 10$, $MC = 8$, and $c_L = c_S = 1.5$.

Numerical analysis. With intensified Cournot competition ($n \uparrow$), the incentive to acquire information weakly decreases. This is illustrated in Figure 24a. First, when $n \leq 4$, an increase

in n reduces the measure of informed S-traders, who have a relatively smaller incentive to acquire information. Second, when $4 < n \leq 18$, S-traders quit from acquiring information and trading on private information, while all L-traders choose to acquire information. Third, when $n \geq 18$, an increase in n further reduces the incentive for L-traders to acquire information.

Correspondingly, Figure 24b depicts total welfare when the number of firms n increases. When $n \leq 4$, total welfare first increases and then decreases and reaches a local minimum when all S-traders abstain from information production. However, when $n \geq 4$, total welfare increases strictly in the number of firms, indicating a dominant role of the market concentration channel.

Understanding the impact of cross-asset learning. By Lemma B.4, cross-asset learning affects L-traders differently from S-traders. Here, we show that this complexity is primarily caused by the combination of the trading opportunity effect and the cross-asset learning effect.

(i) Cross-asset learning effect.

Specifically, with cross-asset learning, market makers can observe the order flow of all stocks, enabling more efficient pricing against informed speculators. Thus, trading profits decrease for both L-traders and S-traders and are lower than those without cross-asset learning. Indeed, given $\widetilde{\gamma}^{LS}$ (or equivalently, $\widetilde{\alpha}_{L,C} + \widetilde{\alpha}_{S,C}$), we have:

$$\frac{\Pi_{L,C}}{\Pi_L} = \frac{\Pi_{S,C}}{\Pi_S} = f_C(n) \tag{B.8}$$

where $f_C(n) = \frac{(n+1)}{2(\widetilde{\gamma}^{LS})^{1-n}+(n-1)}$. Obviously, $f_C(n) \in (0,1)$ and $f_C'(n) < 0$. Therefore, the trading profits of an informed L-trader and an informed S-trader will shrink proportionally by a ratio of $f_C(n)$ when market makers can observe the order flow of all stocks, and this effect is more pronounced when n is large.

(ii) Trading opportunity effect.

This effect arises from the opportunity to access all stock, and thus only exists for L-traders. Unlike an S-trader with small trading opportunities, an L-trader can earn a higher trading profit by acquiring costly information, i.e., $\Pi_L = n\Pi_S$ and $\Pi_{L,C} = n\Pi_{S,C}$. Therefore, the expected trading profit of an L-trader can increase with n, especially when n is small. For example, we can verify that $\frac{\partial \Pi_L}{\partial n} > 0$ for $n = 1$, which differs from the case with an S-trader whose expected trading profit always decreases in n. However, note that $\frac{\partial \Pi_L}{\partial n} < 0$ when n is large enough. Figure 25 illustrates the pattern of trading profits with (blue dashed line) and without (red solid line) cross-asset learning by market makers.

We now examine how cross-asset learning affects the incentive for information production. We first consider S-traders, whose expected trading profits Π_S strictly decrease in n and are further reduced by cross-asset learning (i.e., $\frac{d\Pi_{S,C}}{dn} < 0$). Note that $\Pi_S = \Pi_{S,C}$ when $n = 1$ or $n \to \infty$. Then, one would expect that when n is relatively small, $\Pi_{S,C}$ decreases relatively faster than $\Pi_{S,C}$ as n increases. This is illustrated in panel (a) of Figure 25. Therefore, with cross-asset learning, the expected trading profit of an informed S-trader exhibits a higher level of sensitivity in the number of firms (n), which implies that intensified market competition can further reduce the incentive for S-traders to trade on proprietary information compared to the case without cross-asset learning.

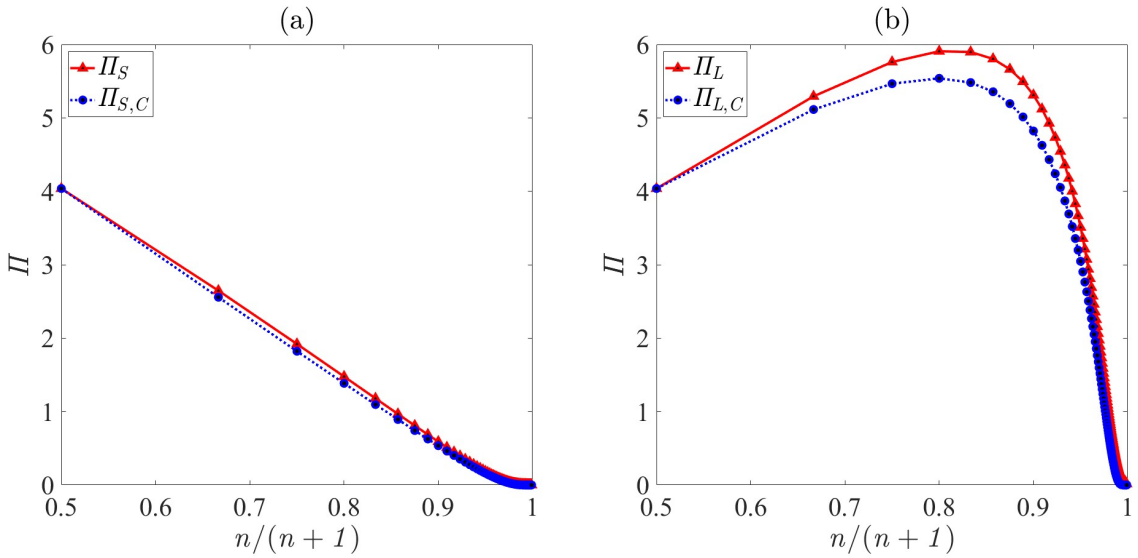


Figure 25: Trading profits with/witout cross-asset learning

Parameters: $\theta = 0.75$, $b = 2.5$, $A_H = 20$, $A_L = 10$, $MC = 6.5$, and $\widetilde{\alpha}_{L,C} + \widetilde{\alpha}_{S,C} = 0.1$.

In other words, it reinforces the informational feedback channel, leading to a stronger (negative) effect of competition on real efficiency.

Next, we consider L-traders, whose expected trading profits Π_L are non-monotonic in n. Specifically, due to the trading opportunity effect, Π_L first increases and then decreases, generating an inverted U-shape pattern when n increases. Similarly, cross-asset learning also decreases the expected trading profit $\Pi_{L,C}$ for L-traders and flattens the inverted U-shape pattern, as shown in panel (b) of Figure 25. Thus, with cross-asset learning by market makers, the expected trading profit of an informed L-trader becomes less sensitive to the number of firms (n) when n is relatively small, leading to weaker informational feedback effects. Therefore, the non-monotonic link between competition and total welfare fails because the trading opportunity effect and cross-asset learning reinforce each other.

As a final remark, Figure 25 appears to indicate that the expected trading profits Π_L and $\Pi_{L,C}$ for L-traders are relatively more sensitive to changes in n when n is large, compared to those of S-traders Π_S and $\Pi_{S,C}$. However, this does not mean that a change in n affects L-traders more than S-traders when it comes to information production. More formally, recall that $\Pi_L = n\Pi_S$ and $\Pi_{L,C} = n\Pi_{S,C}$, which further implies that: $\frac{\partial \Pi_L}{\partial \alpha_L} = n\frac{\partial \Pi_S}{\partial \alpha_S} < 0$ and $\frac{\partial \Pi_{L,C}}{\partial \alpha_L} = n\frac{\partial \Pi_{S,C}}{\partial \alpha_S} < 0$. It then follows that for L-traders, we have:

$$\frac{d\widetilde{\alpha}_L}{dn} = -\frac{1}{n} * \frac{\frac{\partial \Pi_L}{\partial n}}{\frac{\partial \Pi_S}{\partial \alpha_S}} \quad \text{and} \quad \frac{d\widetilde{\alpha}_{L,C}}{dn} = -\frac{1}{n} * \frac{\frac{\partial \Pi_{L,C}}{\partial n}}{\frac{\partial \Pi_{S,C}}{\partial \alpha_S}}$$

In contrast, for S-traders, we have:

$$\frac{d\widetilde{\alpha}_S}{dn} = -\frac{\frac{\partial \Pi_S}{\partial n}}{\frac{\partial \Pi_S}{\partial \alpha_S}} \quad \text{and} \quad \frac{d\widetilde{\alpha}_{S,C}}{dn} = -\frac{\frac{\partial \Pi_{S,C}}{\partial n}}{\frac{\partial \Pi_{S,C}}{\partial \alpha_S}}$$

Furthermore, from $\Pi_L = n\Pi_S$, we know that $\frac{\partial \Pi_L}{\partial n} = n\frac{\partial \Pi_S}{\partial n} + \Pi_S$. It follows that

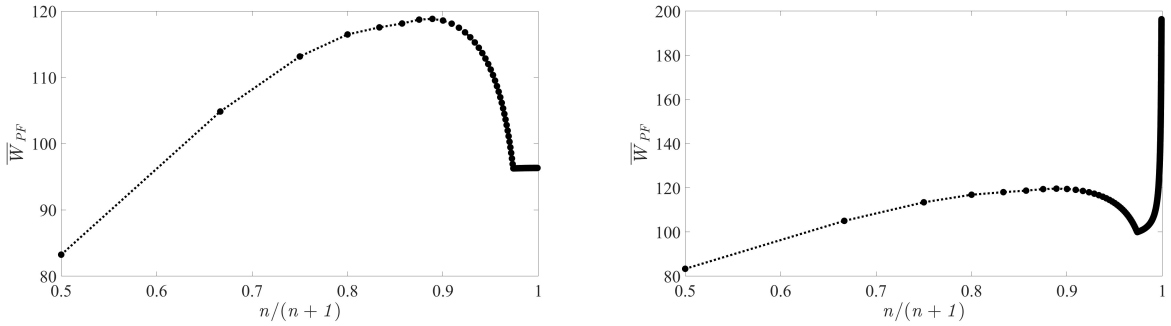
$$\frac{d\widetilde{\alpha}_L}{dn} = \frac{d\widetilde{\alpha}_S}{dn} - \frac{\Pi_S/n}{\frac{\partial \Pi_S}{\partial \alpha_L}} > \frac{d\widetilde{\alpha}_S}{dn}$$

Since $\frac{d\widetilde{\alpha}_S}{dn} < 0$, we have $\left|\frac{d\widetilde{\alpha}_L}{dn}\right| < \left|\frac{d\widetilde{\alpha}_S}{dn}\right|$, when $\frac{d\widetilde{\alpha}_L}{dn} < 0$. Similarly, with cross-asset learning, we also have: $\left|\frac{d\widetilde{\alpha}_{L,C}}{dn}\right| < \left|\frac{d\widetilde{\alpha}_{S,C}}{dn}\right|$, when $\frac{d\widetilde{\alpha}_{L,C}}{dn} < 0$. Thus, intensified market competition will negatively affect S-traders more than L-traders in terms of information production.

B.6 Formal Analysis for Section 5.3

This section provides a formal analysis for Section 5.3. Specifically, we first present a non-monotonic welfare result and then depict the relationship between competition and total welfare when investor welfare is included. Recall that $\Phi(m)$ is defined in Proposition 3, and define $m_0 = \inf\{m \in \mathbb{N} : \Phi(m) \geq 1\}$. Define $\tilde{c} = \frac{2bc}{(A-MC)^2}$.

Lemma B.5 (Informational Feedback & Over-Competition)**.** *Assume $B(n) = B_0$ for some constant B_0. Suppose that $\Phi(m) - m * \tilde{c} - m > 0$ for some $m \geq m_0$. Then, for any $n \geq N(m) > m$, $\overline{W}(\widehat{\alpha}_m, m) > \overline{W}(\widehat{\alpha}_n, n)$ holds for any $c \in [\bar{c}_n, \underline{c}_m)$ with $\bar{c}_n < \underline{c}_m$.*



(a) $B(n) = 1$ (b) $B(n) = 0.1 * n$

Figure 26: Competition & Total Welfare (with Investor Welfare)

Parameters: $\theta = 0.75$, $b = 1.5$, $A_H = 30$, $A_L = 10$, $MC = 3$, and $c = 1.5$.

Figure 26 illustrates the relationship between product competition and total welfare when investor welfare is included in the calculation. Specifically, when the aggregate benefit of liquidity trading is fixed, Figure 26a demonstrates a non-monotonic pattern between competition and total welfare, which is similar to Figure 4. In particular, total welfare first increases and then decreases, and is maximized at $n = 8$. Similarly, Figure 26b illustrates the relationship by specifying the aggregate benefit of liquidity trading as an increasing function of the number of stocks, i.e., $B(n) = 0.1 * n$. The total welfare is also non-monotonic and becomes infinitely large due to the unbounded return from liquidity trading.

B.7 Skipped Proofs in the Online Appendix

B.7.1 Proof of Lemma B.1

Proof. We first state two properties: (a) We compute the following derivatives, including:

$$\frac{\partial \Pi_L(\alpha_L, \alpha_S)}{\partial \alpha_L} = -\frac{n(A_H - A_L)\left(\bar{A} - MC\right)(2\theta - 1)^2\left(2 + n(n-1)(\gamma^{LS})^{n-1}\right)}{2b(n+1)^2} < 0;$$
$$\frac{\partial \Pi_S(\alpha_L, \alpha_S)}{\partial \alpha_S} = -\frac{(A_H - A_L)\left(\bar{A} - MC\right)(2\theta - 1)^2\left(2 + n(n-1)(\gamma^{LS})^{n-1}\right)}{2b(n+1)^2} < 0.$$

and (b) Note that $\Pi_L(\alpha_L, \alpha_S) = n\Pi_S(\alpha_L, \alpha_S)$.

Now, we prove the lemma. First, consider $c_L \geq \Pi_L(0,0)$. Obviously, $\widetilde{\alpha}_L = 0$. Meanwhile, since $c_S \geq c_L$ and $\Pi_L(0,0) \geq \Pi_S(0,0)$, $\widetilde{\alpha}_S = 0$.

Second, consider $\Pi_L(\lambda, 0) < c_L < \Pi_L(0,0)$. By the derivative $\frac{\partial \Pi_L(\alpha_L, \alpha_S)}{\partial \alpha_L} < 0$ and continuity, there exists a unique $\widetilde{\alpha}_L$ such that $\Pi_L(\widetilde{\alpha}_L, 0) = c_L$. Furthermore, given $\widetilde{\alpha}_L$, $\frac{\partial \Pi_S(\alpha_L, \alpha_S)}{\partial \alpha_S} < 0$ implies that $\Pi_S(\widetilde{\alpha}_L, 0) > \Pi_S(\widetilde{\alpha}_L, \alpha_S)$ for any $\alpha_S > 0$. Thus, $c_S \geq c_L = \Pi_L(\widetilde{\alpha}_L, 0) \geq \Pi_S > \Pi_S(\widetilde{\alpha}_L, \alpha_S)$ for any $\alpha_S > 0$. Therefore, $\widetilde{\alpha}_S = 0$.

Third, consider $c_L < \Pi_L(\lambda, 0)$ and $c_S \geq \Pi_S(\lambda, 0)$. Obviously, $(\widetilde{\alpha}_L, \widetilde{\alpha}_S) = (\lambda, 0)$. Furthermore, this is also the unique equilibrium. If not, consider any equilibrium $(\widetilde{\alpha}_L, \widetilde{\alpha}_S)$ with $\widetilde{\alpha}_S > 0$. Note that by property (b), we can infer: $\Pi_L(\widetilde{\alpha}_L, \widetilde{\alpha}_S) > \Pi_S(\widetilde{\alpha}_L, \widetilde{\alpha}_S) \geq c_S \geq c_L$, which implies that $\widetilde{\alpha}_L = \lambda$, which in turn implies that $\widetilde{\alpha}_S = 0$.

Fourth, consider $c_L < \Pi_L(\lambda, 0)$ and $\Pi_S(\lambda, 1-\lambda) < c_S < \Pi_S(\lambda, 0)$. We have shown above that if $\widetilde{\alpha}_S > 0$, then $\widetilde{\alpha}_L = \lambda$. Given that $c_L < \Pi_L(\lambda, 0)$, we can infer that $\widetilde{\alpha}_L = \lambda$. Given this and the assumed condition $\Pi_S(\lambda, 1-\lambda) < c_S < \Pi_S(\lambda, 0)$, by the monotonicity and continuity of $\Pi_S(\alpha_L, \alpha_S)$, there is a unique $\widetilde{\alpha}_S \in (0, 1-\lambda)$ such that $\Pi_S(\lambda, \widetilde{\alpha}_S) = c_S$.

Fifth, consider $c_L < \Pi_L(\lambda, 0)$ and $c_S \leq \Pi_S(\lambda, 1-\lambda)$. Obviously, by the facts $c_S \geq c_L$ and $\Pi_L \geq \Pi_S$, we have: $\widetilde{\alpha}_L = \lambda$ and $\widetilde{\alpha}_S = 1 - \lambda$. The proof concludes. □

B.7.2 Proof of Lemma B.2

Proof. **Case 1: $\widetilde{\alpha}_L = \lambda$.** We can rewrite $\widetilde{W}(\widetilde{\alpha}_L, \widetilde{\alpha}_S, n)$ and $\Pi_S(\widetilde{\alpha}_L, \widetilde{\alpha}_S)$ as:

$$\widetilde{W}(\widetilde{\alpha}_S, n) = \frac{n(n+2)}{8b(n+1)^2}\left(4(\bar{A} - MC)^2 + (1 - {\gamma_S}^n)(A_H - A_L)^2\right),$$
$$\Pi_S(\widetilde{\alpha}_S, n) = \frac{\gamma_S(2\theta - 1)(A_H - A_L)(\bar{A} - MC)\left(2 + (\gamma_S)^{n-1}(n-1)\right)}{2b(n+1)^2}$$

where $\gamma_S = 1 - (\lambda + \widetilde{\alpha}_S)(2\theta - 1)$.

Then, we can calculate the following partial derivatives:

$$\frac{\partial \widetilde{W}}{\partial \widetilde{\alpha}_S} = \frac{n^2(n+2)\gamma_S^{n-1}(2\theta-1)(A_H-A_L)^2}{8b(n+1)^2},$$
$$\frac{\partial \widetilde{W}}{\partial n} = \frac{n(n+2)\gamma_S^n(A_H-A_L)^2\ln(1/\gamma_S)}{8b(n+1)^2} + \frac{2\left((A_H-MC)^2+(A_L-MC)^2\right)-\gamma_S^n(A_H-A_L)^2}{4b(n+1)^3}$$
$$\frac{\partial \Pi_S}{\partial \widetilde{\alpha}_S} = -\frac{(2\theta-1)^2\left((A_H-MC)^2-(A_L-MC)^2\right)\left(2+n(n-1)\gamma_S{}^{n-1}\right)}{4b(n+1)^2}$$
$$\frac{\partial \Pi_S}{\partial n} = -\frac{(2\theta-1)\left((A_H-MC)^2-(A_L-MC)^2\right)\left(4\gamma_S+\gamma_S^n\left(n-3-(n^2-1)\ln\gamma_S\right)\right)}{4b(n+1)^3}$$

By the implicit function theorem, we have:

$$\frac{\partial \widetilde{\alpha}_S}{\partial n} = -\frac{\partial \Pi_S/\partial n}{\partial \Pi_S/\widetilde{\alpha}_S} = -\frac{\gamma_S^n \times \left(\left(4\gamma_S^{1-n}+(n-3)\right)/(n+1)+(n-1)\ln(1/\gamma_S)\right)}{(2\theta-1)\left(2+n(n-1)\gamma_S^{n-1}\right)}$$

which further implies:

$$\frac{d\widetilde{W}\left(\widetilde{\alpha}_{S,C},n\right)}{dn} = \frac{\partial \widetilde{W}}{\partial n} + \frac{\partial \widetilde{W}}{\partial \widetilde{\alpha}_S}\frac{\partial \widetilde{\alpha}_S}{\partial n} = \frac{(A_H-A_L)^2\left(G_1-g_S\left(\widetilde{\alpha}_S,n\right)\right)}{8b(n+1)^3},$$

Thus, $\frac{d\widetilde{W}(\widetilde{\alpha}_{S,C},n)}{dn} < 0$ if and only if $g_S\left(\widetilde{\alpha}_S,n\right) > G_1$.

Case 2: $\widetilde{\alpha}_S = 0$**.** We can rewrite $\widetilde{W}(\widetilde{\alpha}_L,\widetilde{\alpha}_S,n)$ and $\Pi_L(\widetilde{\alpha}_L,\widetilde{\alpha}_S)$ as:

$$\widetilde{W}(\widetilde{\alpha}_L,n) = \frac{n(n+2)}{8b(n+1)^2}\left(4(\bar{A}-MC)^2+(1-(\gamma_L)^n)(A_H-A_L)^2\right),$$
$$\Pi_S(\widetilde{\alpha}_L,n) = \frac{\gamma_S(2\theta-1)(A_H-A_L)(\bar{A}-MC)\left(2+(\gamma_L)^{n-1}(n-1)\right)}{2b(n+1)^2}$$

where $\gamma_L = 1-\widetilde{\alpha}_L\times(2\theta-1)$.

Then, we can calculate the following partial derivatives:

$$\frac{\partial \widetilde{W}}{\partial \widetilde{\alpha}_L} = \frac{n^2(n+2)\gamma_L^{n-1}(2\theta-1)(A_H-A_L)^2}{8b(n+1)^2},$$
$$\frac{\partial \widetilde{W}}{\partial n} = \frac{n(n+2)\gamma_L^n(A_H-A_L)^2\ln(1/\gamma_L)}{8b(n+1)^2} + \frac{2\left((A_H-MC)^2+(A_L-MC)^2\right)-\gamma_L^n(A_H-A_L)^2}{4b(n+1)^3}$$

$$\frac{\partial \Pi_L}{\partial \widetilde{\alpha}_L} = -\frac{n(2\theta-1)^2\left((A_H-MC)^2-(A_L-MC)^2\right)\left(2+n(n-1)\gamma_L{}^{n-1}\right)}{4b(n+1)^2}$$
$$\frac{\partial \Pi_L}{\partial n} = -\frac{(2\theta-1)\left((A_H-MC)^2-(A_L-MC)^2\right)\left(2(1-n)\gamma_L+\gamma_L^n\left((3n-1)+n(n^2-1)\ln\gamma_L\right)\right)}{4b(n+1)^3}$$

By the implicit function theorem, we have:

$$\frac{\partial \widetilde{\alpha}_L}{\partial n} = -\frac{\partial \Pi_L/\partial n}{\partial \Pi_L/\widetilde{\alpha}_L} = \frac{2\gamma_L\times(1-n)+(\gamma_L)^n\left((3n-1)-n(n^2-1)\ln(1/\gamma_L)\right)}{n(n+1)(2\theta-1)\left(2+n(n-1)\gamma_L^{n-1}\right)}$$

which further implies:

$$\frac{d\widetilde{W}\left(\widetilde{\alpha}_L,n\right)}{dn}=\frac{\partial\widetilde{W}}{\partial n}+\frac{\partial\widetilde{W}}{\partial\widetilde{\alpha}_L}\frac{\partial\widetilde{\alpha}_L}{\partial n}=\frac{n\left(A_H-A_L\right)^2\left(G_1-g_L\left(\widetilde{\alpha}_L,n\right)\right)}{8bn(n+1)^3},$$

Thus, $\frac{d\widetilde{W}(\widetilde{\alpha}_L,n)}{dn}<0$ if and only if $g_L\left(\widetilde{\alpha}_L,n\right)>G_1$. The proof concludes. □

B.7.3 Proof of Lemma B.3

Proof. We first state two important properties: (a) $\Pi_{L,C}\left(\alpha_L,\alpha_S\right)=n\Pi_{S,C}\left(\alpha_L,\alpha_S\right)$; and (b) we compute the following derivatives, including $\frac{\partial\Pi_{L,C}(\alpha_L,\alpha_S)}{\partial\alpha_{L,C}}$ and $\frac{\partial\Pi_{S,C}(\alpha_L,\alpha_S)}{\partial\alpha_{S,C}}$. Based on the expressions for trading profits of an informed L-trader and an informed S-trader, we have:

$$\frac{\partial\Pi_{L,C}\left(\alpha_L,\alpha_S\right)}{\partial\alpha_{L,C}}=-\frac{n^2\left(\gamma^{LS}\right)^{n-1}\left(2\theta-1\right)^2\left(\bar{A}-MC\right)\left(A_H-A_L\right)}{2(n+1)b}<0$$
$$\frac{\partial\Pi_{S,C}\left(\alpha_L,\alpha_S\right)}{\partial\alpha_{S,C}}=-\frac{n\left(\gamma^{LS}\right)^{n-1}\left(2\theta-1\right)^2\left(\bar{A}-MC\right)\left(A_H-A_L\right)}{2(n+1)b}<0$$

Next, we prove the lemma. First, consider $c_L\geq\Pi_{L,C}(0,0)$. Obviously, $\widetilde{\alpha}_{L,C}=0$. Meanwhile, since $c_S\geq c_L$ and $\Pi_{L,C}(0,0)=n\Pi_{S,C}(0,0)$, we can deduce that $\widetilde{\alpha}_{S,C}=0$.

Second, consider $\Pi_{L,C}(\lambda,0)<c_L<\Pi_{L,C}(0,0)$. By the derivative $\frac{\partial\Pi_{L,C}(\alpha_L,\alpha_S)}{\partial\alpha_L}<0$, and continuity, there exists a unique $\widetilde{\alpha}_{L,C}$ such that $\Pi_{L,C}\left(\widetilde{\alpha}_{L,C},0\right)=c_L$. By solving the equation $\Pi_{L,C}\left(\widetilde{\alpha}_{L,C},0\right)=c_L$, we have $\widetilde{\alpha}_{L,C}=\nu$. Furthermore, given $\widetilde{\alpha}_{L,C}$, $\frac{\partial\Pi_{S,C}(\alpha_L,\alpha_S)}{\partial\alpha_S}<0$ implies that $\Pi_{S,C}\left(\widetilde{\alpha}_{L,C},0\right)>\Pi_{S,C}\left(\widetilde{\alpha}_{L,C},\alpha_S\right)$ for any $\alpha_S>0$. Thus, $c_S\geq c_L=\Pi_{L,C}\left(\widetilde{\alpha}_{L,C},0\right)>\Pi_{S,C}\left(\widetilde{\alpha}_{L,C},\alpha_S\right)$ for any $\alpha_S>0$. Therefore, $\widetilde{\alpha}_{S,C}=0$.

Third, consider $c_L\leq\Pi_{L,C}(\lambda,0)$ and $c_S\geq\Pi_{S,C}(\lambda,0)$. Obviously, $(\widetilde{\alpha}_{L,C},\widetilde{\alpha}_{S,C})=(\lambda,0)$. Furthermore, this is also the unique equilibrium. If not, consider any equilibrium $(\widetilde{\alpha}_{L,C},\widetilde{\alpha}_{S,C})$ with $\widetilde{\alpha}_{S,C}>0$. Note that by property (b), we can infer: $\Pi_{L,C}\left(\widetilde{\alpha}_{L,C},\widetilde{\alpha}_{S,C}\right)>\Pi_{S,C}\left(\widetilde{\alpha}_{L,C},\widetilde{\alpha}_{S,C}\right)\geq c_S\geq c_L$, which implies that $\widetilde{\alpha}_{L,C}=\lambda$, which in turn implies that $\widetilde{\alpha}_{S,C}=0$.

Fourth, consider $c_L\leq\Pi_{L,C}(\lambda,0)$ and $\Pi_{S,C}(\lambda,1-\lambda)<c_S<\Pi_{S,C}(\lambda,0)$. We have shown above that if $\widetilde{\alpha}_{S,C}>0$, then $\widetilde{\alpha}_{L,C}=\lambda$. Given that $c_L\leq\Pi_{L,C}(\lambda,0)$, we can infer that $\widetilde{\alpha}_{L,C}=\lambda$. Given this and the assumed condition $\Pi_{S,C}(\lambda,1-\lambda)<c_S<\Pi_{S,C}(\lambda,0)$, by the monotonicity and continuity of $\Pi_{S,C}\left(\widetilde{\alpha}_{L,C},\widetilde{\alpha}_{S,C}\right)$, there is a unique $\widetilde{\alpha}_{S,C}\in(0,1-\lambda)$ such that $\Pi_{S,C}\left(\lambda,\widetilde{\alpha}_{S,C}\right)=c_S$. By solving $\Pi_{S,C}\left(\lambda,\widetilde{\alpha}_{S,C}\right)=c_S$, we have $\widetilde{\alpha}_{S,C}=\xi$.

Fifth, consider $c_L\leq\Pi_{L,C}(\lambda,0)$ and $c_S\leq\Pi_{S,C}(\lambda,1-\lambda)$. Obviously, by the facts $c_S\geq c_L$ and $\Pi_{L,C}>\Pi_{S,C}$, we have: $\widetilde{\alpha}_{L,C}=\lambda$ and $\widetilde{\alpha}_{S,C}=1-\lambda$. The proof concludes. □

B.7.4 Proof of Lemma B.4

Proof. We first state two important properties: (a) $\Pi_{L,C}\left(\alpha_L,\alpha_S\right)=n\Pi_{S,C}\left(\alpha_L,\alpha_S\right)$; and (b) we compute the following derivatives, including $\frac{\partial\Pi_{L,C}(\alpha_L,\alpha_S)}{\partial\alpha_{L,C}}$ and $\frac{\partial\Pi_{S,C}(\alpha_L,\alpha_S)}{\partial\alpha_{S,C}}$. Based on the expres-

sions for trading profits of an informed L-trader and an informed S-trader, we have:

$$\frac{\partial \Pi_{L,C}\left(\alpha_L, \alpha_S\right)}{\partial \alpha_{L,C}} = -\frac{n^2\left(\gamma^{LS}\right)^{n-1}(2\theta-1)^2\left(\bar{A}-MC\right)\left(A_H-A_L\right)}{2(n+1)b} < 0$$
$$\frac{\partial \Pi_{S,C}\left(\alpha_L, \alpha_S\right)}{\partial \alpha_{S,C}} = -\frac{n\left(\gamma^{LS}\right)^{n-1}(2\theta-1)^2\left(\bar{A}-MC\right)\left(A_H-A_L\right)}{2(n+1)b} < 0$$

Now, we prove the lemma.

Case 1: $\widetilde{\alpha}_{L,C} = \lambda$**.** We can rewrite $\widetilde{W}_{LS}\left(\widetilde{\boldsymbol{\alpha}}_n, n\right)$ and $\Pi_{L,C}(\boldsymbol{\alpha}_n)$ as:

$$\widetilde{W}_{LS}(\widetilde{\alpha}_S, n) = \frac{n(n+2)}{8b(n+1)^2}\left(4(\bar{A}-MC)^2 + (1-\gamma_S^n)(A_H-A_L)^2\right),$$
$$\Pi_{S,C}(\widetilde{\alpha}_S, n) = \frac{\gamma_S^n(2\theta-1)(A_H-A_L)(\bar{A}-MC)}{2b(n+1)}$$

where $\gamma_S = 1-(\lambda+\widetilde{\alpha}_S)(2\theta-1)$.

Then, we can calculate the following partial derivatives:

$$\frac{\partial \widetilde{W}_{LS}}{\partial \widetilde{\alpha}_{S,C}} = \frac{\gamma_S^{n-1}n^2(n+2)(2\theta-1)(A_H-A_L)^2}{8b(n+1)^2},$$
$$\frac{\partial \widetilde{W}_{LS}}{\partial n} = \frac{\gamma_S^n n(n+2)(A_H-A_L)^2\ln(1/\gamma_S)}{8b(n+1)^2} + \frac{2\left((A_H-MC)^2+(A_L-MC)^2\right)-\gamma_S^n(A_H-A_L)^2}{4b(n+1)^3}$$
$$\frac{\partial \Pi_{S,C}}{\partial \widetilde{\alpha}_{S,C}} = -\frac{n\gamma_S^{n-1}\left(\bar{A}-MC\right)\left(A_H-A_L\right)(2\theta-1)^2}{2b(n+1)}$$
$$\frac{\partial \Pi_{S,C}}{\partial n} = -\frac{\gamma_S^n(2\theta-1)\left(\bar{A}-MC\right)\left(A_H-A_L\right)\left(1+(n+1)\left(\ln 1/\gamma_S\right)\right)}{2b(n+1)^2}$$

By the implicit function theorem, we have:

$$\frac{\partial \widetilde{\alpha}_{S,C}}{\partial n} = -\frac{\partial \Pi_{S,C}/\partial n}{\partial \Pi_{S,C}/\widetilde{\alpha}_{S,C}} = -\frac{\gamma_S\left(1+\ln(1/\gamma_S)\right)}{n(2\theta-1)}$$

which further implies:

$$\frac{d\widetilde{W}_{LS}\left(\widetilde{\alpha}_{S,C}, n\right)}{dn} = \frac{\partial \widetilde{W}_{LS}}{\partial n} + \frac{\partial \widetilde{W}_{LS}}{\partial \widetilde{\alpha}_{S,C}}\frac{\partial \widetilde{\alpha}_{S,C}}{\partial n} = \frac{\left(A_H-A_L\right)^2\left(G_1 - g_{S,C}\left(\gamma_S, n\right)\right)}{8b(n+1)^3}.$$

Thus, $\frac{d\widetilde{W}_{LS}\left(\widetilde{\alpha}_{S,C}, n\right)}{dn} < 0$ if and only if $g_{S,C}(\gamma_S, n) > G_1$.

Case 2: $\widetilde{\alpha}_{S,C} = 0$**.** We can rewrite $\widetilde{W}_{LS}\left(\widetilde{\boldsymbol{\alpha}}_n, n\right)$ and $\Pi_{L,C}(\boldsymbol{\alpha}_n)$ as:

$$\widetilde{W}_{LS}(\widetilde{\alpha}_{L,C}, n) = \frac{n(n+2)}{8b(n+1)^2}\left(4(\bar{A}-MC)^2 + (1-\gamma_L^n)(A_H-A_L)^2\right),$$
$$\Pi_{S,C}(\widetilde{\alpha}_{L,C}, n) = \frac{n\gamma_L^n(2\theta-1)\left(\bar{A}-MC\right)\left(A_H-A_L\right)}{2b(n+1)}$$

where $\gamma_L = 1-\widetilde{\alpha}_L \times (2\theta-1)$.

Then, we can calculate the following partial derivatives:

$$\frac{\partial \widetilde{W}_{LS}}{\partial \widetilde{\alpha}_L} = \frac{\gamma_L^{n-1} n^2 (n+2)(2\theta-1)(A_H - A_L)^2}{8b(n+1)^2},$$
$$\frac{\partial \widetilde{W}_{LS}}{\partial n} = \frac{n(n+2)\gamma_L^n (A_H - A_L)^2 \ln(1/\gamma_L)}{8b(n+1)^2} + \frac{2\left((A_H - MC)^2 + (A_L - MC)^2\right) - \gamma_L^n (A_H - A_L)^2}{4b(n+1)^3}$$
$$\frac{\partial \Pi_{L,C}}{\partial \widetilde{\alpha}_L} = -\frac{n^2 \gamma_L^{n-1}\left(\bar{A} - MC\right)(A_H - A_L)(2\theta-1)^2}{2b(n+1)}$$
$$\frac{\partial \Pi_{L,C}}{\partial n} = \frac{\gamma_L^n (2\theta-1)\left(\bar{A} - MC\right)(A_H - A_L)\left(1 - n(n+1) ln\left(1/\gamma_L\right)\right)}{2b(n+1)^2}$$

By the implicit function theorem, we have:

$$\frac{\partial \widetilde{\alpha}_{L,C}}{\partial n} = -\frac{\partial \Pi_{L,C}/\partial n}{\partial \Pi_{L,C}/\partial \widetilde{\alpha}_{L,C}} = \frac{\gamma_L \left(1 - n(n+1)\ln(1/\gamma_L)\right)}{n^2(n+1)(2\theta-1)}$$

which further implies:

$$\frac{d\widetilde{W}_{LS}\left(\widetilde{\alpha}_{L,C}, n\right)}{dn} = \frac{\partial \widetilde{W}_{LS}}{\partial n} + \frac{\partial \widetilde{W}_{LS}}{\partial \widetilde{\alpha}_{L,C}} \frac{\partial \widetilde{\alpha}_{L,C}}{\partial n} = \frac{4\left((A_H - MC)^2 + (A_L - MC)^2\right) + n\gamma_L^n (A_H - A_L)^2}{8b(n+1)^3}$$

Obviously, $\frac{d\widetilde{W}_{LS}\left(\widetilde{\alpha}_{L,C}, n\right)}{dn} > 0$. The proof concludes. □

B.7.5 Proof of Lemma B.5

Proof. First, note that $B(n) = B_0$ eliminates the impact of the benefits of liquidity trading and thus we can focus on the information cost. Second, $\Phi(m) - m * \tilde{c} > 0$ holds for some $m \geq m_0$ for $\frac{2b}{(\bar{A}-MC)^2}$ sufficiently small since $\Phi(m) > 1$ for $m \geq m_0 + 1$. Third, note that

$$\frac{\overline{W}\left(\widehat{\alpha}_m, m\right)}{\overline{W}\left(\widehat{\alpha}_n, n\right)} = \frac{\left(1 - \frac{1}{(m+1)^2}\right) * \left(1 + \mu * \left(1 - (2-2\theta)^m\right)\right) - m * \tilde{c}}{\left(1 - \frac{1}{(n+1)^2}\right)}$$

Then, the remaining proof follows from that of Proposition 3. The proof concludes. □

Reinforcement learning in a dynamic limit order market*

Amy Kwan† **Richard Philip**‡

March 6, 2025

Abstract. What drives the value of limit orders? We use a novel machine learning approach to investigate optimal limit order management and the factors affecting the expected value of an order. A limit order is more valuable if positioned towards the front of the queue and when there is a large queue resting behind the order. When trading is constrained by minimum tick size requirements, volatility decreases the value of the order, but increases its value when trading is unconstrained. Further, the option to cancel an order is economically meaningful, contributing approximately 19% of the order's total expected value. This study uncovers pervasive market dynamics, advancing our understanding of financial markets.

Key words: Limit order markets, machine learning, big data, queue size, optimal limit order

JEL: G10; G20

*This paper has benefited from the comments of Michael Brolley, James Brugler, David Cimon, Vincent van Kervel, Pete Kyle, Tom McInish, Albert Menkveld, Ryan Riordan, Andriy Shkilko, Ester Felez Vinas, Gideon Saar, Wing Wah Tham, Yajun Wang, Ying Wu, Chen Yao, Marius Zoican, and the audiences at the SFS Cavalcade Asia Pacific, FIRN annual meeting, and the Microstructure Exchange. Thank you to seminar participants at The University of Sydney, The University of New South Wales, University of Wollongong, Wilfrid Laurier University, and the team at Vivienne Court Trading for their insightful comments.

†University of New South Wales, Australia, e-mail: amy.kwan@unsw.edu.au

‡University of Sydney, Australia, e-mail: richard.philip@sydney.edu.au

1 Introduction

Limit order submissions and cancellations make up a staggering 95% of trading activity in modern markets.[1] In response, exchanges and regulators have proposed measures to curb message frequency by imposing limitations on the order to trade ratio, enforcing minimum order resting times, and introducing message taxes or cancellation fees. Despite these initiatives, we know very little about how liquidity providers should manage their limit orders. What is the value of a limit order? At what price level should we submit a limit order? When should an order be canceled? How often should an order be cancelled? How important is this option to cancel? Answering these questions is non-trivial; the dimensionality of the problem is extremely large and decisions are path dependent.

Despite the complexity of the problem, theory has shed some light on the way traders manage their orders. Among others, Parlour (1998), Foucault (1999), Goettler et al. (2005), Foucault et al. (2005), Goettler et al. (2009), Rosu (2009), Ricco et al. (2020), Rosu (2020), and Bhattacharya and Saar (2022) propose multi-period equilibrium models, which represent limit order markets as sequential games. In these models, traders arrive sequentially and submit, or update, the optimal order that maximizes their gains from trade. However, these models differ in the features that are modeled. For example, some models highlight the importance of volatility (Foucault (1999)) while others demonstrate the importance of queue size (Parlour (1998)). In some models, traders can only submit to one price level (Parlour (1998)) while in other models, traders can submit to prices beyond the best quotes (e.g., Goettler et al. (2005)). In Goettler et al. (2005), Foucault et al. (2005), and Ricco et al. (2020) traders can enter the market once, while in Goettler et al. (2009), Rosu (2009) and Bhattacharya and Saar (2022) traders can reenter the market. However, which market features are most important to a trader's optimal order decisions? Many of the features of these models have not been empirically tested due to the lack of technologies available to researchers.

In this study, we uncover the most important features influencing a liquidity provider's limit order decisions using a novel machine learning (ML). For over 18,000 unique market states, we compute the expected value of a resting limit order for each of these market states, conditional

[1] See Brogaard et al. (2019). Market orders, which have been the focus of much of the existing literature, make up less than 5% of all activity.

on the optimal management of the order over its life cycle. Our technique allows us to identify important features of limit order management and in doing so, provide several stylized facts about limit orders that is new to the literature. First, we quantify the value of a resting limit order under a broad combination of different market conditions. This allows us to identify when it is optimal to leave or cancel a resting limit order under different market conditions. Second, we uncover pervasive market dynamics and show which features drive the underlying value of the limit order and how their interactions interplay. Finally, we quantify the value of the option to cancel an order and identify the market conditions when this option is most valuable.

To solve this problem, we cast limit order management as a sequential Markovian decision process within a reinforcement learning (RL) framework. RL is a type of machine learning that enables an agent to learn the optimal action, given the current environment, using feedback from the agent's own actions and experiences. We emphasize that our RL framework is not a conventional theoretical model, which typically models trader behavior to arrive at equilibrium outcomes. Rather, our RL framework imposes a structure onto the vast amount of empirical data to identify the features of theoretical models that contribute most to the trader's order submission decision.

In our RL framework, at short periodic time intervals, our risk neutral liquidity provider faces the same decision: to leave or cancel their resting limit order. This decision making process repeats until the trader's limit order executes or is canceled. For each periodic decision, our liquidity provider maximizes expected profit and leaves (cancels) their limit order if the order has a positive (negative) expected value conditional on 1) the current market conditions and 2) the future optimal management of the limit order. Thus, our framework captures the endogenous option to cancel based on the future expected value of the order's payoff. As a result, the limit order's conditional expected value at time t is a recursive estimate based on all future conditional expected values and their corresponding likelihoods. To overcome the recursive nature of the problem, we empirically estimate the conditional expected value via an iterative update function, known as Q-learning.

The key estimate in our liquidity provider's decision making process is the limit order's conditional expected value. The expected value of a limit order is driven by a tradeoff between two opposing dynamics: the order's probability of execution, which enhances its value, and its risk of

adverse selection, which diminishes the order's value. We draw insights from existing theoretical literature to identify the variables or market conditions that influence adverse selection risk or its execution probability—thus, contributing to the limit order's overall conditional expected value. Parlour (1998) provides theoretical arguments that strategic traders should consider queue lengths on both sides of the limit order book. Further, Yueshen (2021), Li et al. (2020) and Yao and Ye (2018) argue that there is an advantage to being at the front of the queue, due to the time priority rule. Last, Foucault (1999) finds that volatility is a main determinant for limit order management. Using these concepts, we define a state space for a bid order, which considers the lengths of the queues on the first three levels of the bid side of the order book and the length of the queue on the best ask price. The bid limit order can sit at the best bid, one tick behind the bid, or two ticks behind the bid. We also consider the limit order's position within the queue and volatility. For tractability, we estimate a model in which we discretize these features, resulting in a state space of 18,001 unique market states. At any point in time, the limit order exists in one of the market states, which then transitions to a different market state in the future. Because our model is completely data driven, our framework provides the flexibility to use alternate features to define the state space. For example, the framework can be adapted to investigate a trader's choice between a market or limit order, determine the optimal order size given current market conditions, or consider factors such as a trader's inventory, risk tolerance, and private information. We further explore these capabilities in Section 5.

Our approach is also the first attempt in the literature to estimate the impact of factors that could affect the limit order's value. We show that the average expected value of a limit order resting at the best bid is approximately one quarter of a tick. The value of the limit order drops off substantially as we move away from the best quotes: the expected value of limit orders resting at one and two levels behind the best bid are 0.10 ticks and 0.03 ticks, respectively. There is also large variation in the expected value of a limit order. For example, our findings reveal that for all price levels, a resting limit order loses almost half its expected value when it transitions from the front to the back of the queue, which supports the theoretical predictions of Yueshen (2021), Li et al. (2020) and Yao and Ye (2018), who show that queue priority is advantageous. Second, we extend the findings of Parlour (1998), and quantify the importance of queue size. Specifically, we

show that the expected value of an order increases with the queue size resting behind the order, and decreases with the queue size in front of the order. Third, the expected value of a limit order resting at the best price decreases with an increase in the opposing queue size.

Last, we show that volatility is important for the limit order's value, with its effect contingent on whether the stock is tick constrained, consistent with Li et al. (2020). Foucault (1999) predicts that volatility has two opposing forces on a limit order's profitability. The first force suggests an increase in volatility *decreases* the expected value of a limit order via an increase in the risk of adverse selection. However, the second force suggests an increase in volatility *increases* the expected profit of a limit order as liquidity providers counteract losses from an increase in adverse selection risk by widening the bid ask spread. Further, Li et al. (2020) show that the minimum tick size also plays an important role. If the breakeven bid ask spread is always less than the one tick-mandated spread, liquidity providers do not widen the bid ask spread to compensate for the increased picking off risk even in times of high volatility. Thus, volatility decreases the expected value of the limit order as the compensation for providing liquidity (i.e., the difference between the quoted and breakeven bid ask spread) falls but remains positive.

Conversely, if volatility increases such that the breakeven bid ask spread widens beyond the one tick-mandated spread, liquidity providers react by widening their quoted bid ask spread to compensate themselves for the additional risk. In doing so, an increase in volatility increases the value of the limit order. Consistent with these predictions, we find that volatility has mixed effects depending on whether the stock is tick constrained. For stocks that are most tick constrained, an increase in volatility decreases the expected value of a limit order at the best price. On the other hand, we show that the value of a limit order increases with volatility for stocks that are most tick constrained.

Our RL approach also enables a comparative analysis of these market features while simultaneously accounting for the intricate interdependencies among them. Our analysis ranks the price level at which a limit order is placed as the most critical variable for traders to consider. This finding suggests that it is important for theory models to consider order submissions at or away from the quote as in Goettler et al. (2005). Following the price level, the subsequent factors in order

of importance are queue sizes at different price levels, market volatility, and the queue position of the order.

How valuable is the option to cancel a limit order? We find the option to cancel represents 19% of a limit order's total expected value, on average. This option becomes even more valuable during periods of high *ex-ante* adverse selection risk. In the most extreme case, we demonstrate that a limit order, which would otherwise have a negative expected value, can have a positive expected value purely because of the option to cancel the order at a later time.

The advantage of our approach is four-fold. First, similar to Goettler et al. (2005) and Goettler et al. (2009), our approach can handle a state-action space with large dimensionality. In Section 5, we demonstrate that our general framework can be extended to encompass a wide range of scenarios. For instance, this framework allows us to explore decisions involving the choice between limit and market orders, various order sizes, and the inclusion of factors like the liquidity provider's risk aversion or current inventory levels.

Second, we can estimate the option value in cancelling a limit order because our limit order's expected value estimates are conditional on the future endogenous option to cancel as in Goettler et al. (2009). By considering the option to cancel, we are able to determine the trader's optimal limit order placement, conditional on the optimal management of the order over its life. This approach differs from most of the previous empirical work that uses probability models, which capture the outcome of these order placement strategies, regardless of its optimal management (e.g., Griffiths et al. (2000), Ranaldo (2004), Ellul et al. (2007), Goldstein et al. (2023)).

Third, we complement traditional theory models in that our approach is completely driven by data enabling us to empirically assess existing theories. Similar to Sandås (2015), who tests Glosten (1994) via a structural model, we use a structural RL model, which removes the need for assumptions about trader behavior or market dynamics. By removing assumptions about trader behaviour, we determine optimal order management under real market conditions, where traders may not necessarily behave rationally, or follow a stylized set of assumptions. Last, our analysis draws variables from the theoretical literature and thus, we overcome the 'black-box' nature of ML techniques, which can obscure economic intuition (see Chinco et al. (2019)).

We contribute to the literature in three ways. First, our approach is the first attempt to quantify the value of a limit order and to systematically assess the factors influencing its value. In doing so, we can explore previously untested theoretical predictions and uncover new interactions in financial markets. While existing theoretical models highlight the importance of queue size and queue priority (Parlour (1998), Yueshen (2021), Li et al. (2020) and Yao and Ye (2018)), we quantify the value of queue position to a liquidity provider across various market conditions. We empirically show that volatility has mixed effects on the value of a limit order depending on the degree of tick constraint, which is consistent with the predictions in Foucault (1999) and Li et al. (2020). Importantly, because ML techniques are well suited to complex environments characterized by multiple variables, we can evaluate these relations while holding all other market conditions constant.

Second, our research contributes to the literature on order cancellations. In Copeland and Galai (1983), when a market maker posts a bid or offer, they effectively write an option. However, this order also grants the market maker an option to cancel the order at some future point in time. Despite the substantial increase in order cancellations, constituting 47% of all messages (Brogaard et al. (2019)), understanding the implications and value associated with the option to cancel an order remains limited. Our study complements Dahlström et al. (2023), who investigate the determinants of order cancellations by liquidity providers, by highlighting the economic significance of the option to cancel.

Finally, we contribute to the growing literature applying learning algorithms to financial markets. Similar to this study, Bhattacharya and Saar (2022) use a recursive procedure to solve their model of dynamic limit order markets and Ait-Sahalia and Saglam (2023) model the high frequency trader's optimization problem as a Markov Decision Process. Dou et al. (2024) explore how AI-powered trading algorithms, specifically those combining algorithmic trading with reinforcement learning, impact price efficiency in a theoretical framework. In Colliard et al. (2022), algorithmic market makers set quotes using Q-learning algorithms and their trading outcomes are compared to the outcomes predicted by theory. We also apply a Q-learning approach to evaluate the decision making process of liquidity providers but in contrast to Colliard et al. (2022) and Dou et al. (2024), our analysis tests existing theory using empirical data. O'Hara (2015) highlights that in the modern era, markets and trading have changed, with limit orders now playing a more crucial role. Similarly,

Easley et al. (2021) issue a call to update the learning models and empirical methods used. Our paper answers this call by proposing a novel technique that provides a deeper understanding of limit order management than traditional learning models or empirical methods allow.

2 Method

2.1 Intuition

Consider a liquidity provider, or trader, who wants to optimally manage their limit orders to ensure that only limit orders with a positive expected value execute.[2] The dynamics of the limit order book make this task non-trivial, as the trader must constantly monitor their resting limit orders and cancel an order if it is expected to lose money. To achieve this task, the trader must estimate the expected value of a limit order conditional on the current state of the market *and* the future optimal management of the order over its life cycle.

Estimating the expected value of a limit order, conditional on its future optimal management, requires the trader to consider the evolution and likelihood of various market conditions. The trader must evaluate these conditions until one of two events occurs: 1) the order is executed, or 2) the order is canceled. The decision to cancel an order is endogenous and should be made when the limit order has a negative expected value.

Figure 1 illustrates the trader's problem. Initially, the limit order book is in a certain state at time t_0. The gray rectangles represent the volume available at the ask prices, and the white rectangles represent the volume available at the bid prices. The best bid and ask prices are 13 and 14, respectively, resulting in a bid ask spread of 1. In Figure 1, we assume a trader submits a limit buy order at t_0 at a price of 12 (one tick behind the best available bid) and depict this order as a black rectangle.

[Insert Figure 1]

[2]Similarly, automated market makers use a learning algorithm to pick the price that generates the largest expected profit in Colliard et al. (2022).

The trader then monitors the limit order book until the volume on the current best bid is removed, which occurs at t_1. For illustrative purposes, we assume the market evolves into one of only two possible states at t_1: State A or State B. In State A, since t_0, other market participants have submitted buy limit orders at 12, causing our trader's order to move up the queue at 12. Further, market participants have added buy limit orders at 11, and some of the sell limit orders at 14 have been removed, either due to cancellations or executions. In contrast, in State B, no new market participants have submitted additional buy limit orders. Instead, a large sell limit order at 13 has been submitted, removing the bid at 13 that existed at t_0.

If the volume available on the bid side of the order book is significantly larger (or smaller) than the volume available on the ask side of the order book, the midprice is more likely to increase (or decrease) in the near future (see Cao et al. (2009)). Therefore, the order in State A has a positive expected value, as the volume on the bid side is much larger than the volume on the ask side, suggesting a future price rise. In contrast, the order in State B has a negative expected value, as the volume available on the ask side is much larger than the volume available on the bid side, indicating the price is likely to decline in the future and the order would be adversely selected.

The expected value of the limit order submitted at t_0, if left unmonitored, is the sum of the expected values in State A and B, each weighted by their respective probabilities. Therefore, if the probability of transitioning to State B is much higher than the probability of transitioning to State A, the expected value of the unmonitored limit order at t_0 could be negative. However, if monitoring and the option to cancel the limit order are allowed, the expected value of the order becomes positive. This is because the trader will cancel the order if the market transitions to State B, resulting in a profit of 0, and leave the order if the market transitions to State A, where it has a positive expected value. This oversimplified example demonstrates that the option to cancel can transform an order from having a negative expected value to a positive one.

In this illustrative example, we make two tenuous assumptions. First, we assume the market can only transition to two possible states once the trader's order is submitted. In reality, the market can transition to an almost infinite number of states. Second, we arbitrarily assert that the limit order has a positive expected value in State A and a negative expected value in State B. Instead of

relying on these arbitrary assertions, we can achieve a precise estimate of the true expected values for State A and State B at time t_1 by calculating the expected value of the limit order within both states, conditional on the order's optimal management throughout its life cycle. This problem presents the same challenge we are attempting to address at t_0.

To overcome these limitations, we use a recursive state space technique known as reinforcement learning (RL). This technique allows us to accommodate numerous states and capture the inherently recursive nature of the problem.

2.2 Reinforcement Learning

Typically in an RL framework, an agent has knowledge of the current state, s, and then makes an action, a. Jointly, we refer to this state-action pair as an experience tuple defined as $\langle s, a \rangle$. If there are S states and A actions, then the agent has the choice of making A possible actions in S different states, which implies there are $S \times A$ unique experience tuples. We assume that each experience tuple can transition the agent to a new state, s', with probability $T(\langle s, a \rangle, s')$. For each action in a given state, the agent receives an immediate reward, $R(s, a)$. The agent's objective function is to maximize the total future reward by choosing the appropriate actions for each state that maximize the long-run discounted sum of all the immediate rewards received for each action in the future.

More formally, if we define the rules or policy an agent must follow as π, the optimal value of a state is computed as follows:

$$V^*(s) = \max_{\pi} E\big(\sum_{t=0}^{\infty} \gamma^t E[R(s_t, a_t)]\big), \tag{1}$$

where $E[R(s_t, a_t)]$ is the expected immediate reward at time t and γ is a discount factor bound between 0 and 1. $V^*(s)$ is the expected infinite discounted sum of reward the agent receives if they start in state s and execute the optimal policy defined by π^* moving forward. In our setup, the optimal policy, π^*, defines how the trader should optimally manage their limit order moving

forward (i.e., the action the trader should take given current market conditions and current order positioning). Similarly, the reward is the profit generated from earning the spread or favorable price movements after the order executes.

For every experience tuple, there is an associated Q-value, $Q^*(s,a)$, which is the expected infinite discounted sum of reward the agent gains if the agent takes action a while in state s, then subsequently follows the optimal policy path. Using (1), we note that $Q^*(s,a)$ can be expressed recursively as:[3]

$$\underbrace{Q^*(s,a)}_{\substack{\text{long run expected value}\\ \text{from taking action } a}} = \underbrace{E[R(s,a)]}_{\substack{\text{expected immediate}\\ \text{value from}\\ \text{taking action a}}} + \gamma \underbrace{\sum_{s' \in S} \overbrace{T(\langle s,a\rangle, s')}^{\substack{\text{probability of}\\ \text{transitioning to}\\ \text{future state } s'\\ \text{by taking}\\ \text{action } a}} \overbrace{\max_{a'}(Q^*(s',a'))}^{\substack{\text{expected long}\\ \text{run value from}\\ \text{taking optimal}\\ \text{action } a' \text{ when}\\ \text{in state s'}}}}_{\substack{\text{expected future value}\\ \text{from taking future optimal}\\ \text{actions, } a', \text{ while in future states, } s'}}, \tag{2}$$

where s' and a' define future states and actions, respectively. Equation (2) is the basis of our framework. In our setup, $Q^*(s,a)$ is the expected long run value of the limit order if the trader takes action a while in state s and in all future states s' takes the optimal action a'. We observe that this expected long run value equals any immediate value for taking action a plus the expected long run value the trader receives in future state s' if they make optimal future action a'. Recognizing that the future state s' is not known with certainty, our RL model assigns different transition probabilities, $T(\langle s,a\rangle, s')$, for all possible future states. Equation (2) is recursive because both the right hand side and the left hand contain a $Q^*(s,a)$ term. Thus, for estimation we use an iterative learning rule known as Q-learning.[4]

Estimating (2) requires us to first define a state-action space that reflects the problem of optimal limit order management. Specifically, the states should capture current market conditions and information about the order, while the actions should reflect the decisions available to the trader. Next, estimation requires two key input variables: the immediate reward and the transition probabilities. In the following sections, we describe how we cast the optimal limit order management

[3] See Watkins and Dayan (1992) for a full derivation.

[4] We provide a detailed illustrative example of the learning rule in Appendix C

problem within the RL framework. We explain the basic timing of our trader's decision process, define our state and action space, and describe how we empirically estimate the input variables: the immediate reward and the transition probabilities.

2.2.1 Timing

Figure 2 depicts the timing of our trader's decisions. In essence, the trader follows a recursive Markovian decision making system. At the start of each interval, the trader makes a decision based on observations of the current market conditions, for example, the existing shape of the order book and their own private information about their limit order's status. The trader decides whether to leave or cancel their existing limit order. At the start of the subsequent interval, the trader repeats the same decision making process. This decision-making process repeats continuously until the limit order executes or the order is canceled. If the limit order executes, the trader continues to monitor market conditions to observe the long-term value of the executed order.

[Insert Figure 2]

This recursive decision-making system allows the trader to keep the same limit order active for multiple consecutive intervals. During this time, the trader can monitor the order's queue position and market conditions. If at any point the order appears to have a high chance of adverse selection, indicated by a negative expected value, the trader cancels the order.

In our empirical section, we select a short time interval of 100ms. Choosing a short time interval offers three advantages. First, a short interval more closely reflects a trader who continuously monitors their orders. Second, a shorter interval provides more data points for model estimation. Third, it allows us to produce more accurate estimates of the likelihood of transitioning to future market conditions, as dramatic changes are less likely to occur over short intervals.

2.2.2 Actions

The A actions available define all possible decisions or individual actions, a, a trader can make given the current state. In our setup, the trader can make two possible actions. The trader can either cancel their resting limit order, which we define as C, or the trader can leave their existing limit order in the queue by taking no action, which we define as NA. Taken together, the trader's action space is defined by

$$a \in \{C, NA\}. \tag{3}$$

Figure 2 depicts the timing of the actions. Specifically, the trader decides on an action at the beginning of the interval. To ensure that the trader's limit order remains at price levels within our defined state space, we make the following adjustment: when the market transitions to a state in which limit order lies outside the state space, then the action C supersedes action NA. This implies that the resting limit order is canceled. This adjustment forces the trader to cancel resting limit orders if the best bid and offer has diverged away from the trader's resting limit order.

2.2.3 States

The state, s_t, reflects information available to the trader about the environment at time t. We decompose the environment into two sets of variables that reflect the current state: private and public. The public variables represent current market conditions available to all market participants. Parlour (1998) suggests that queue sizes in the limit order book influence the strategic behavior of traders. For this reason, we include the size of the queue at the best bid, one tick below the best bid and two ticks below the best bid, which we define as q^{B_0}, q^{B_1} and q^{B_2}, in our state space. Similarly, we include the size of the queue on the opposing side of the book (the best ask), which we define as q^{A_0}. Given queue sizes are essentially continuous, for tractability, we reduce the dimensionality of the state space by discretizing queue sizes. Specifically, we categorize queue lengths into five quintiles; extremely long (ELo), long (Lo), normal (No), short (Sh) and extremely short (ESh).[5]

[5]To further reduce dimensionality, we discretize the queue size at q^{B_2} to only three terciles.

Moreover, Foucault (1999) finds that volatility is a main determinant for limit order management. For this reason, we also include volatility, V, as a public variable, which we discretize into terciles; low (Low), medium (Med) and high (Hi).[6]

The private variables we use to define our state space capture information that is unique to the trader. Specifically, we capture the trader's current inventory position, I, which in our model is either 0 (no position) or 1 (long). We also include a variable, L, which captures the price level of the trader's limit order. We let L take on the value of $i \in 0, 1, 2$ if the trader has a resting limit order submitted at level i of the order book. Finally, because there is an advantage to being at the top of queue as the order has time priority (see Yueshen (2021), Li et al. (2020) and Yao and Ye (2018)), we include the queue position of any resting limit orders in our state space, which we define by Q. Similar to our previous variables, for tractability, we reduce the dimensionality of our queue position to five quintiles, which we define as *top*, *top-middle*, *middle*, *middle-back* and *back*.

Last, to ensure we estimate the expected value of a single limit order in isolation, we include a state that captures when the trader cancels their order. This state is a terminal absorbing state where the trader remains once they cancel their order. We define this terminal state by setting $Q = X$ and $L = X$. Taken together, these definitions let us express the current market state, s, as a vector

$$s = [I, L, Q, q^{B_0}, q^{B_1}, q^{B_2}, q^{A_0}, V] \tag{4}$$

[6]To proxy for volatility, we compute the difference between the log of the highest and the log of the lowest traded price over the last 100 trades in the stock.

where

$$I \in \{0, 1\}$$

$$L \in \{0, 1, 2, X\}$$

$$Q \in \{top, top\text{-}middle, middle, middle\text{-}back, back, X\}$$

$$q^j \in \{ELo, Lo, No, Sh, ESh\}, \forall j \in \{B_0, B_1, B_2, A_0\}$$

$$V \in \{Low, Med, Hi\}$$

In our setup, we restrict the trader to executing only one limit order. We achieve this restriction by ensuring no additional orders exist once a long position is achieved. As a result, the states when the trader is long are only defined by the four public limit order book information variables and volatility $(q^{B_0}, q^{B_1}, q^{B_2}, q^{A_0}, V)$. This restriction implies there are m possible states when the trader is long.[7] In contrast, when the trader has no inventory and is working their limit order, the state is defined by the public variables (i.e., queue sizes and volatility) plus the private variables (i.e., queue position and the price level of the resting limit order). The additional private variables results in n possible states when the trader has no inventory.[8] Collectively, in this setup, we have n states when the trader has no inventory, m states when the trader is long and one absorbing state for when the trader cancels their order, thereby resulting in $m + n + 1$ total possible states, where $n > m$ and $m + n + 1 = S$.

2.2.4 Transition matrix

With the states and action defined, we require estimates of the transition probabilities. Recall that if the limit order is currently in state s and the trader makes action a, the order transitions to states s' with probability $T(\langle s, a\rangle, s')$. Since the transition probabilities from state i to all other

[7] In our setup $m = 1,125$ as we have three public limit order book information variables $(q^{B_0}, q^{B_1}, q^{A_0})$, each with five possible values, one order book information variable with three possible values (q^{B_2}) and one volatility variable (V) with three possible values. Thereby resulting in $5^3 \times 3 \times 3$ possible combinations.

[8] In our setup $n = 16,875$. We have 1,125 possible public states, plus the private price level and queue position variables, which have three and five possible values respectively. Collectively, these variables result in $1,125 \times 5 \times 3$ possible combinations.

states must sum to 1 for a given action, for all i and a, $\sum_{j=1}^{S} T(\langle s_i, a\rangle, s_j) = 1$.

Because our framework has S unique market states, each action has an $S \times S$ transition probability matrix. When the trader makes no action (i.e., action NA), which leaves their resting limit order, the future state the limit order transitions to is not known with certainty. Thus, we empirically estimate the $S \times S$ transition probabilities for action NA. To estimate $T(\langle s_i, NA\rangle, s_j)$ we determine the number of times we observe a limit order in state s_i, followed by the limit order being in state s_j in the subsequent interval, and express this number as a fraction of all observations of limit orders in state s_i. More formally, if we define $N_{i,j}|NA$ as the number of times a limit order in state i transitions to state j, it is straightforward to show that the MLE estimate of $T(\langle s_i, NA\rangle, s_j)$ is

$$T(\langle s_i, NA\rangle, s_j) = \frac{N_{i,j}|NA}{\sum_{j=1}^{S} N_{i,j}|NA}. \tag{5}$$

In contrast, when a trader cancels their limit order, they transition to the absorbing cancel state with certainty. For this reason, we do not require empirical estimates for the $S \times S$ transition probabilities for action C, as the probability of transitioning to the absorbing cancel state is always 1. To ensure the trader only has one resting limit order, we restrict any state where the trader has an inventory position, or has already canceled their order, to not having a resting order. Because of this restriction, the action to cancel is prohibited and has a zero probability for all states where the trader has a long inventory position, or has canceled their order.[9]

To generate the full transition matrix, T, that captures all state actions, we vertically stack the $S \times S$ transition matrix for action NA on top of the $S \times S$ transition matrix for action C.

2.2.5 Immediate reward

An action from a given state can transition the trader to a new state and produce an immediate reward in the process. In our setup, the immediate reward captures any value, or profit, generated

[9] In Appendix A, we provide a detailed description of the structure and design of the transition matrices for each action.

during the transition from the current state to the next. This profit is derived from two sources: 1) price movements while carrying inventory, and 2) earning the spread through limit order execution. If the trader holds inventory when transitioning from state s to s', the immediate reward for this transition is the observed change in the midpoint during the transition (first component of equation (6)). Alternatively, if the trader's limit order executes during the transition from s to s', they profit by earning the spread. In this case, the immediate reward is the difference between the midpoint price observed in state s' and the execution price of the limit order (second component of equation (6)). Finally, if the trader has no inventory in state s and no order executes during the transition from state s to s', then the immediate reward must be zero. Formally, defining the midpoint price in state i as mid_i, the immediate reward from making action a while in state s that results in a transition to state s' is

$$R(\langle s, a\rangle, s') = \underbrace{(mid_{s'} - mid_s) \times I_s}_{\substack{\text{Profit from carrying} \\ \text{inventory}}} + \underbrace{(mid_{s'} - execPrice_s) \times Exec_{s,s'}}_{\text{Profit from execution}}, \tag{6}$$

where I_s equals 1 if the the trader has a long inventory position when in state s and 0 otherwise and $execPrice_s$ equals the price of the limit order and $Exec_{s,s'}$ equals 1 if the limit order executes during the transition from state s to s' and zero otherwise.

To compute the immediate reward for each transition, we require empirical estimation when the trader leaves their order (action NA). To obtain these estimates, we first compute the immediate reward using equation (6) for every observation in the data. Then, for each state-action transition, we compute the average immediate reward across all observations that belong to that state-action transition.[10] In contrast, when the trader cancels their order, the immediate reward must be zero as they have no limit orders executed and no inventory position. Therefore, for action C, the $S \times S$ immediate reward matrix contains only zeros.

Similar to the transition matrix, we create the immediate reward matrix for all experience

[10] In Appendix B, we provide a detailed description of the structure and design of the immediate reward matrices for action NA.

tuples by vertically stacking the immediate reward matrix for action NA and the immediate reward matrix for action C, resulting in a matrix of dimension $2S \times S$. We compute the expected immediate reward for taking action a while in state s as

$$E[R(s,a)] = \sum_{s' \in S} T(\langle s,a \rangle, s') \times R(\langle s,a \rangle, s'). \tag{7}$$

3 Data

We use ITCH data for the Australian Securities Exchange (ASX) extracted from the SIRCA database for the period July 3, 2017 to September 29, 2017. Table 1 contains summary statistics for the 20 sample stocks analyzed, ranging from the lowest price stock of Santos (STO), with a price of approximately 3.50 over the sample period to CSL Ltd. (CSL) with an average price of 129.86. The sample stocks also cover a wide range of average bid ask spreads, from 1.00 tick to 2.59 ticks.

[Insert Table 1]

The ITCH data provides comprehensive order book information with nanosecond-level timestamps, allowing us to fully reconstruct the order book at all price levels. We extract detailed information for each resting limit order, including its queue position. To facilitate our analysis, we process the data according to the following steps. First, we reconstruct the limit order book, enabling us to replay the market activity throughout a trading day. Second, for each trading day, we create 210,000 consecutive intervals, each 100 milliseconds long. The first interval begins at 10:10, coinciding with the start of continuous trading, and the last interval concludes at 16:00 when continuous trading ends.

At the beginning of each interval, we assume there is a series of hypothetical limit orders positioned at various price levels and queue positions. To align with our RL model, we consider hypothetical bids for one share placed at the prevailing best bid, one tick behind the best bid,

and two ticks behind the best bid. Additionally, at each of these price levels, we assume there is a hypothetical order positioned at the top of the queue, three-quarters of the way up the queue, halfway up the queue, one-quarter up the queue, and at the very back of the queue.[11]

Next, using the granularity of the data, we track these hypothetical limit orders over the next 100ms and determine if any of the orders execute.[12] In the event that the hypothetical limit order remains unexecuted within the 100ms timeframe, we monitor its progress by tracking the execution of real orders that precede it, as well as the cancellation and submission of real orders during the interval. This approach enables us to determine the position of the hypothetical order within the order book.

For each hypothetical order, we record information on the state space at both the beginning and the end of the interval. Specifically, at the start of the interval, we capture the volatility, initial queue position, and the total volume available at the first three best bid and best ask prices. At the end of the interval, we note whether the order executes. If the order does not execute, we report the order's new queue position. Further, regardless of execution, we record the volatility and total volume available at the first three best bid and best ask prices at the end of the interval.

With the extracted information, we can identify each order's initial starting state and its state at the end of the interval. This information allows us to estimate the transition matrix and immediate reward matrix using the process outlined in Section 2.

4 Results

In this section, we estimate our model using four public state variables based on the limit order book queue sizes (q^{B0}, q^{B1}, q^{B2} and q^{A0}), each with five possible values (except q^{B2}, which only has three states for tractability reasons).[13] Additionally, we include a public volatility state variable

[11] We assume each order consists of only one share to ensure it does not have a significant economic impact. Additionally, the price and queue position of the hypothetical orders are selected to ensure our observations cover the state space defined by our RL framework.

[12] We assume a hypothetical order executes if, during the 100ms interval, a real order positioned behind it in the queue executes, or if a trade occurs at a price worse than the hypothetical order's price.

[13] Queue size quantiles are formed for each stock at each price level.

with three possible values, where volatility is defined as the difference between the log of the highest traded price and the log of the lowest traded price over the last 100 trades. We also consider two private state variables related to the trader's resting limit order: L and Q, which have 3 and 5 possibles values, respectively.

This state space results in 16,875 different states when the trader has no inventory and is executing a limit order, 1,125 unique market states when the trader's order has executed and they hold an inventory position, and 1 absorbing cancel state. In total, this gives us $m = 16,875$, $n = 1,125$ and $o = 1$, resulting in 18,001 unique states.[14]

In Section 4.2 , we investigate the effect of each market feature on the expected value of a limit order. In Section 4.3, we assess the relative contribution of each market feature to the order's expected value. Finally, in Section 4.4, we evaluate the value of the option to cancel a limit order.

4.1 The expected value of a limit order

The results presented in Table 2, Panel A show that an average limit order submitted within two ticks of the best bid and ask price has an expected value of 0.146 ticks, conditional on the order's optimal management over its life.

Next, we investigate the relation between a limit order's price level and its expected value. The relation between a resting limit order's price and expected value is not immediately clear due to two opposing forces. First, the further a limit order is from the best bid or offer, the more favorable the execution price. However, this price improvement comes at the cost of lower execution probabilities (see Handa and Schwartz (1996)).

Figure 3 presents a boxplot of expected value for all markets states at each of the three price levels defined in our state space (best bid, one tick behind and two ticks behind the best bid). In Figure 3, we observe that the expected value of a limit order is positive, on average. This result is consistent with the empirical findings of Handa and Schwartz (1996), who report that a randomly submitted limit order is profitable and supports the hypothesis that liquidity providers

[14] For further clarity, we demonstrate the full estimation process via a detailed illustrative example in Appendix C.

who accommodate purchases (sales) should be compensated with a higher (lower) price than the fundamental value (see Scholes (1972)).

[Insert Figure 3]

Table 2, Panel B reports the summary statistics for our expected value estimates. The first row reports summary statistics for all market states, whereas rows 2 to 4 report summary statistics for limit orders conditional on their price level. Consistent with Figure 3, when the order is resting at the best bid, its mean expected value is highest at 0.262 ticks. When an optimally managed limit order moves away from the best bid, its expected value drops to 0.105 ticks when it is one tick behind the best bid, and drops further to 0.031 ticks, when it is two ticks behind the best best bid. Similarly, the variance in a limit order's conditional expected value decreases as the order moves away from the best price. The expected value of an order located at the best bid has a standard deviation of 0.134 ticks, but this value drops to only 0.01 ticks when the order is two behind the best bid.

[Insert Table 2]

The observation that the mean expected value and variance of expected value decreases as the order moves further away from the best bid or offer may explain why the majority of order cancellations occur at the best bid or ask (see Fong and Liu (2010)). Intuitively, a trader has little incentive to cancel a limit order resting far from the best price. Such an order likely has a small positive expected value because it carries minimal execution risk and could gain favorable queue priority in the future. However, if the market moves toward the resting limit order, increasing its probability of execution, the expected value could turn negative. At that point, the trader should consider canceling the order.

4.2 Features influencing a limit order's expected value

4.2.1 Queue position

In this section, we investigate the effect of a limit order's queue position on the order's expected value. Some argue that there is an advantage to being at the top of the queue, due to the time priority rule (see Yueshen (2021), Li et al. (2020) and Yao and Ye (2018)). In contrast, some literature suggests that small incoming market orders are more informed (see Brogaard et al. (2014)). Thus, orders at the top on the queue execute against these small informed orders, whereas orders further back in the queue can only execute against larger, less informed orders. To determine the effect of queue position on the expected value of a limit order, we estimate the following regression:

$$Q_s = \beta_1 QueuePos_s + State\ Fixed\ Effects + \epsilon_s, \tag{8}$$

where Q_s is the expected value of a limit order in state s and $QueuePos_s$ is the order's queue position (0 being the top and 1 being the back) in state s. We use fixed effects for all other variables that define our state space to isolate the effect of queue position.

Table 3 presents the mean coefficient for orders resting at three different positions: the best bid (column 1), one level behind the best bid (column 2), and two levels behind the best bid (column 3), across all 20 sample stocks. Additionally, Table 3 reports the number of stocks with significantly positive or negative coefficients at the 5% level, as in Engle and Patton (2004).[15]

Our results strongly indicate that queue priority benefits the liquidity provider. The coefficient for queue position is negative and significant across all 20 sample stocks and all price levels. This implies that the further back a limit order is in the queue, the lower its expected value. The magnitude of these coefficients suggests that queue position has a substantial economic impact on the expected value of a limit order. For instance, an order's expected value at the best bid decreases

[15]Each regression has 5,625 observations, encompassing the following states: five queue position states, five queue size states at the best bid, five queue size states one tick behind the best bid, three queue size states two ticks behind the best bid, five queue size states at the best ask, and three volatility states. This results in a total of $5 \times 5 \times 5 \times 3 \times 5 \times 3 = 5,625$ observations.

by 0.12 ticks when it moves from the top of the queue ($QueuePos_s = 0$) to the back of the queue ($QueuePos_s = 1$). This decrease is economically significant, representing almost half of the average value of a limit order resting at the best bid, which is 0.262 ticks.

We also observe that the magnitude of the mean coefficient decreases as the order moves further from the best bid. Specifically, for orders at the best bid, the mean coefficient is -0.12. For orders one level behind the best bid, the mean coefficient is -0.05, and for orders two levels behind the best bid, the mean coefficient further decreases to -0.01. This pattern indicates that queue priority becomes more critical as the order moves closer to the best price, where execution is most likely.

[Insert Table 3]

Our results highlight the advantages of having orders positioned at the front of the queue, which is consistent with Yueshen (2021), Li et al. (2020) and Yao and Ye (2018). Orders at the front of the queue have higher execution probabilities and lower adverse selection risk compared to those at the back. This lower risk occurs because an order at the back of the queue can only execute against large incoming market orders, which have a significant adverse price impact when all resting limit orders at the current price level are removed. In contrast, orders at the front of the queue can execute against small incoming market orders, which do not exhaust the liquidity at the current price level. The combination of higher execution likelihood and lower adverse selection risk results in a higher expected value for orders at the front of the queue.

Our findings also support the argument by Lo et al. (2002) that the simulated profits from placing a network of buy and sell limit orders, as reported by Handa and Schwartz (1996), may be overstated. This is because their assumption that the orders are placed at the top of the queue does not fully account for the critical importance of queue priority.

4.2.2 Queue size

Existing theoretical literature suggests that queue sizes affect the value of a limit order (see Parlour (1998), Goettler et al. (2005), Goettler et al. (2009)). However, there are few empirical tests. In this

section, we empirically investigate how queue size affects the expected value of a limit order. To investigate the relation between queue sizes and expected value, we estimate the following regression for orders at different price levels:[16]

$$Q_s = \beta_1 q_s^{B_0} + \beta_2 q_s^{B_1} + \beta_3 q_s^{B_2} + \beta_4 q_s^{A_0} + \textit{State Fixed Effects} + \epsilon, \tag{9}$$

where Q_s is the expected value of a limit order in state s, $q_s^{B_i}$ is the size of the bid queue at level i in state s and $q_s^{A_0}$ is the size of the queue on the ask. To isolate the effect of queue sizes, we use fixed effects for all other variables that define our state space.

Table 4 presents the results for orders resting at the three different prices levels defined in our state space (best bid, one tick behind the best bid, two ticks behind the best bid). For each variable, we report the mean coefficient across all 20 sample stocks. To ensure the mean coefficients are not driven by one stock, we also report the number of stocks with statistically positive or negative coefficients.

Overall, our results suggest that the larger the queue size *behind* a resting limit order, the higher the expected value of the order. Conversely, the larger the queue size *in front* of a resting limit order, the lower the expected value of the order. Observing the results for orders resting at the best bid, we find that the mean coefficients for q^{B_0}, q^{B_1}, q^{B_2} are all positive at 0.06, 0.05 and 0.04 respectively, suggesting that an increase in queue lengths at or behind the resting order's price level increases the limit order's expected value. This relation weakens at price levels further away from the best bid. Not only do the average coefficients drop monotonically from 0.06 to 0.04 as we transition from q^{B_0} to q^{B_2}, but we also see the number of stocks with positive and significant coefficients drop from 19 to 18 to 14 as we transition from q^{B_0} to q^{B_1} to q^{B_2}.

In contrast to our findings for orders resting at the best bid, for orders behind the best bid (columns 2 and 3), an increase in queue sizes at price levels ahead of the resting limit order

[16]There is no existing theory suggesting that price levels have a linear affect on limit order value. Thus, we estimate a regression for all orders at each price level individually.

decreases the order's expected value. For example, the coefficient for queue sizes at the best bid (q^{B_0}) is negative and significant for all 20 sample stocks for orders resting one level behind the best bid (column 2). Similarly, for orders resting two levels behind the best bid in column 3, the coefficients for queue lengths at the best bid (q^{B_0}) and one level behind the best bid (q^{B_0}) are negative and significant for all 20 sample stocks.

[Insert Table 4]

Our findings manifest in two ways. First, when the volume in front of the limit order increases, the order's probability of execution worsens. This is because the volume in front of the order exerts order book pressure that can drive the price away from the order. Cao et al. (2009) demonstrate that if the volume on the bid side of the order book is significantly larger (smaller) than the volume available on the ask side of the order book, then the midpoint price is more likely to increase (decrease) in the near future. Thus, if a resting limit order has a large volume ahead of it, that limit order is more likely to be on the thick side of the book. As a result, the price is likely to shift away from the order, resulting in non-execution.

Second, a limit order with more volume in front of the order faces higher adverse selection risk. This is because the volume in front of the order must first execute before the limit order can execute. For example, a limit order behind a large block of volume can only immediately execute when a larger incoming market order enters to first remove the large block of volume. These large market orders cause the largest adverse selection (see Hasbrouck (1991)). In contrast, an order with no volume in front of it can execute against the next incoming market order, regardless of how small it is.

The relation between a limit order's expected value and the volume on the opposite side of the book also depends on the order's price level. In Table 4, the coefficient for the volume on the opposite side of the book (q_{A_0}) is negative and significant for all sample stocks when the order is on the best bid. However, the sign becomes positive and significant for orders behind the best bid. This finding suggests an increase in volume on the opposite side of the order book decreases (increases) the expected value of the limit order if it is at (behind) the best price.

This difference in effect is due to a trade off between adverse selection and execution probability. Cao et al. (2009) document that a large volume on the opposite side of the book creates book pressure that causes shifts in the midpoint towards the limit order, which increases both the likelihood of adverse selection and the probability of execution. When a resting limit order is on the best bid, an increase in ask volume increases the likelihood of a downtick, thereby increasing the expected losses from adverse selection. This negative effect is stronger than the potential gains resulting from a higher probability of order execution. In contrast, when the order is behind the best bid, the expected value from increased execution probability outweighs the expected losses from adverse selection risk. Adverse selection risk remains low for orders behind the best bid as the order can be subsequently canceled if market conditions worsen.

Taken together, our results provide strong support for Parlour (1998); we find that the larger the queue size *behind* a resting limit order, the *higher* the expected value of the order, and the larger the queue size *in front* of a resting limit order, the *lower* the expected value of the order. We also document that the queue size on the opposite side of the book has mixed effects due to a trade off between adverse selection and execution probability. As the queue size on the other side of the book increases, the risk of adverse selection and execution probability both increase. For orders resting at the best price, the losses from adverse selection outweigh the gains from higher execution probability. In contrast, for orders resting behind the best price, the gains from higher execution probability outweigh the loses from adverse selection. Overall, our findings provide support for the predictions of Parlour (1998), Goettler et al. (2005) and Goettler et al. (2009) that strategic traders should consider queue sizes at multiple price levels and demonstrate pervasive features that exist for orders at different price levels.

4.2.3 Volatility

In this section, we explore the impact of volatility on the expected value of a limit order. While existing theoretical models provide some insights into the relation between volatility and the expected value of a limit order, there is no clear consensus due to two opposing forces identified in the literature.

The first force suggests that an increase in volatility decreases the expected value of a limit order. Specifically, Foucault (1999) predicts that when volatility increases, the picking off risk for a limit order increases, and the losses that ensue are larger, decreasing the expected value of the limit order. The second force acts as a response to the first; to compensate for the higher likelihood of adverse selection and the corresponding reduction in the expected value of a limit order, liquidity providers widen the bid ask spread when volatility increases (as discussed in Copeland and Galai (1983) and Foucault (1999)). In a continuous order book where limit orders can be placed at any price level, liquidity providers can adjust the bid ask spread precisely to offset the anticipated losses due to the increased picking-off risk.

However, in practice, price levels are discrete, and liquidity providers may not always be able to set a breakeven bid ask spread that perfectly offsets the increase in volatility, as described in Li et al. (2020). In the first scenario, the breakeven bid ask spread is below one tick, as shown in Figure 4, Panel A. Here, the liquidity provider receives the difference between the one tick mandated bid ask spread and the breakeven bid ask spread as compensation for providing liquidity. The bottom of Figure 4, Panel A illustrates the effect of an increase in volatility. As volatility increases, the breakeven bid ask spread widens but still remains within the one-tick mandated spread, resulting in a decrease in the liquidity provider's compensation. This reduction in compensation lowers the expected value of the limit order. This scenario is particularly pronounced in stocks that are constrained by the minimum tick size.

[Insert Figure 4]

In the second scenario, shown in Figure 4, Panel B, an increase in volatility leads to a breakeven bid ask spread that exceeds the mandated minimum tick size. In response, liquidity providers widen their quoted bid ask spread to a level that is at least as wide as the breakeven spread. As a result, an increase in volatility raises the expected value of a limit order resting at the best bid price. This scenario is most common in stocks that are least constrained by the minimum tick size.

In summary, due to price discretization, volatility can either increase or decrease the expected value of a limit order, depending on whether the stock's trading is constrained by minimum tick

size requirements. Our initial analysis examines the impact of market volatility on the expected value of a limit order across all stocks. We then further explore the effects of volatility on individual stocks, focusing on whether their trading is constrained by minimum tick size requirements.

For our initial investigation using the full sample of stocks, we conduct the following regression analysis:

$$Q_s = \beta_1 Volatility_s + State\ Fixed\ Effects + \epsilon_s, \tag{10}$$

where Q_s is the expected value of a limit order in state s, and $Volatility_s$ is the discretized volatility in state s.[17] To isolate the effect of volatility, we use fixed effects for all other variables that define our state space. Since our primary interest is in the effect of volatility on orders at the best bid or offer, and this effect may vary across price levels, we estimate (10) on the subset of limit orders at the first price level of our defined state space.

Table 5, Column 1 reports an average coefficient of 0.38 across all sample stocks. While this average coefficient suggests that volatility increases the expected value of a limit order, a closer inspection of the individual coefficients for each stock reveals a more complex picture. Specifically, we find positive coefficients for 9 out of the 20 sample stocks, while 11 out of 20 stocks show negative coefficients. Thus, the effect of volatility on the expected value of a limit order at the best bid is not consistent across all stocks.

[Insert Table 5]

Because liquidity providers can widen the bid ask spread to offset the increase in picking off risk, volatility could have mixed effects on the expected value of a limit order (Foucault (1999)). According to Li et al. (2020), the impact of volatility should vary depending on whether a stock's trading is tick constrained. For stocks that are typically tick constrained, we propose that an

[17] Volatility is measured as the highest traded price minus the lowest traded price over the last 100 trades and discretized into terciles.

increase in volatility negatively affects the expected value of a limit order, as the breakeven bid ask spread is narrower than the one tick mandated spread. Conversely, for stocks that are less tick constrained, we predict that an increase in volatility will raise the expected value of a limit order. In these cases, liquidity providers can widen their quoted bid ask spread to compensate for the increased picking off risk during volatile periods.

To test whether volatility has different effects on the expected value of a limit order for tick constrained and unconstrained stocks, we split the sample stocks into two subsamples. Table 5, column 2 reports results for the quartile of the most tick constrained stocks, while column 3 reports results for the quartile of the least tick constrained stocks. Consistent with our hypothesis, we find that volatility decreases the expected value of a limit order for tick constrained stocks. Specifically, *Volatility* is negative and significant for all stocks within the tick constrained subset. Conversely, for all stocks not constrained by the tick size in Column 3, we find that volatility increases the expected value of a limit order.

Overall, our results support the findings of Foucault (1999) and the tick size channel proposed by Li et al. (2020). For stocks that are most tick constrained, the breakeven spread lies within the one-tick mandated spread. In these cases, liquidity providers are unable to widen the quoted bid ask spread, leading to a decrease in the expected value of a limit order as volatility increases. Conversely, for stocks that are least tick constrained, an increase in volatility can push the breakeven spread beyond the one-tick minimum. In response, liquidity providers widen the quoted bid ask spread to compensate for the additional picking-off risk. As a result, volatility increases the expected value of a limit order for these less constrained stocks.

4.3 How important are the variables?

The results so far indicate that price levels, queue sizes, queue position, and volatility all influence the expected value of a limit order. In this section, we assess the importance of these variables using a technique from the machine learning literature known as Mean Decreased Accuracy (MDA). This method has been applied to the finance literature by Easley et al. (2021) and Kwan et al. (2024). In our context, MDA measures the decrease in accuracy of the forecasted expected value of a limit

order when one of the variables defining our states is intentionally measured with error.

Estimating the MDA requires two parameters. The first parameter is the true expected value of a resting limit order, $Q(s, NA)$, which we estimate using the RL model described in Section 2. Specifically, we have 18,001 $Q(s, NA)$ estimates corresponding to 16,875 different states when the trader has no inventory and is executing a limit order, 1,125 unique market states when the trader's order has executed and they hold an inventory position, and 1 absorbing cancel state.

The second parameter is the randomized expected value of a resting limit order, $Q(s_R^k, NA)$, which we estimate by randomizing one of the seven variables that define the state space while keeping all other variables constant.[18] $Q(s_R^k, NA)$ represents the expected value associated with the randomly altered state, s_R^k, created by randomizing variable k. This randomization helps isolate the effect of variable k on the $Q(s, NA)$ estimate.

Using these two parameters, we estimate the MDA for variable k as follows:

$$MDA^k = \sum_{s=1}^{S} \left(\frac{|(Q(s, NA) - Q(s_R^k, NA))|}{Q(s, NA)} \right) / S. \tag{11}$$

The MDA measures the error in expected value estimates that occurs when a variable is measured with error. Therefore, the larger a variable's MDA, the more important that variable is in determining the expected value of a limit order. For each variable, k, we estimate the MDA and repeat this process 100 times.

Table 6 presents the mean and standard deviation of the MDA for each variable. Our findings indicate that the most influential factor affecting the expected value of a limit order is the price level at which the order rests. The next most important variable is the queue size on the same side of the order book as the limit order. However, the importance of queue size decreases as the queues move further from the best bid. Specifically, the queue size at the best bid (MDA = 1.22) is the most important, followed by the queue size one tick behind the best bid (MDA = 1.17), and then the queue size two ticks behind the best bid (MDA = 0.68).

[18] The seven variables include the price level, queue position, queue sizes at different price levels (best bid, one tick behind the best bid, two ticks behind the best bid, best ask), and volatility.

After considering queue sizes on the same side of the book as the order, the next most important variable is the queue size on the opposite side of the order book (MDA = 0.68), followed by volatility (MDA = 0.56), and lastly, queue position (MDA = 0.2).

[Insert Table 6]

4.4 The option to cancel

Despite the prevalence of order cancellations, the option to cancel has received little attention in the literature. In this section, we investigate the value of the option to cancel and identify the market conditions under which this option is most valuable. For our analysis, we estimate a constrained version of our RL model, which restricts the trader to only one action, NA, meaning the trader is unable to cancel their order.[19] Thus, the estimated Q values from this restricted model represent the expected value of a limit order that is not optimally managed. To determine the value of the option to cancel, we calculate the difference between the Q value of the unrestricted model, which includes the option to cancel, and the Q value of the restricted model, which lacks this option.

Table 7 reports the summary statistics on the value of the option to cancel. The first row reports summary statistics for limit orders across all market states, while rows 2 to 4 focus on limit orders conditional on their price level. On average, the value of the option to cancel a limit order on the best bid is 0.049 ticks. This means that, on average, the option to cancel a limit order from the best level is worth approximately 19% of the total value of an optimally managed limit order.[20] This finding suggests that the endogenous option to cancel a limit order contributes an economically meaningful amount towards the overall expected value of an optimally managed limit order.

[Insert Table 7]

Table 7 also suggests that the value of the option to cancel varies significantly depending on

[19] Recall in the full model, the trader can cancel the order, C, or take no action, NA.
[20] Table 2 reports a mean value of an optimally managed limit order at the best bid of 0.258 ticks.

prevailing market conditions. Regardless of the price level at which orders are resting, the data shows a substantial disparity between the means and medians, indicating a pronounced right skew. Over the full sample, the mean value of the option to cancel is 0.024 ticks, whereas the median is only 0.008 ticks.

Theoretical considerations suggest that the option to cancel is most valuable when the limit order is most likely to be adversely selected. However, it is difficult for a trader to know when adverse selection risk is high *ex-ante*. To proxy for an *ex-ante* measure of adverse selection risk, we draw on Cao et al. (2009), who show that order book pressure can be used to predict short term price movements, and Goldstein et al. (2023) who show that limit orders with high ex-ante adverse selection risk are more likely to rest on the thin side of the book. A higher volume of buy orders (bids) in the market creates buying pressure in the stock, and the price is more likely to rise. As a result, the adverse selection risk of a resting buy limit order decreases and the value of having the option to cancel the buy order also decreases. On the other hand, when volume on the ask side increases, the selling pressure is more likely to result in a price fall. The adverse selection risk of a resting buy order rises and the option to cancel the order becomes more valuable.

Drawing from these studies, we use order book pressure as a proxy for *ex-ante* adverse selection and estimate the following regression for the subset of limit orders at the front of the queue at each price level:

$$value\ of\ option\ to\ cancel = \beta_0 + \beta_1 q^{B_0} + \beta_2 q^{B_1} + \beta_3 q^{B_2} + \beta_4 q^{A_0} + \epsilon. \tag{12}$$

If the value of the option to cancel increases when adverse selection increases, we expect an increase in volume on the *same* side of the order (i.e., $q^{B_0}, q^{B_1}, q^{B_2}$) to reduce the value of the option to cancel. Similarly, we expect an increase in volume on the *opposite* side of the order (i.e., q^{A_0}) to increase the value of the option to cancel. Table 8 confirms this hypothesis and demonstrates the option to cancel is most valuable when book pressure is going against the order (i.e., adverse selection is high). Specifically, for all regressions, Table 8 reports a negative relation between the

queue size at *any* price level on the bid side and the value of the option to cancel. Similarly, for all regressions, the value of the option to cancel has a positive relation with the queue size on the opposing ask (q^{A_0}). In other words, when there is greater trading volume on the same side as the resting limit order, the option to cancel holds a lower value. Conversely, if there is more volume present on the opposite side of the limit order, the option to cancel carries a higher value.

[Insert Table 8]

5 Extensions

In this section, we describe possible extensions to the basic RL framework. Such enhancements include the exploration of strategic decisions related to selecting between limit orders and market orders, incorporating considerations such as the liquidity provider's risk tolerance or existing inventory levels, and determining order sizes.

5.1 Market vs. limit orders

A substantial body of theoretical work explores the dynamics behind traders' preferences for market versus limit orders. For instance, Kaniel and Liu (2006) reveals that, contrary to previous assumptions, informed traders are more inclined to use limit orders rather than market orders. In a fully dynamic framework, Bhattacharya and Saar (2022) show that the liquidity of the market significantly influences informed traders' decision to place limit orders in less liquid markets and marketable orders in more liquid markets. From an empirical standpoint, Ranaldo (2004) demonstrates that a trader's decision on order aggressiveness is dependent on the market's depth, spread, and volatility. The flexibility of our proposed RL framework allows researchers to study a trader's order choice while also taking into account factors such as their existing inventory levels, risk aversion, private information, as well as the prevailing market conditions.

The RL specification specified in section 2 can be modified to investigate the decision between limit and market order submissions. Specifically, we can extend the action space to include an

additional action that simultaneously cancels the resting limit order, and executes a market order by crossing the spread, M. This augmentation results in three possible actions 1) to leave the existing limit order (NA), 2) to cancel the existing limit order (C) or 3) to cancel the existing limit order and immediately execute a market order (M). As before, we solve Equation (2) via the Q-learning rule for all $Q(s, a)$ where $a \in \{C, NA, M\}$. This modification allows us to estimate the expected profit for each action, enabling us to compare the potential outcomes of leaving a resting limit order or sending a market order in a given state.

Moreover, we can modify the RL framework to examine the order submission strategies of an impatient trader who penalizes wait time until execution. Specifically, we can adjust the discount factor γ in Equation 2. Recall that γ is a discount factor between 0 and 1. Values close to 1 apply minimal discounting to future payoffs or rewards, while values close to 0 place little weight on future payoffs (i.e., future payoffs are heavily discounted). Therefore, using a value close to zero places little emphasis on rewards from limit order executions that occur in the distant future. This setup effectively represents an impatient trader who underweights the potential payoffs from long-lived limit orders.

5.2 Inventory and risk aversion

Inventory is also a crucial factor in managing limit orders. For example, Garriott et al. (2024) demonstrate that inventory levels and adverse selection constraints similarly affect limit order sizes. Consequently, a market maker holding a long position of 1,000 shares will place a higher value on a limit sell order compared to a market maker with a short position of 1,000 shares, assuming both market makers aim to minimize their inventory positions due to the associated risk.

We can adapt the existing RL framework to model the preferences of a risk-averse market maker with inventory considerations through two adjustments. First, we include the market maker's inventory position in the state space. Second, we modify the reward function to account for the profits related to the varying inventory position and any associated risk aversion. Specifically, equation 6 can be modified to:

$$R(\langle s, a\rangle, s') = \underbrace{(mid_{s'} - mid_s) \times Inv_s}_{\substack{\text{Profit from carrying} \\ \text{inventory}}} + \underbrace{(mid_{s'} - execPrice_s) \times Exec_{s,s'}}_{\text{Profit from execution}} - \underbrace{\lambda(|Inv_s|)}_{\text{Risk aversion}}, \tag{13}$$

where Inv_s is the market maker's inventory position in state s, $execPrice_s$ equals the price of the limit order and $Exec_{s,s'}$ is the volume of the limit order that executes during the transition from state s to s'. $\lambda(|Inv_s|)$ is a risk aversion penalty applied to the market maker's inventory position. This penalty can be defined using various risk aversion utility functions due to the flexibility of the RL framework.

5.3 Order size

In our RL framework, a risk-averse market maker (liquidity provider) trades a hypothetical order of unit size. This assumption ensures that the order is not economically meaningful and thus does not cause any permanent price impact. However, in practice, many limit orders are larger and do have an associated price impact (see Brogaard et al. (2019) and Kwan et al. (2024)). Here, we extend the RL framework by modifying the private variables in the state space to include the size of the limit order. This modification allows us to capture any permanent price impact that arises from submitting a limit order that is large enough to be economically meaningful.

While the estimation of expected profits via Q-learning remains unchanged, two modifications are required in the process. First, instead of using hypothetical limit orders of unit size, the transition matrix should incorporate appropriately sized orders that can alter the queue sizes in the state space. Second, the reward matrix must account for the volume executed, considering both partial and complete executions.

5.4 Private information

Many theoretical models assume that a trader has access to private information. Our reinforcement learning (RL) framework can be adapted to capture this feature. Specifically, we can introduce a private valuation variable to the state space, p, which represents the trader's private information. If we assume that the trader's private information is valuable, this modification will lead to limit buy orders having higher expected values when the private information indicates a potential price increase, and lower expected values when it suggests a potential price decrease.

6 Conclusion

In modern markets, where limit order submissions and cancellations constitute an overwhelming majority of trading activity, understanding the optimal management of limit orders is crucial. Despite its importance, our understanding of the dynamics of the limit order book and order management strategies remains limited due to the complexity and high dimensionality of the problem (see Parlour and Seppi (2008)).

To address this issue, we propose a recursive sequential framework for limit order management which allows us to empirically uncover the most important features contribution to the value of a limit order. In our framework, the expected value of a limit order is determined by current market conditions and future market condition expectations. The liquidity provider exercises the option to cancel a limit order if its expected value becomes negative.

Our findings reveal that the average expected value of a limit order resting at the best bid is approximately one quarter of a tick. However, this value is influenced by various market factors. Specifically, we demonstrate that queue size, order position, volatility, and order price significantly impact the expected value of a limit order. Using Mean Decreased Accuracy (MDA) to rank the importance of these variables, we find that price level is the most critical factor, followed by queue sizes, volatility, and queue position.

Finally, we show that the endogenous option to cancel is economically meaningful: On average,

this option to cancel represents 19% of a limit order's total expected value. During periods of high adverse selection risk, this option becomes even more valuable.

References

Ait-Sahalia, Y. and Saglam, M. (2023). High Frequency Market Making: The Role of Speed. *Journal of Econometrics*, Forthcoming.

Bhattacharya, A. and Saar, G. (2022). Limit Order Markets under Asymmetric Information. Working paper, Available at SSRN: https://ssrn.com/abstract=3688473.

Brogaard, J., Hendershott, T., and Riordan, R. (2014). High-frequency trading and price discovery. *Review of Financial Studies*, 27(8):2267–2306.

Brogaard, J., Hendershott, T., and Riordan, R. (2019). Price discovery without trading: Evidence from limit orders. *The Journal of Finance*, 74(4):1621–1658.

Cao, C., Hansch, O., and Wang, X. (2009). The information content of an open limit-order book. *Journal of Futures Markets*, 29(1):16–41.

Chinco, A., Clark-Joseph, A., and Ye, M. (2019). Sparse signals in the cross-section of returns. *The Journal of Finance*, 74(1):449–492.

Colliard, J.-E., Foucault, T., and Lovo, S. (2022). Algorithmic Pricing and Liquidity in Securities Markets. Working paper, Available at SSRN: https://ssrn.com/abstract=4252858.

Copeland, T. E. and Galai, D. (1983). Information effects on the bid-ask spread. *The Journal of Finance*, 38(5):1457–1469.

Dahlström, P., Hagströmer, B., and Nordén, L. L. (2023). The determinants of limit order cancellations. *Financial Review*, Forthcoming.

Dou, W. W., Goldstein, I., and Ji, Y. (2024). AI-Powered Trading, Algorithmic Collusion, and Price Efficiency. The Wharton School Research Paper.

Easley, D., López de Prado, M., O'Hara, M., and Zhang, Z. (2021). Microstructure in the Machine Age. *The Review of Financial Studies*, 34(7):3316–3363.

Ellul, A., Holden, C. W., Jain, P., and Jennings, R. (2007). Order dynamics: Recent evidence from the nyse. *Journal of Empirical Finance*, 14(5):636–661.

Engle, R. F. and Patton, A. J. (2004). Impacts of trades in an error-correction model of quote prices. *Journal of Financial Markets*, 7(1):1–25.

Fong, K. and Liu, W.-M. (2010). Limit order revisions. *Journal of Banking and Finance*, 34:1873–1885.

Foucault, T. (1999). Order flow composition and trading costs in a dynamic limit order market. *Journal of Financial Markets*, 2(2):99 – 134.

Foucault, T., Kadan, O., and Kandel, E. (2005). Limit order book as a market for liquidity. *The Review of Financial Studies*, 18(4):1171–1217.

Garriott, C., van Kervel, V., and Zoican, M. (2024). Queuing in limit-order markets.

Glosten, L. R. (1994). Is the electronic open limit order book inevitable? *The Journal of Finance*, 49(4):1127–1161.

Goettler, R. L., Parlour, C. A., and Rajan, U. (2005). Equilibrium in a dynamic limit order market. *The Journal of Finance*, 60(5):2149–2192.

Goettler, R. L., Parlour, C. A., and Rajan, U. (2009). Informed traders and limit order markets. *The Journal of Financial Economics*, 93:67–87.

Goldstein, M., Kwan, A., and Philip, R. (2023). High Frequency Trading Strategies. *Management Science*, 69(8):4413–4434.

Griffiths, M., Smith, B., Turnbull, D., and White, R. (2000). The costs and determinants of order aggressiveness. *Journal of Financial Economics*, 56:65–88.

Handa, P. and Schwartz, R. (1996). Limit order trading. *The Journal of Finance*, 51(5):1835–1861.

Hasbrouck, J. (1991). Measuring the information content of stock trades. *The Journal of Finance*, 46(1):179–207.

Kaniel, R. and Liu, H. (2006). So what orders do informed traders use? *The Journal of Business*, 79(4):1867–1913.

Kwan, A., Philip, R., and Shkilko, A. (2024). The conduits of price discovery: A machine learning approach.

Li, S., Wang, X., and Ye, M. (2020). Who provides liquidity and when? *Journal of Financial Economics, Fourthcoming*.

Lo, A. W., MacKinlay, A., and Zhang, J. (2002). Econometric models of limit-order executions. *Journal of Financial Economics*, 65(1):31 – 71.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, 116(2):257 – 270.

Parlour, C. and Seppi, D. (2008). Limit order markets: A survey. *Handbook of Financial Intermediation and Banking*, 5:63–95.

Parlour, C. A. (1998). Price dynamics in limit order markets. *The Review of Financial Studies*, 11(4):789–816.

Ranaldo, A. (2004). Order aggressiveness in limit order book markets. *Journal of Financial Markets*, 7(1):53–74.

Ricco, R., Rindi, B., and Seppi, D. (2020). Information, Liquidity, and Dynamic Limit Order Markets. Working paper, Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3032074.

Rosu, I. (2009). A dynamic model of the limit order book. *The Review of Financial Studies*, 22(11):4601–4641.

Rosu, I. (2020). Liquidity and information in limit order markets. *Journal of Financial and Quantitative Analysis*, page 1–48.

Sandås, P. (2015). Adverse Selection and Competitive Market Making: Empirical Evidence from a Limit Order Market. *The Review of Financial Studies*, 14(3):705–734.

Scholes, M. S. (1972). The market for securities: Substitution versus price pressure and the effects of information on share prices. *The Journal of Business*, 45(2):179–211.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3):279–292.

Yao, C. and Ye, M. (2018). Why Trading Speed Matters: A Tale of Queue Rationing under Price Controls. *The Review of Financial Studies*, 31(6):2157–2183.

Yueshen, B. (2021). Queuing uncertainty of limit orders. Working paper, Available at SSRN: http://dx.doi.org/10.2139/ssrn.2336122.

Figure 1. Limit order book evolution

Figure 1 depicts the possible evolution of the limit order book from t_0 to two possible future states at t_1 (A and B). The white rectangles represent the bid volume and the grey rectangles represent the ask volume. Prices are shown on the x-axis, with the best bid at 13 and the best offer at 14 at t_0. The trader's limit order is in black and starts at the back of the queue at t_0 at price 12.

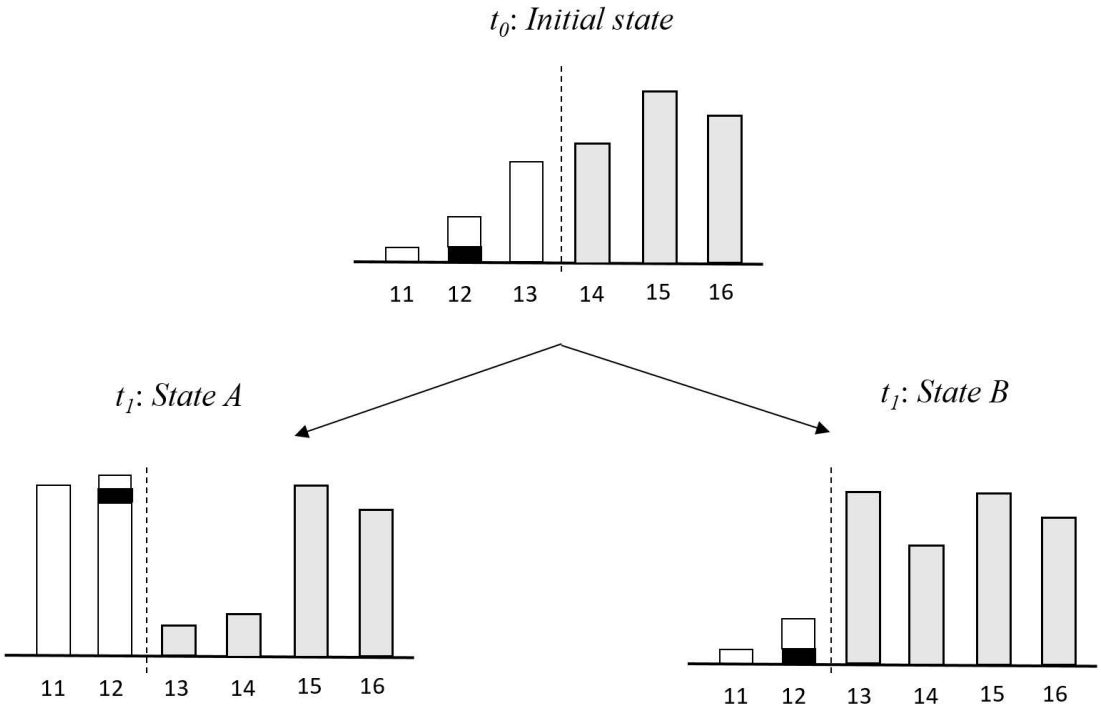


Figure 2. Traders sequential decision making process

Figure 2 depicts the time line of the liquidity provider's decision making process when monitoring their limit order. At the end of each interval, the liquidity provider observes current market conditions and decides to leave or cancel their order. This process repeats until the order is either executed or canceled.

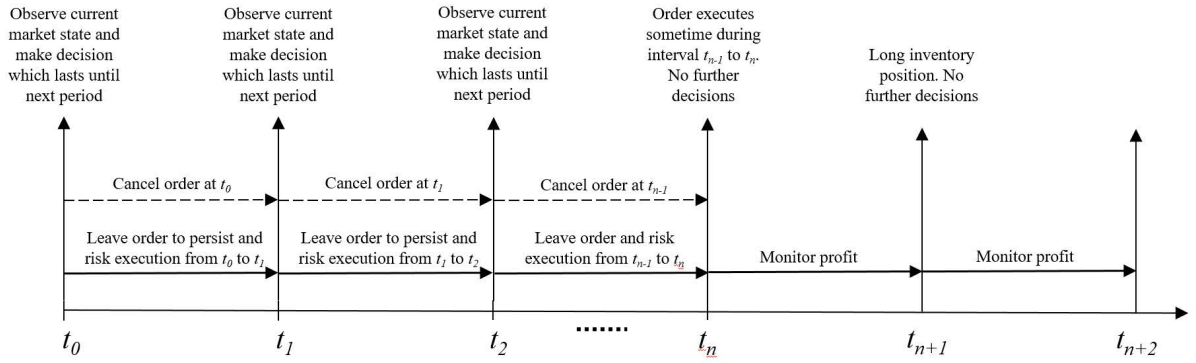


Figure 3. Boxplot of the expected value of a limit order

This figure plots a boxplot of the expected value of a limit order, estimated via our RL model. The figure contains the estimates from all 20 sample stocks. The figure depicts a boxplot for three subsamples conditional on the price level the limit order is resting at.

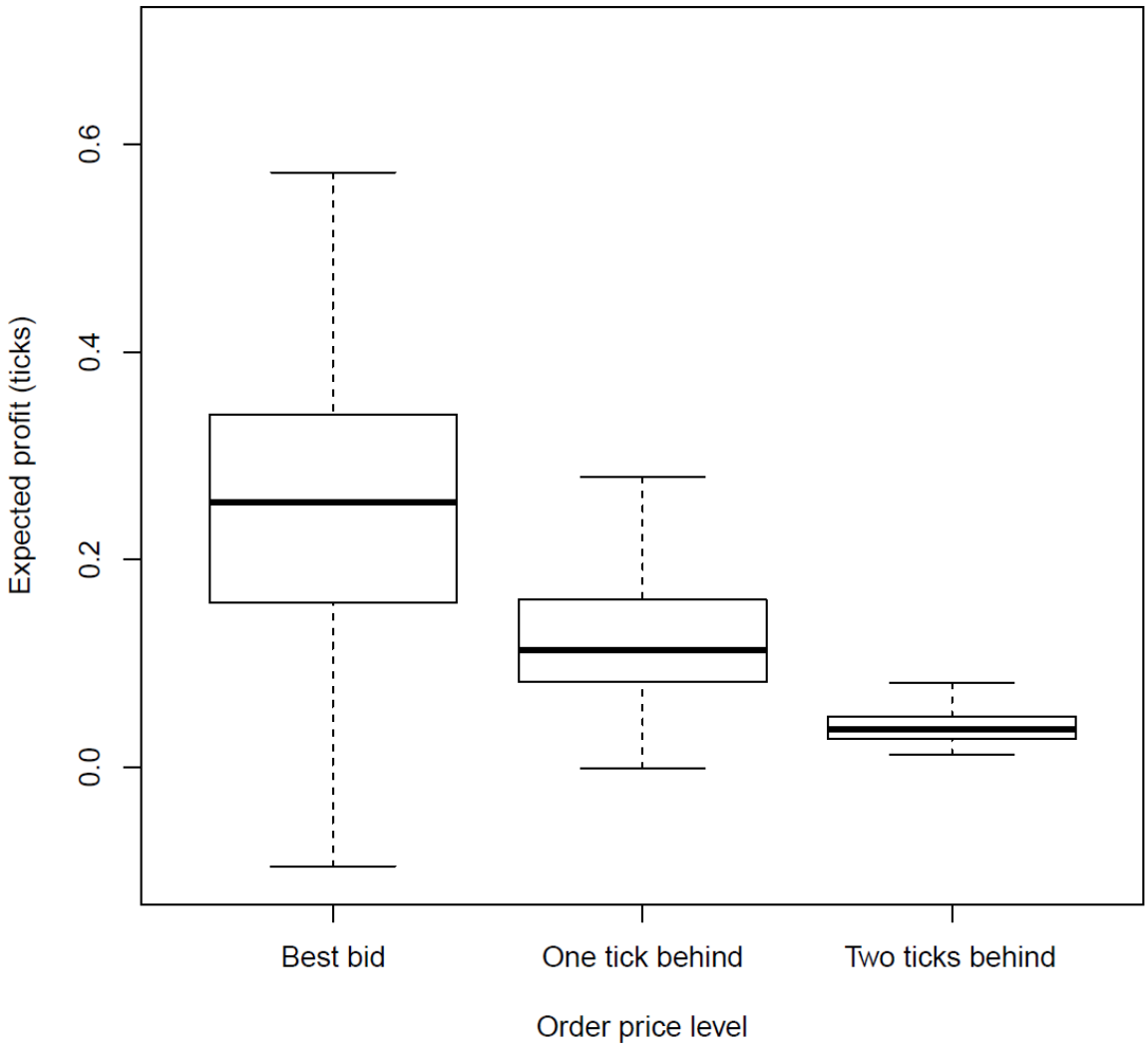


Figure 4. Volatility, bid ask spreads and the expected value of a limit order

This figure shows the predicted effects of volatility on the quoted bid ask spread, breakeven bid ask spread, and the expected value of a limit order. The top (bottom) figure in the panels depicts a low (high) volatility environment. Panel A shows a constrained stock in which the breakeven bid ask spread is always less than the one tick-mandated spread, even when volatility is high. In Panel B, when volatility is high, the breakeven bid ask spread increases beyond the quoted bid ask spread and the liquidity provider widens the quoted spread. The expected compensation to the liquidity provider (i.e., the difference between the quoted bid ask spread and the breakeven bid ask spread) is depicted in green.

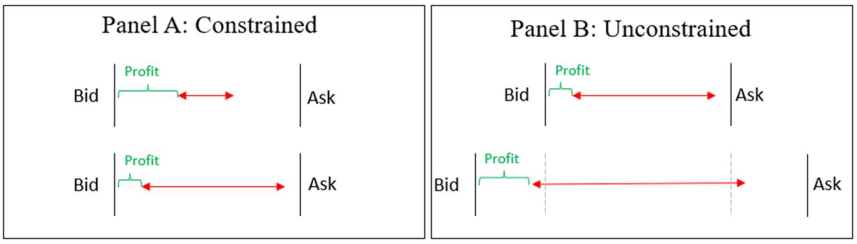


Table 1
Summary statistics

This table reports summary statistics for our sample stocks. Our sample period covers July 3, 2017 to Septemeber 29, 2017 for 20 actively traded stocks on the ASX. We report the average bid ask spread in cents (*Spread*), the average trade price in AUD (*Price*), and the average number of daily trades, order deletions and order submissions labelled *No. trades*, *No. deletions* and *No. submissions*, respectively.

	Spread	Price	No. trades	No. deletions	No. submissions
AMC	1.03	15.72	6970	9247	21142
AMP	1.01	5.11	3026	4279	9318
ANZ	1.08	29.51	11326	68791	88536
BHP	1.06	25.79	14268	20304	44994
BXB	1.04	9.40	5484	6686	16001
CBA	1.61	79.39	21498	33810	71510
CSL	2.59	129.87	16198	42372	70900
IAG	1.01	6.54	3530	5273	11142
MQG	1.97	87.20	13999	32589	57422
NAB	1.06	30.40	11714	69080	89394
NCM	1.12	21.34	10735	17888	36675
ORG	1.02	7.29	4637	6220	14191
QBE	1.08	11.14	7779	9758	22926
RIO	1.70	65.61	15955	30138	57912
STO	1.01	3.51	3255	4668	10058
SUN	1.02	13.61	7920	10994	24647
TLS	1.00	3.94	3999	4754	11186
WBC	1.08	31.62	13469	38652	62169
WOW	1.05	26.04	9208	16856	32784
WPL	1.08	29.27	11203	22482	41672

Table 2
Summary statistics

This table reports the summary statistics on the expected value of an optimally managed limit order. The first row reports summary statistics for orders placed at all price levels, whereas rows 2 to 4 report summary statistics for limit orders conditional on their price level.

Order Location	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. dev.
All prices	-0.098	0.031	0.123	0.146	0.252	0.569	0.098
Best bid	-0.098	0.153	0.258	0.262	0.339	0.569	0.134
One tick behind best bid	0.002	0.071	0.098	0.105	0.155	0.276	0.067
Two ticks behind best bid	0.011	0.023	0.029	0.031	0.044	0.074	0.010

Table 3
Queue position and the expected value of a limit order

This table reports estimation results for the following OLS regression:

$$Q_s = \beta_1 QueuePos_s + State\ Fixed\ Effects + \epsilon_s,$$

where Q_s is the expected value of a limit order estimated via our RL model. The independent variable is $QueuePos$, with fixed effects controlling for all other variables. $QueuePos$ takes the value of 0 if the order is at the front of the queue and 1 if it is at the back of the queue. Columns 1, 2 and 3 present the regression results for subsamples in which the order rests at the best bid, one tick behind the best bid and two ticks behind the best bid, respectively. We report the mean coefficient across all sample stocks (*Mean*), along with the number of significantly positive (*No. +*) and negative (*No. -*) coefficients at the 5% significance level, out of the full sample of 20 stocks. We also report the mean R-squared value across the 20 regressions, along with the number of observations used in each regression.

	Best bid	1 behind best bid	2 behind best bid
Mean	-0.12	-0.05	-0.01
No. +	0	0	0
No. -	20	20	20
Mean R^2	0.89	0.88	0.89
No. obs. per stock	5,625	5,625	5,625

Table 4
Queue size and the expected value of a limit order

This table reports estimation results for the following OLS regression:

$$Q_s = \beta_1 q_s^{B_0} + \beta_2 q_s^{B_1} + \beta_3 q_s^{B_2} + \beta_4 q_s^{A_0} + \textit{State Fixed Effects} + \epsilon,$$

where Q_s is the expected value of a limit order estimated via our RL model, q^{B_i} is the queue size on the best bid at price level i and q^{A_0} is the queue size on the best ask. q^{B_0}, q^{B_1}, q^{A_0} take values from 0 to 4 to depict the queue size quintile, extremely short, short, normal, long, and extremely long, respectively. q^{B_2} takes values from 0 to 2 to represent the three queue size terciles (short, normal long) at the price two ticks below the best bid. Columns 1, 2 and 3 present the regression results for subsamples in which the order rests at the best bid, one level behind the best bid and two levels behind the best bid, respectively. We report the mean coefficient across all sample stocks (*Mean*), along with the number of significantly positive (*No. +*) and negative (*No. -*) coefficients at the 5% significance level, out of the full sample of 20 stocks. We also report the mean R-squared value across the 20 regressions, along with the number of observations used in each regression.

		Best bid	1 behind best bid	2 behind best bid
q^{B_0}	Mean	0.06	-0.05	-0.02
	No. +	19	0	0
	No. -	0	20	20
q^{B_1}	Mean	0.05	0.01	-0.01
	No. +	18	14	0
	No. -	2	5	20
q^{B_2}	Mean	0.04	0.03	0.00
	No. +	14	16	12
	No. -	3	4	6
q^{A_0}	Mean	-0.04	0.01	0.01
	No. +	0	19	20
	No. -	20	0	0
Mean R^2		0.88	0.89	0.83
No. obs. per stock		5,625	5,625	5,625

Table 5
Volatility and the expected value of a limit order

This table reports estimation results for the following OLS regression:

$$Q_s = \beta_1 Volatility_s + \textit{State Fixed Effects} + \epsilon_s,$$

where Q_s is the expected value of a limit order estimated via our RL model for orders at the best bid. The independent variable is $Volatility$, with fixed effects controlling for all other variables. Volatility is calculated as the log of the highest traded price minus the log of the lowest traded price over the last 100 trades. $Volatility$ takes the value of 0, 1, or 2 for low, medium, and high volatility states, respectively. Column 1 presents the regression results for all stocks. Column 2 (3) presents results on the subsample of stocks that are most (least) tick constrained. We report the mean coefficient across all sample stocks (*Mean*), along with the number of significantly positive (*No. +*) and negative (*No. -*) coefficients at the 5% significance level, out of the full sample of 20 stocks. We also report the mean R-squared value across the 20 regressions, the number of stocks in each sample, along with the number of observations used in each regression.

	All stocks	Constrained	Unconstrained
Mean	0.38	-2.39	5.82
No. +	9	0	5
No. -	11	5	0
Mean R^2	0.86	0.86	0.87
No. stocks	20	5	5
No. obs. per stock	5,625	5,625	5,625

Table 6
Relative importance of variables

For each variable, k, that partially defines the market state (i.e., *Price level*, *Queue position*, queue size at best bid (*Bid size 1*), queue size one level behind best bid (*Bid size 2*), queue size two levels behind best bid (*Bid size 3*), *Ask size*, *Volatility*), this table reports the Mean Decreased Accuracy (MDA) estimated as follows:

$$MDA^k = \sum_{s=1}^{S} \left(\frac{|(Q(s, NA) - Q(s_R^k, NA))|}{Q(s, NA)} \right) / S,$$

where $Q(s, NA)$ is the expected value of a limit order while in state s and taking action NA, and $Q(s_R^k, NA)$ is the estimate associated with state s_R when variable k is randomized. For each variable k, we repeat this process 100 times and report the mean and standard deviation of the MDA.

	Price level	Queue position	Bid size 1	Bid size 2	Bid size 3	Ask size	Volatility
Mean	2.54	0.20	1.22	1.17	0.68	0.68	0.56
St. dev.	1.34	0.07	0.34	0.71	0.48	0.19	0.33

Table 7
Summary statistics for the value of the option to cancel

Table 7 reports the summary statistics on the expected value of the option to cancel a limit order. The first row reports summary statistics for orders placed at all price levels, whereas rows 2 to 4 report summary statistics for limit orders conditional on their price level.

Order Location	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St Dev.
All prices	0.000	0.003	0.008	0.024	0.019	0.483	0.052
Best bid	0.000	0.004	0.012	0.049	0.051	0.483	0.081
One tick behind best bid	0.001	0.005	0.010	0.017	0.020	0.161	0.020
Two ticks behind best bid	0.000	0.002	0.004	0.007	0.009	0.059	0.007

Table 8
The value of the option to cancel

This table reports estimation results for the following OLS regression:

$$value\ of\ option\ to\ cancel = \beta_0 + \beta_1 q^{B_0} + \beta_2 q^{B_1} + \beta_3 q^{B_2} + \beta_4 q^{A_0} + \epsilon,$$

where the dependent variable is the option value to cancel a limit order estimated via our RL model, q^{B_i} is the queue size on the best bid at price level i and q^{A_0} is the queue size on the best ask. q^{B_0}, q^{B_1}, q^{A_0} take values from 0 to 4 to depict the queue sizes, extremely short, short, normal, long, and extremely long, respectively. q^{B_2} takes values from 0 to 2 to represent the three queue size terciles (short, normal long) at the price two ticks below the best bid. Columns 1, 2 and 3 present the regression results for subsamples in which the order rests at the best bid, one level behind the best bid and two levels behind the best bid, respectively. We report the mean coefficient across all sample stocks (*Mean*), along with the number of significantly positive (*No. +*) and negative (*No. -*) coefficients at the 5% significance level, out of the full sample of 20 stocks.

		Best bid	1 behind best bid	2 behind best bid
q^{B_0}	Mean	-0.034	-0.015	-0.019
	No. +	0	1	0
	No. -	20	19	20
q^{B_1}	Mean	-0.044	-0.032	-0.021
	No. +	1	0	2
	No. -	19	20	18
q^{B_2}	Mean	-0.034	-0.031	-0.028
	No. +	1	0	0
	No. -	19	20	20
q^{A_0}	Mean	0.014	0.025	0.035
	No. +	20	20	20
	No. -	0	0	0
	Mean R^2	0.16	0.24	0.34
	No. obs per stock	625	625	625

7 Appendix

A Transition Matrix

Transition matrix for action NA

Figure A.1 illustrates the section of the transition matrix, T, when the action is NA (i.e., leave the limit order), which is a $S \times S$ matrix that requires empirical estimation. The states $s_1(0), \ldots, s_n(0)$ reflect the n possible states when the trader has no inventory and is working an order. The states $s_1(1), \ldots, s_m(1)$ reflect the m possible states the market can exist, when the trader has a long position and is no longer working a limit order. $s^C(0)$ reflects the absorbing state once the trader cancels their order.

Figure A.1. Transtion matrix for NA

Figure A.1 depicts the $S \times S$ transition matrix for the experience tuples in which the action is to leave the resting limit order, or do nothing, NA. States $s_i(0)$ represent states when the trader is working their limit order, while states $s_j(1)$ represent states when the trader's order has been executed. The state $s^C(0)$ represents the absorbing order cancellation state.

Current state with action NA	Future State: $s_1(0)$	$s_2(0)$	...	$s_n(0)$	$s_1(1)$	...	$s_m(1)$	$s^C(0)$
$s_1(0)$	$p_{1,1}$	$\cdots$		$p_{1,n}$	$p_{1,n+1}$	$\cdots$	$p_{1,n+m}$	$p_{1,n+m+1}$
$s_2(0)$	$\vdots$	$\ddots$			$\vdots$	$\ddots$		$\vdots$
	Unexecuted				Executed			
$s_{n-1}(0)$	$p_{n-1,1}$				$p_{n-1,n+1}$			
$s_n(0)$	$p_{n,1}$	$\cdots$		$p_{n,n}$	$p_{n,n+1}$		$p_{n,n+m}$	$p_{n,n+m+1}$
$s_1(1)$	0	$\cdots$	0	0	$p_{n+1,n+1}$	$\cdots$	$p_{n+1,n+m}$	0
$\vdots$	$\vdots$	Prohibited		0	$\vdots$	Long	$\vdots$	$\vdots$
$s_m(1)$	0	0	0	0	$p_{n+m,n+1}$	$\cdots$	$p_{n+m,n+m}$	0
$s^C(0)$	0	$\cdots$	0	0	0	$\cdots$	0	1

The top left quadrant of the transition matrix, labeled "Unexecuted", contains the transition probabilities for a limit order that does not execute during the transition from one state to the next. These transition probabilities reflect the changes in market conditions and the movement of the limit order within the order book. For example, they capture the likelihood of the limit order advancing in the queue or how other market participants might respond to current market conditions. We estimate these probabilities empirically using equation (5).

The block of the transition matrix titled "Executed" contains the probability of limit order execution during the state transition. In our setup, once an order is executed, the trader has no remaining limit orders. Consequently, the trader must transition to one of m positive inventory states, $s_j(1)$, where j represents different possible states based on the public information reflected in the order book variables. Again, we estimate these probabilities empirically via (5).

After execution, the trader remains in one of the m positive inventory states and cannot submit another order. To ensure the trader does not hold another limit order while being long and remains in a positive inventory state, we define the "Prohibited" block in Figure A.1 to contain only zeros. The block labeled "Long" captures the transition probabilities for a trader who is long in one market state and transitions to another market state while remaining long; these probabilities are also estimated empirically via equation (5).

The final column of the matrix reports the probability that the trader transitions to the absorbing state by canceling their order. The absorbing nature of the state is represented by the transition probability of 1 in the bottom right of Figure A.1. If the trader is currently in the absorbing cancel state, the probability of remaining in that state in the subsequent period is 1. Given that the action for this section of the matrix is NA, we may expect the probability to enter the absorbing cancel state to be zero for all market states when the trader has a resting limit order. However, we assume that if the resting limit order transitions into an undefined state (more than three ticks from the best bid), the trader's action NA is overridden, and the order is canceled. Therefore, there can be a non-zero probability of the order being canceled, which we estimate empirically.

Transition matrix for action C

Figure A.2 illustrates the $S \times S$ section of the transition matrix, T, when the action is to cancel the resting limit order (C). Unlike the section of the transition matrix when the action is NA, this section of the transition matrix is deterministic and does not require any empirical estimation of the transition probabilities. If the trader cancels their limit order, they transition to the absorbing cancel state with certainty. Therefore, the probability of entering the absorbing cancel state, which is captured in the final column of Figure A.2, is 1 for all current states where the trader has a resting

limit order. Further, once the order is canceled, the market cannot transition to any state where the limit order still exists or executes. Thus, the "Unexecuted" and "Executed" blocks contain only zeros.

To ensure the trader only has one resting limit order at a time, we impose a restriction that any state where the trader has an inventory position or has already canceled their order cannot have another resting order. Due to this restriction, taking the action to cancel an order in a state where the trader has a long inventory position, or has canceled their order, is prohibited and has a zero probability of occurring.

Figure A.2. Transition matrix for C

Figure A.2 depicts the $S \times S$ transition matrix for the experience tuples in which the action is to cancel the resting limit order, C. States $s_i(0)$ represent states when the trader is working their limit order, whereas states $s_j(1)$ represent states when the trader's order has been executed. State $s^C(0)$ represents the absorbing order cancellation state.

Future State

Current state with action C	$s_1(0)$	$s_2(0)$	...	$s_n(0)$	$s_1(1)$	...	$s_m(1)$	$s^C(0)$
$s_1(0)$	0	...		0	0	...	0	1
$s_2(0)$	⋮	⋱			⋮	⋱		1
		Unexecuted				Executed		1
$s_{n-1}(0)$								1
$s_n(0)$	0	...		0	0		0	1
$s_1(1)$	0	...	0	0	0	...	0	0
⋮	⋮	Prohibited		0		Prohibited		0
$s_m(1)$	0	0	0	0	0	...	0	0

Full transition matrix

In Figures A.1 and A.2 we present two $S \times S$ sections of the full $2S \times S$ transition matrix, T. Specifically, Figure A.1 (A.2) is a transition matrix for all experience tuples when the action is NA (C). To generate the full transition matrix, T, we vertically stack the 2 subsections, each with dimension $S \times S$, resulting in the full transition matrix of dimension $2S \times S$. For notational convenience, we refer to $T(\langle s, a \rangle, s')$ as the probability a limit order transitions to state s' given the trader makes action a while the limit order is in state s.

B Immediate Reward

Figure B.3 shows the matrix of immediate rewards for all experience tuples that occur when the action is to do nothing, NA. If the trader's limit order remains unexecuted during the transition, the immediate reward is zero, as indicated in the upper left quadrant titled "Unexecuted". Conversely, if the trader's limit order executes, the immediate reward corresponds to the profit generated. This profit is empirically calculated using (6) and is defined as the difference between the execution price and the midpoint in the future state s'. These profits are shown in the block of Figure B.3 titled "Executed". The block of Figure B.3 titled "Long" contains the immediate profits that occur when the trader is long and the market transitions from one state to the next. We empirically estimate these immediate profits via (6), and they reflect any profit generated via a change in midpoint over a state transition.

Figure B.3. Immediate reward matrix

Figure B.3 depicts the $S \times S$ immediate reward matrix for transitioning from one state to the next. States $s_i(0)$ represent states when the trader is working their limit order, whereas states $s_j(1)$ represent states when the trader's order has been executed. State $s^C(0)$ represents the absorbing order cancellation state.

Future State

Current state with action NA	$s_1(0)$	$s_2(0)$	...	$s_n(0)$	$s_1(1)$	...	$s_m(1)$	$s^C(0)$
$s_1(0)$	0	...		0	$r_{1,n+1}$	...	$r_{1,n+m}$	0
$s_2(0)$	⋮	⋱			⋮	⋱		0
		Unexecuted				Executed		0
$s_{n-1}(0)$	0				$r_{n-1,n+1}$			0
$s_n(0)$	0	...		0	$r_{n,n+1}$		$r_{n,n+m}$	0
$s_1(1)$	0	...	0	0	$r_{n+1,n+1}$	...	$r_{n+1,n+m}$	0
⋮	⋮	Prohibited		0	⋮	Long	⋮	0
$s_m(1)$	0	0	0	0	$r_{n+m,n+1}$	...	$r_{n+m,n+m}$	0
$s^C(0)$	0	0	0	0	0	...	0	0

When the action is NA, the limit order can execute or there can be an existing long position. Either scenario can result in a non-zero immediate reward. In contrast, when the trader cancels their order (action C), the immediate reward must be zero, as no limit orders are executed and there is no inventory position. Consequently, the $S \times S$ immediate reward matrix for the action C contains only zeros.

C Illustrative example

In this appendix, we provide a simple example to illustrate the empirical estimation process of our framework via an iterative learning rule known as Q-learning defined as:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha\Big(E[R(s,a)] + \gamma \sum_{s' \in S} T(\langle s,a \rangle, s') \max_a Q_t(s',a) - Q_t(s,a)\Big), \tag{14}$$

where α is the learning rate and t is the iteration number. The Q-learning rule is a value iteration update. Watkins and Dayan (1992) show that the Q values will converge to Q^* with probability 1 if all actions are repeatedly sampled in all states and the action-values are represented discretely.

To simplify this illustrative example, we first define a simplified state-action space. We then illustrate how to empirically estimate our simplified transition probability matrix and immediate reward matrix. We conclude with a demonstration of the Q-learning rule.

C.1 State action space

Similar to the model we formulated in Section 2, in this illustrative example, our trader has two available actions. The first action is to cancel the existing limit order (C). The second action is to do nothing (NA), allowing the existing limit order to remain in the queue. However, to simplify our example, we reduce the state space by considering only the queue size at the best bid (q^{B_0}) and best ask (q^{A_0}), ignoring the queue sizes at levels behind the best bid (q^{B_1},q^{B_2}). Moreover, we discretize queue size into only two groups, $large$ and $small$.

To further reduce dimensionality, we reduce the private state variable, queue position (Q) to two possible states: $front$ and $back$, representing whether the order is in the front half or back half of the queue, respectively. These simplifications result in a state space with 8 possible states when the trader has no inventory and is executing a limit order (m), 4 possible market states when the trader has an inventory position and is no longer executing an order (n), and 1 absorbing state

when the trader cancels their order (o). Altogether, our setup consists of a total of 13 possible unique market states (S). More formally, $m = 8$, $n = 4$ $o = 1$ and $S = 13$. We define each state as follows:

$$s_k^j(I) = [I, L, Q, q^{B_0}, q^{A_0}] = \begin{cases} s_1^f(0) = [0, 0, front, small, small] \\ s_2^f(0) = [0, 0, front, small, large] \\ s_3^f(0) = [0, 0, front, large, small] \\ s_4^f(0) = [0, 0, front, large, large] \\ s_1^b(0) = [0, 0, back, small, small] \\ s_2^b(0) = [0, 0, back, small, large] \\ s_3^b(0) = [0, 0, back, large, small] \\ s_4^b(0) = [0, 0, back, large, large] \\ s_1^X(1) = [1, X, X, small, small] \\ s_2^X(1) = [1, X, X, small, large] \\ s_3^X(1) = [1, X, X, large, small] \\ s_4^X(1) = [1, X, X, large, large] \\ s^C(0) = [0, X, X, -, -] \end{cases} \tag{15}$$

where k is an index of the public market state, which is reflected by q^{B_0} and q^{A_0}. j takes on the value of f (b) if the limit order is at the front (back) half of the queue, a value of X if the trader has an inventory position and no limit order, or a value of C if the trader has canceled their order. The X term captures our restriction that no additional limit orders can be submitted once the trader has a positive inventory position or cancels their order. I captures the trader's current inventory position and $L \in \{0, X\}$ depending on whether the trader has an order resting at the best

bid, or their order is canceled of executed.[21] For each of the 8 states, where the trader is working a limit order, the trader has the choice of making the action to do nothing, NA, or cancel their existing order, C. For the states where the trader is long or has canceled their order, they can only make action NA. With the state action space defined, the input variables for (14) are the transition matrix and immediate reward matrix, which we empirically estimate.

C.2 Transition probabilities

$T(\langle s, a\rangle, s')$ represents the probability that the limit order transitions the market to state s' under action a while in state s. For example, $T(\langle s_1^f(0), NA\rangle, s_2^f(0))$ is the probability that a limit order at the front of the queue which exists when the best bid and best ask both have short queue lengths transitions to a subsequent period where the order is still at the front half of a queue and it remains unexecuted, but market conditions have changed such that the bid volume is small and the ask volume is now large.

We compute these transition probabilities empirically using the MLE estimate defined by (5). For example, to estimate $T(\langle s_1^f(0), NA\rangle, s_2^f(0))$, we observe the subsample of observations that capture state $s_1^f(0)$ (i.e., the observations that have small queue sizes on both the bid and the ask and the limit order is at the front half of the bid). Next, we compute the proportion of observations that transition to the subsequent state $s_2^f(0)$, which is reflected by the limit order still remaining in the top half of the book, but under new market conditions (i.e., the bid queue size is small and the ask queue size is large). Table C.4, reports empirical estimates of the transition probabilities using data defined in Section 3.

Figure C.4 has a distinct structure. The upper left block of the transition matrix represents states when the trader has no inventory and completes the action of do nothing, NA. This area has a strong diagonal, which reflects that an uncanceled limit order is most likely to remain in the same state in the subsequent 100ms period. For example, observing the transition probabilities for the state $s_1^f(0)$, which reflects a resting limit order at the front half of the queue when the queue

[21] In this simplified example, L is redundant because the order can only be placed at the best bid. However, we have included L for consistency with our main analysis, in which the order can rest at multiple price levels.

Figure C.4. Transition matrix

Figure C.4 depicts the $SA \times S$ transition matrix for the experience tuple in which the action is to leave the resting limit order, NA, or cancel the order, C. States $s_i(0)$ represent states when the trader is working their limit order, whereas states $s_j(1)$ represent states when the trader's order has been executed. State $s^C(0)$ represents the absorbing order cancellation state.

		Future State												
		$s_1^f(0)$	$s_2^f(0)$	$s_3^f(0)$	$s_4^f(0)$	$s_1^b(0)$	$s_2^b(0)$	$s_3^b(0)$	$s_4^b(0)$	$s_1^X(1)$	$s_2^X(1)$	$s_3^X(1)$	$s_4^X(1)$	$s^C(0)$
Current state with action NA	1 $s_1^f(0)$	**0.86**	0.02	0.02	0.00	0.02	0.00	0.00	0.00	**0.05**	0.00	0.01	0.00	0.01
	2 $s_2^f(0)$	0.03	**0.82**	0.00	0.02	0.00	0.02	0.00	0.00	0.01	**0.07**	0.01	0.01	0.01
	3 $s_3^f(0)$	0.01	0.00	**0.89**	0.03	0.01	0.01	0.01	0.00	0.01	0.00	**0.02**	0.00	0.02
	4 $s_4^f(0)$	0.00	0.01	0.02	**0.90**	0.00	0.01	0.00	0.01	0.00	0.00	0.00	**0.03**	0.02
	5 $s_1^b(0)$	0.06	0.00	0.01	0.00	**0.87**	0.03	0.02	0.00	**0.01**	0.00	0.01	0.00	0.01
	6 $s_2^b(0)$	0.00	0.07	0.00	0.01	0.03	**0.84**	0.00	0.02	0.01	**0.01**	0.01	0.01	0.01
	7 $s_3^b(0)$	0.00	0.00	0.02	0.00	0.03	0.01	**0.90**	0.03	0.00	0.00	**0.00**	0.00	0.02
	8 $s_4^b(0)$	0.00	0.00	0.00	0.02	0.00	0.02	0.02	**0.92**	0.00	0.00	0.00	**0.00**	0.02
	9 $s_1^X(1)$									**0.94**	0.03	0.03	0.00	0
	10 $s_2^X(1)$									0.04	**0.91**	0.01	0.04	0
	11 $s_3^X(1)$				0					0.03	0.01	**0.92**	0.04	0
	12 $s_4^X(1)$									0.01	0.03	0.03	**0.94**	0
	13 $s^C(0)$				0						0			1
Current state with action C	14 $s_1^f(0)$													1
	15 $s_2^f(0)$													1
	16 $s_3^f(0)$													1
	17 $s_4^f(0)$				0						0			1
	18 $s_1^b(0)$													1
	19 $s_2^b(0)$													1
	20 $s_3^b(0)$													1
	21 $s_4^b(0)$													1
	22 $s_1^X(1)$													0
	23 $s_2^X(1)$													0
	24 $s_3^X(1)$				0						0			0
	25 $s_4^X(1)$													0
	26 $s^C(0)$				0						0			0

sizes on the best bid and best ask are small, there is an 86% chance the subsequent state will be the same. However, there is also a 2% chance the subsequent state is either $s_2^f(0)$ or $s_3^f(0)$, which implies either 1) the best ask has grown to become large and the market has transitioned to $s_2^f(0)$, or 2) the best bid has grown and the market has transitioned to $s_3^f(0)$.

The section of the transition matrix for transitions from state $s_i(0)$ to state $s_j(1)$ with action NA, reports the probabilities that a resting limit order executes during the transition to the subsequent state. We observe that resting limit orders at the front of the queue (rows 1-4) have a higher probability of execution than resting limit orders at the back of the queue (rows 5-6). Further, the

probability of execution for $s_2^f(0)$ is 0.1 $(0.01 + 0.07 + 0.01 + 0.01)$, which is higher than the probability of execution for any of the other states with a resting limit order. State $s_2^f(0)$ occurs when the trader has a resting limit order at the front half of the best bid and the bid queue size is small, while the ask queue size is large. Cao et al. (2009) demonstrate that when the ask volume is larger than the bid volume, aggressive sell orders are more likely to occur and prices will decrease in the near future. Therefore, it is consistent with the literature that the highest probability of execution occurs for state $s_2^f(0)$. Moreover, the strong diagonal component of this section of the transition matrix reflects that when a resting limit order executes during the transition to the subsequent period, it is most likely that the state of the order book in the subsequent period is in the same state as the current period.

Rows 9 to 12 of Table C.4 represent the transition probabilities when the trader has an inventory position. The left block of the rows take the value of zero to ensure the trader does not have additional limit orders once a long inventory position occurs. The middle block captures the probability the trader transitions to a subsequent market state with their inventory position remaining unchanged. Given the trader has no resting limit orders, we estimate these transition probabilities using only the public state variables, which in this example are the size of the best bid and ask (q^{B_0} and q^{A_0}).

As discussed in Appendix A, we do not need to estimate transition probabilities when the action is C. When the action is C, the transition probability to any state with a resting limit is 0 and the transition probability to the absorbing state is 1. Moreover, if the trader has a long position, or is already in the absorbing state, they are prohibited to make action C, as they have no order to cancel. To uphold this constraint, rows 22 to 26 all sum to zero, which ensures there is a 0 probability that action C occurs when in these states.

We note that in rows 1-8 of Figure C.5 we report non-zero values for the probability to transition to the absorbing order cancellation state, $s^C(0)$, despite the action being NA. These non-zero values maintain our assumption that if the market transitions to a state space where the resting limit is not recognized, the action NA is over ruled by action C. Specifically, in this case, the state space only contains limit orders at the best bid. Thus, if the best bid increases during the market transition,

so that the existing limit order is no longer at the best bid, the trader will be forced to cancel the order.

C.3 Immediate rewards

Next, we require the immediate reward for all possible transitions via (6). To empirically estimate the immediate reward when the trader has a long position, we take the average change in midpoint for the subset of observations that capture the correct transition from one state to the next. For example, to estimate $R(\langle s_1^X(1), NA\rangle, s_1^X(1))$, we create a subset of observations from our full sample of data by using observations when the market is in an initial state of $s_1^X(1)$ (i.e., the queue size of the best bid and ask are both small) and the subsequent market state is the same, $s_1^X(1)$. For this subset of observations, we then take the average of (6), which is the average change in midpoint price.

To estimate the immediate reward for the execution of a limit order, we use a similar approach. For example, to estimate the immediate reward for $R(\langle s_1^f(0), NA\rangle, s_1^X(1))$ we create a subset of observations that only include observations where the trader is in state $s_1^f(0)$ (i.e., the trader has a resting limit order at the front of the best bid during market conditions where the size of the best bid and ask are small) and transitions to the subsequent state $s_1^X(1)$ (i.e., the trader has a long position when the best bid and ask queue sizes are small). For this subset of observations, we use the average immediate reward, computed via (6), which is the midpoint price in the new state less the limit order's execution price.

Figure C.5 reports the empirically estimated immediate reward for all possible transitions. Figure C.5 only reports non zero values when the trader transitions to a long position. This segmentation ensures the trader only receives an immediate reward when a limit order is executed or a the trader has a long position. Otherwise, the trader receives no immediate reward.

The reported immediate rewards are the potential gains or losses that immediately occur during the transition from one market state to the next. For example, we report the immediate rewards for state $s_1^f(0)$ in row 1. When the limit order in state $s_1^f(0)$ executes and the trader transitions to

Figure C.5. Immediate reward matrix

Figure C.5 depicts the $SA \times S$ transition matrix for the experience tuple in which the action is to leave the resting limit order, NA, or cancel the order, C. States $s_i(0)$ represent states when the trader is working their limit order, whereas states $s_j(1)$ represent states when the trader's order has been executed. State $s^C(0)$ represents the absorbing order cancellation state.

Current state with action		Future State: $s_1^f(0)$	$s_2^f(0)$	$s_3^f(0)$	$s_4^f(0)$	$s_1^b(0)$	$s_2^b(0)$	$s_3^b(0)$	$s_4^b(0)$	$s_1^X(1)$	$s_2^X(1)$	$s_3^X(1)$	$s_4^X(1)$	$s^C(0)$
NA	1	$s_1^f(0)$								**0.32**	0.25	−0.19	−0.05	0
	2	$s_2^f(0)$								−0.30	**0.47**	−0.49	−0.00	0
	3	$s_3^f(0)$								0.20	0.16	**0.43**	0.31	0
	4	$s_4^f(0)$			0					−0.40	0.43	−0.36	**0.47**	0
	5	$s_1^b(0)$								**-0.07**	−0.02	−0.24	−0.09	0
	6	$s_2^b(0)$								−0.45	**0.33**	−0.49	−0.04	0
	7	$s_3^b(0)$								−0.18	−0.15	**-0.15**	−0.14	0
	8	$s_4^b(0)$								−0.50	0.23	−0.49	**0.16**	0
	9	$s_1^X(1)$								0	0.23	−0.20	0.08	0
	10	$s_2^X(1)$								−0.23	0	−0.84	−0.25	0
	11	$s_3^X(1)$			0					0.20	0.84	0	0.24	0
	12	$s_4^X(1)$								−0.08	0.25	−0.24	0	0
	13	$s^C(0)$			0						0			0
C	14	$s_1^f(0)$												0
	15	$s_2^f(0)$												0
	16	$s_3^f(0)$												0
	17	$s_4^f(0)$			0						0			0
	18	$s_1^b(0)$												0
	19	$s_2^b(0)$												0
	20	$s_3^b(0)$												0
	21	$s_4^b(0)$												0
	22	$s_1^X(1)$												0
	23	$s_2^X(1)$												0
	24	$s_3^X(1)$			0						0			0
	25	$s_4^X(1)$												0
	26	$s^C(0)$			0						0			0

state $s_1^X(1)$, the immediate reward is 0.32, which implies the trader makes an immediate gain of 0.32 ticks, on average.

C.4 Estimation

We initialize our Q values, or long run expected profits forecasts, for each experience tuple to zero. Using the Q-learning rule defined by (14), we update our Q values for each experience tuple

recursively. For example, we update our estimate for $Q(s_1^f(0), NA)$ for the first iteration via:

$$Q_1(s_1^f(0), NA) = E[R(s_1^f(0), NA)] + \gamma \sum_{s' \in S} T(\langle s_1^f(0), NA \rangle, s') \max_{a_{t+1}} Q_t(s', a_{t+1}), \tag{16}$$

where the first term is the immediate profit for taking action NA which we compute via (7). The second term is the expected future profit conditional on taking action NA now. We observe the second term multiplies the probability of arriving in future state s' with the maximum Q value the trader can achieve by picking the optimal action a_{t+1} while in state s'. Because we have initialized all Q values to zero, on the first iteration, the $\max_{a_{t+1}} Q_t(s', a_{t+1})$ term in (16) will be zero for all s' and the trader will be indifferent to all choices of a_{t+1}. Thus, the second term of (16) is zero and we update our estimate for $Q(s_1^f(0), NA)$ for the first iteration as follows:

$$\begin{aligned} Q_1(s_1^f(0), NA) &= E[R(s_1^f(0), NA)] + \sum_{s' \in S} T(\langle s_1^f(0), a \rangle, s') \times R(\langle s_1^f(0), a \rangle, s') \\ &= (0.05 \times 0.32) + (0 \times 0.25) + (0.01 \times -0.19) + (0 \times -0.05) + \cdots + 0 \\ &= 0.0141 \end{aligned}$$

Applying the same process, we update the associated Q values for all experience tuples, which we report in Column 1 of Table C.1. Given the Q values were all initialized to 0, these first iteration values are the expected immediate profits.

On iteration two, the input values for our learning rule remain the same except for the Q value estimates, which are updated to the new values estimated in iteration 1. As a consequence, unlike in iteration 1, the $\max_{a_{t+1}} Q_t(s', a_{t+1})$ term in (16) will no longer be zero for all s' and the trader will have the option to pick the optimal action a_{t+1} conditional on the future state s' they transition to. For example, for the experience tuple $\langle s_1^f(0), NA \rangle$, the trader makes action NA, which can transition the trader to the future state $s_1^f(0)$ with probability 0.86. In this future state, the trader can make action NA or action C. Given the current Q value estimate for taking action NA while in state $s_1^f(0)$ is 0.0141, while the current Q value estimate for taking action C while in state $s_1^f(0)$ is 0, if the trader transitions to future state $s_1^f(0)$, it is optimal for the trader to take future action NA as this action results in a higher Q value.

An alternative scenario when it is not optimal for the trader to make future action NA occurs when the trader transitions to future state $s_1^b(0)$, which occurs with probability 0.02. In this state, the trader's future optimal action now differs, as it is optimal to take future action C and cancel. If the trader makes future action C while in future state $s_1^b(0)$, the associated current Q value, or long term profit, is zero. Whereas, if the trader makes future action NA, while in future state $s_1^b(0)$, the associated current Q value, or long term profit, is -0.0031.

This ability for the trader to select the optimal action when in a future state is the critical component of a reinforcement learning algorithm, allowing us to model a traders optimal management over the life-cycle of a limit order. Applying this logic, we update our second iteration estimate for $Q(s_1^f(0), NA)$ as follows:

$$
\begin{aligned}
Q_1(s_1^f(0), NA) &= E[R(s_1^f(0), NA)] + \gamma \sum_{s' \in S} T(\langle s_1^f(0), NA \rangle, s') \max_{a_{t+1}} Q_t(s', a_{t+1}) \\
&= E[R(s_1^f(0), NA)] \\
&\quad + \gamma T(\langle s_1^f(0), NA \rangle, s_1^f(0)) \max\{Q_t(s_1^f(0), NA), Q_t(s_1^f(0), C_0)\} \\
&\quad + \gamma T(\langle s_1^f(0), NA \rangle, s_2^f(0)) \max\{Q_t(s_2^f(0), NA), Q_t(s_2^f(0), C_0)\} \\
&\quad + \gamma T(\langle s_1^f(0), NA \rangle, s_3^f(0)) \max\{Q_t(s_3^f(0), NA), Q_t(s_3^f(0), C_0)\} \\
&\quad + \gamma T(\langle s_1^f(0), NA \rangle, s_4^f(0)) \max\{Q_t(s_4^f(0), NA), Q_t(s_4^f(0), C_0)\} \\
&\quad + \dots \\
&\quad + \gamma T(\langle s_1^f(0), NA \rangle, s_3^X) Q_t(s_3^X, NA) \\
&\quad + \gamma T(\langle s_1^f(0), NA \rangle, s_4^X) Q_t(s_4^X, NA) \\
&= 0.0141 \\
&\quad + 0.99\big(0.86 \times \max(0.0141, 0)\big) + 0.99\big(0.02 \times \max(0.0201, 0)\big) \\
&\quad + 0.99\big(0.02 \times \max(0.0106, 0)\big) + 0.99\big(0 \times \max(0.0141, 0)\big) + \dots \\
&\quad + 0.99\big(0.01 \times 0.0240\big) + 0.99\big(0.00 \times -0.005\big) \\
&= 0.0270
\end{aligned}
$$

Table C.1 reports the progression of our Q values estimates for each iteration of the learning rule. At iteration 200, the Q value estimates exhibit a minor deviation of less than 0.0001 from the value computed in the previous iteration. This stability indicates the Q-learning rule has converged and we can terminate the iterative process of the learning rule. We can observe the learning process of our estimation method via the progression of $Q(s_1^b(0), NA)$. In iteration 1, $Q(s_1^b(0), NA)$ takes on a value of -0.0031, but at termination, $Q(s_1^b(0), NA)$ is now positive at 0.0763. Recall that iteration 1 reports the expected immediate value if the order executes in the next transition, whereas our final iteration reports the expected value if the order is optimally managed up until execution or cancellation. $Q(s_1^b(0), NA)$ reflects the scenario in which the trader leaves an order at the back half

of the limit order book when both the bid and ask queue sizes are small. If this order was to execute immediately, the order likely faces adverse selection by a large incoming order, hence a negative immediate value. In contrast, if the order does not immediately execute, the trader can wait until favorable market conditions arrive, thereby giving a long term positive expected value.

Table C.1
Q-learning rule

This table shows the Q value estimates of the conditional expected value of a limit order for all experience tuples at the end of each iteration of the Q-learning rule defined by (14). The bottom row labeled *Difference*, reports the sum of the total change in estimates after each iteration.

	Iteration 1	Iteration 2	Iteration 3	...	Iteration 199	Iteration 200
$Q(s_1^f(0), NA)$	0.0141	0.0270	0.0387		0.1492	0.1492
$Q(s_2^f(0), NA)$	0.0201	0.0357	0.0477		0.0737	0.0737
$Q(s_3^f(0), NA)$	0.0106	0.0210	0.0311		0.1868	0.1868
$Q(s_4^f(0), NA)$	0.0141	0.0271	0.0389		0.1686	0.1686
$Q(s_1^b(0), NA)$	-0.0031	-0.0019	-0.0008		0.0763	0.0763
$Q(s_2^b(0), NA)$	-0.0065	-0.0050	-0.0038		0.0125	0.0125
$Q(s_3^b(0), NA)$	0.0000	0.0002	0.0006		0.0622	0.0622
$Q(s_4^b(0), NA)$	0.0000	0.0003	0.0008		0.0527	0.0527
$Q(s_1^X(1), NA)$	0.0009	0.0016	0.0022		-0.0025	-0.0026
$Q(s_2^X(1), NA)$	-0.0276	-0.0522	-0.0742		-0.2657	-0.2657
$Q(s_3^X(1), NA)$	0.0240	0.0456	0.0650		0.2287	0.2287
$Q(s_4^X(1), NA)$	-0.0005	-0.0011	-0.0017		-0.0230	-0.0230
$Q(s^C(0), NA)$	0.0000	0.0000	0.0000		0.0000	0.0000
$Q(s_1^f(0), C)$	0.0000	0.0000	0.0000		0.0000	0.0000
$Q(s_2^f(0), C)$	0.0000	0.0000	0.0000		0.0000	0.0000
$Q(s_3^f(0), C)$	0.0000	0.0000	0.0000		0.0000	0.0000
$Q(s_4^f(0), C)$	0.0000	0.0000	0.0000		0.0000	0.0000
$Q(s_1^b(0), C)$	0.0000	0.0000	0.0000		0.0000	0.0000
$Q(s_2^b(0), C)$	0.0000	0.0000	0.0000		0.0000	0.0000
$Q(s_3^b(0), C)$	0.0000	0.0000	0.0000		0.0000	0.0000
$Q(s_4^b(0), C)$	0.0000	0.0000	0.0000		0.0000	0.0000
Difference	0.1215	0.1024	0.0915		0.00017	0.00015

Table C.2 reports the converged Q value estimates for the states where the trader has a choice to either do nothing, NA, or cancel their order, C. The trader's optimal action is the action that gives the highest Q value. For example, when the market is in state s_1 and the trader has a limit order at the front half of the queue, the long run expected value is 0.1492 if the trader chooses to do nothing, and the long run expected profit is 0 if the trader chooses to cancel their order. Given these two scenarios, it is optimal for the trader to leave their order at the front half of the queue as this action provides a higher long term expected value.

Table C.2
Q-value estimates

Table C.2 reports the conditional expected value estimates for a limit order resting in four possible different market states (s_1,..,s_4) for the actions to leave the order (NA) or cancel the order (C).

	Front half		Back half	
	NA	C	NA	C
s_1 (small bid, small ask)	0.1492	0	0.0763	0
s_2 (small bid, big ask)	0.0737	0	0.0125	0
s_3 (big bid, small ask)	0.1868	0	0.0622	0
s_4 (big bid, bid ask)	0.1686	0	0.0527	0

Passive Ownership and Corporate Bond Lending

Amit Goyal **Yoshio Nozawa** **Yancheng Qiu***

December 2024

Abstract

An increased ownership of corporate bonds by passive funds leads to a decline in demand to borrow bonds, alleviating the short-sale constraints in the corporate bond market. As passive ownership compresses bonds' credit spreads, there is less demand from other investors to buy those bonds, reducing dealers' need to borrow the bonds to sell short for market-making. When a bond's maturity shrinks to cross the maturity cutoffs, a temporary buying pressure from passive funds increases the quantity of bonds borrowed, but the pattern reverses in the medium term due to the decline in borrowing demands from insurers and active funds. These results point to a bright side of passive ownership, which makes it easier for dealers to accommodate buying pressure from customers. Since bond borrowing is done mostly for market making and liquidity provision, the results point against short-sale bans on corporate bonds.

JEL classification: G12, G14, G23

Keywords: Short sales, corporate bonds, fixed income securities, security lending

*Amit Goyal is from Swiss Finance Institute at the University of Lausanne, email: Amit.Goyal@unil.ch, Yoshio Nozawa is from University of Toronto, email: yoshio.nozawa@rotman.utoronto.ca, and Yancheng Qiu is from University of Sydney, email: yancheng.qiu@sydney.edu.au. We thank IHS Markit, Refinitiv, and WRDS for providing us with helpful feedback and advice on understanding the data on security lending and bond holdings.

1 Introduction

Passive ownership of corporate bonds is on the rise as the popularity of exchange-traded funds (ETFs) has gained traction over the past 20 years. These passive funds differ from traditional corporate bond investors because they follow certain bond indices and aim to minimize tracking errors. Whether the rising ownership of corporate bonds by passive funds enhances or reduces the ease of secondary market transactions is widely debated among academia, policymakers, and market participants but has not reached a consensus yet.

In this paper, we examine a particular type of corporate bond transactions, bond lending and shorting, and ask whether an increase in passive ownership makes it easier or more difficult to borrow bonds and short them. We do so because bond lending and shorting is a grossly understudied area, and we do not have a clear understanding of why investors borrow bonds and short them. Market participants' motivation for shorting bonds, in turn, influences how they react to the increased passive ownership. Therefore, we aim to understand how increased ownership influences both the supply of lendable bonds and the demand.

To measure the impact of passive ownership on lending outcomes, we estimate panel regressions of lending outcome variables, such as lendable supply, quantity of bonds on loan, and borrowing fees. To identify exogenous shocks to passive ownership, we include in the regression high-dimensional fixed effects to soak up any variation in firm-level unobservables driving both the outcome and passive ownership. With issuing firm-by-quarter fixed effects, any time-varying firm-level fundamental information is accounted for. In addition, we include bond fixed effects to control for time-invariant characteristics of bonds, such as the bond's covenants and seniority, as well as other time-varying bond-level controls.

In our main estimates, we find that increased passive ownership increases lending supply. This result is expected because passive funds are natural security lenders, as in any asset market. However, we also find that the increase in ownership reduces the demand to borrow bonds. The demand decrease dominates the supply increase, and thus in the new equilibrium, we observe lower fees and lower quantity of bonds lent. Specifically, we find a one-standard-deviation increase in passive ownership significantly decreases the fee by 0.018 percentage points, or 13.85% of its inter-quartile range, while it slightly reduces the quantity of bonds lent by 0.039 percentage points. While the effect on the fee is economically large, the effect on the quantity is small because the supply and demand move in the opposite direction, canceling with each other.

These findings are in contrast to the increased ownership of insurance firms and active mutual funds. We confirm that these other types of institutional ownership increase lendable supply, just like passive ownership does. However, the reaction in demand is small and

dominated by the increased supply. Therefore, an increase in the ownership by insurance firms and active funds leads to an increase in quantity lent and a smaller decrease in borrowing fees. Therefore, passive funds are unique in that its ownership significantly decreases shorting demand.

To understand the decrease in the demand to borrow bonds, we examine how bonds' credit spreads react to the increased ownership. We find that a one-standard-deviation increase in passive ownership reduces credit spreads by 1.9 basis points. Since credit spreads move in the opposite direction with bond prices, the results indicate that the bond becomes more expensive when it is held by passive owners. This finding is consistent with the literature on passive ownership (e.g. Dannhauser 2017; Bretscher, Schmid, and Ye 2024b): since passive funds are forced to hold the security in certain indices, their inelastic demand increases the price of securities held. This reduction in credit spreads is in contrast to the ownership by other types of institutions: we find that an increase in ownership by insurance firms significantly increases credit spreads, while increased ownership by active funds has no effect on credit spreads.

The difference in the reaction of credit spreads explains why the borrowing demand reacts differently for passive ownership and insurance firms' ownership. We argue and show evidence that the main borrowers of bonds are dealers, not end-users such as hedge funds. When active customers send urgent buy orders to dealers, dealers aim to cater to this buying pressure by first looking at their inventory and potential sellers in their network. When they cannot locate the bond, they resort to borrowing bonds and selling them short to the buying customer. Thus, bond borrowing stems from customers' actions to buy rather than sell the bonds.

When passive ownership rises and credit spread declines, active customers are less likely to buy those expensive bonds, reducing their buying pressure. This reduction in turn decrease dealers' need to borrow bonds, reducing the demand to borrow bonds. In contrast, the insurance firms' and active funds' ownership do not reduce buying pressure because their bonds are not expensive.

The key to the argument above is that the main short sellers of bonds are dealers, not customers. This is the opposite of the practice in the stock market, where the main drivers of short selling are informed hedge funds who identify overvalued stocks and sell them short for speculation. In the corporate bond market, such speculative short sale is prohibitively expensive for customers who pay bid-ask spreads each time they sell short the bond and buy it back to cover. For example, the average half spreads of our bond sample is 29 basis points (bps) per transaction, the average loan tenure is roughly three months, and the average borrowing fee is 44 bps per year. Thus, if a customer borrows a bond and sells it short and

buy it back after three months, the round-trip cost is 58 bps, which is very high and more than half of the average three-month corporate bond returns of 1.05%. High bid-ask spreads completely dominate the borrowing fee, which is only 11 bps per three months.

To support this claim, we run a panel regression of the daily customer buy and sell volume on the daily changes in quantity of bonds lent. If customers short selling is the driver of the lending activity, then customer sell volume should be positively correlated with an increase in lending. We find, however, the opposite result. When the amount of bond lent increases, customer sell volume declines and customer buy volume increases. Since customer buy volume is identical to dealer sell volume, this finding suggests that an increase in lending corresponds to dealers selling these bonds. Prior to September 2017, the amount lent increases strongly three business days after dealer sell, reflecting the settlement cycle. After September 2017, the increase occurs two business days after dealer sell because the SEC shortened the standard settlement gap to two business days.

We extend the panel regression of changes in quantity on loan on bond trading volume by including other potential determinants, such as contemporaneous and past bond returns, bid-ask spreads and return volatility. Decomposing the explained variation in bond lending, we find that the contemporaneous dealer trading activity explains by far the largest variation in bond lending. If other variables, such as past returns and half spreads, capture the motives to borrow bonds for non-market-making reasons, the contributions of these factors to bond lending are minimal at best. For example, when we include all explanatory variables in the multivariate panel regression, a one-standard-deviation increase in contemporaneous dealer sell is associated with a 0.05 percentage point increase in daily changes in quantity on loan, corresponding to 23% of its standard deviation. In contrast, a one-standard-deviation increase in the contemporaneous bond return, a measure of speculative short motive, is associated with only a 0.002 percentage point increase.

Therefore, the available evidence indicates that the bond short sellers are mainly dealers and that the demand to borrow corporate bonds is driven by customers' buying pressure, not selling pressure. This finding explains why passive ownership reduces the demand to borrow bonds: the bonds' higher price alleviates speculating customers' incentive to buy them.

Our main results on the impact of passive ownership are obtained by a static comparison of bonds issued by the same firm after controlling for the influence of maturity and bond-specific time-invariant attributes such as covenants. However, the underlying mechanism linking passive ownership and lending outcomes is likely to be more dynamic. When a passive fund invests in a bond, this action itself creates temporary buying pressure, causing dealers to sell short and credit spreads to tighten. Meanwhile, the other speculating customers' aggressive purchase orders gradually diminish over time because the arrival of such

an investment opportunity is sporadic. Thus, there are two forces at work: an increase in passive ownership causes short selling to increase in the short run and decrease in the long run.

To test this conjecture, we study the maturity cutoff events as proposed by Bretscher, Schmid, and Ye (2024b). In this study, we regress changes in lending outcome variables on a dummy that equals one when a bond's remaining time to maturity crosses certain cutoff values, such as three, five, and ten years. Their idea is to use this crossing event as a positive shock to passive ownership. This is because there are more bond index funds that track short-term bond indices than those that track long-term bond indices. Thus, as a bond's time to maturity shrinks, it is more likely to be held by more passive funds, and this action is independent of the bond's fundamental value.

We follow Bretscher, Schmid, and Ye (2024b) to estimate the impact of increased passive ownership on quantity on loan. We find evidence supporting our conjecture. One to three months after the bond crosses the maturity cutoff, its quantity lent increases significantly, reflecting the increasing buying pressure of passive funds during this period. Six months after the cutoff event, the passive ownership stabilizes, the quantity on loan starts to decline, and the cumulative changes turn negative 18 months after the cutoff event. This medium-term outcome reflects the decline in borrowing demand, which is exactly what we identify in the main panel regression analysis. Thus, the dynamics of lending activities provide additional support for our main findings that passive ownership alleviates the short sale frictions and helps improve the dealers' market-making capacity.

In summary, we contribute to the literature by examining the impact of increased passive ownership of corporate bonds on bond lending, which is crucial for bond dealers' market-making activities.

Our paper is related to a strand of literature that examines the effect of changing ownership of stocks on stock lending activities. Prado, Saffi, and Sturgess (2016) investigate the effect of institutional ownership on short selling. Coles, Heath, and Ringgenberg (2022) document that increased index investing causes stocks in the index to have a higher short interest. Sikorskaya (2023), Von Beschwitz, Honkanen, and Schmidt (2023), and Palia and Sokolinski (2024) focus on passive ownership on lending outcome and argue that it is crucial to account for the reactions in both lending supply and demand.[1]

The literature on equity lending and short-sales activity is vast. Starting from Miller (1977), a significant amount of work has been done to understand the stock lending activities and their implications on stock prices and returns (e.g. D'Avolio 2002; Cohen, Diether, and

[1]All the papers find that passive ownership is associated with an *increase* in short interest and lendable supply in the equity space, but provide mixed evidence on the lending fees.

Malloy 2007; Boehmer, Jones, and Zhang 2008; Saffi and Sigurdsson 2011; Blocher, Reed, and Van Wesep 2013; Boehmer and Wu 2013; Boehmer, Jones, and Zhang 2013; Kolasinski, Reed, and Ringgenberg 2013; Engelberg, Reed, and Ringgenberg 2018; Chen, Joslin, and Ni 2018; Muravyev, Pearson, and Pollet 2022, 2023a,b). Multiple papers have argued that in the equity space, short-sellers are informed, and short-sale constraints have an economically significant effect on asset prices and stock anomalies.

In contrast, there are a handful of papers on corporate bond lending. Asquith, Au, Covert, and Pathak (2013) provide an initial look at the bond lending activities using proprietary data and report that the cost of borrowing corporate bonds is not much higher than that of borrowing stocks. Anderson, Henderson, and Pearson (2018) and Hendershott, Kozhan, and Raman (2020) study whether bond lending activities are related to subsequent bond returns. They both find evidence that an increase in bond borrowing is associated with lower subsequent returns in the high-yield bond market, but not among the investment grade bonds. Our paper differs from the three papers as we study the impact of the changing ownership landscape in the corporate bond market on bond lending activities.

This paper also contributes to the literature on the behavior of institutional investors in the corporate bond market (e.g., Becker and Ivashina 2015; Choi and Kronlund 2017; Dannhauser and Dathan 2023; Dannhauser and Karmaziene 2023; Bretscher, Schmid, Sen, and Sharma 2024a; Bretscher, Schmid, and Ye 2024b). There is a recent rise in research interest, specifically in corporate bond ETFs and their implications on valuation effect and market liquidity. Dannhauser (2017) documents that an increase in ETF ownership reduces bond yields using a research design based on the changes to Markit iBoxx index inclusion rules. Pan and Zeng (2019) and Koont, Ma, Ľuboš Pástor, and Zeng (2024) examine the influence of ETF ownership on the liquidity of underlying bonds. Dannhauser and Hoseinzade (2022) and Ma, Xiao, and Zeng (2022) show bond ETF creates flow-induced pressure and exposes the bond market to a source of destabilizing demand in times of distress. Our focus, on the other hand, is on the bond lending activity, which has not been studied.

The rest of the paper is organized as follows: In Section 2, we describe our data set; In Section 3, we present our main empirical findings; In Section 4, we investigate the factors influencing bond lending activities; In Section 5, we use alternative instruments for bond ownership; and in Section 6, we provide a concluding remark.

2 Data and Sample Construction

We compile our sample from multiple data sources: (1) IHS Markit for security lending data, (2) the Thomson Reuters eMAXX database for quarterly holdings of bond investors, (3) the Mergent Fixed Income Securities Database (FISD) database for bond characteristics, (4) the Enhanced Trade Reporting and Compliance Engine (TRACE) database for bond transaction volume and direction, and (5) the Bank of America Merrill Lynch (BAML) database for daily bond returns. This section outlines the construction of our dataset and variables, as well as presents summary statistics.

2.1 Bond Lending Data

We source our bond lending data from the Markit Securities Finance Buy-Side Analytics Data (now part of S&P) via WRDS. This database covers daily data on securities borrowing and lending activity, including the quantity on loan, the active lendable quantity, utilization ratio, rebates and borrow (loan) fees, average loan tenure, and other lending outcome variables. We select our sample based on two filters. First, we require the variables "*QuantityOnLoan*" and "*IndicativeFee*" are not missing. Next, we require the observation to be non-missing in the corporate bond database, created using Mergent FISD and TRACE.[2] The first requirement implies that all bonds in our sample have non-zero quantity on loan. Thus, our study focuses on the intensive margin. However, the requirement is necessary because we want to study the supply and demand that simultaneously drive the quantity on loan and the borrowing fee, and we do not know the fee for bonds with zero quantity on loan.

We scale the quantity on loan and lendable supply by the amount outstanding of bonds, obtained from FISD. Following recent research in the equity lending market (e.g., Muravyev, Pearson, and Pollet 2022, 2023a), we use the variable "*IndicativeFee*" to proxy for direct short-selling cost, which is a buy-side borrowing fee. Specifically, it is Markit's estimate of the expected borrow cost, in fee terms, for a hedge fund on a given day based on both borrow costs between Agent Lenders and Prime Brokers as well as rates from hedge funds to produce an indication of the current market rate. Since our main analysis is conducted at a quarterly frequency, we take the average of the daily lending variables within each bond-quarter observation.

The Markit sample after the data filters above contains 300,282 bond quarters for 17,363 bonds issued by 1,709 firms over 66 quarters from 2006 Q3 to 2022 Q4. Our sample begins

[2]We filter corporate bond data following standard approaches in the literature and provide details on the cleaning procedure in the Internet Appendix A.

in September 2006 because the bond lending data have been available at a daily frequency since then on WRDS.[3]

2.2 Bond Investor Holdings Data

The bond holdings data are from the Thomson Reuters eMAXX database at a quarterly frequency. The database mainly covers the holdings of US insurance companies, US mutual funds, and US pension funds; it does not contain bond holdings from government agencies, banks, and households.[4] To clean the eMAXX data, we start with the dollar-denominated bonds issued by US firms from Mergent FISD and restrict the sample to corporate bonds that have trade records in the Enhanced TRACE database.

Next, we carefully identify and delete duplicate observations. Duplicates might arise for two reasons. First, eMAXX presents the holding data by the time information is reported. For example, a fund's holding as of 2002 Q4 may be reported in 2003 Q1, 2003 Q2, or both. Thus, in some cases, the same bond holdings data appear in multiple reporting quarters, leading to duplicate observations. In such instances, we keep the first vintage of holdings data for each bond-quarter-fund-managing firm pair. Using the example above, for 2002 Q4 holdings, we keep the one reported in 2003 Q1 and delete the observation reported in 2003 Q2.

Second, there are funds managed by multiple managing firms, called co-managed funds. For those funds, eMAXX may create separate entries across different managing firms and another entry for total holdings.[5] To avoid double counting, we delete such duplicates arising from the co-managed funds.

Using the investor type classification codes provided by eMAXX, we group investors into the following categories: insurers (i.e., life insurance, and properties and casualties insurance), mutual funds (i.e., active funds and passive funds), and others (e.g., pension funds).[6] Since eMAXX does not separate active and passive mutual funds, we identify

[3]We have reached out to WRDS and S&P about the missing Markit Securities Finance Analytics bonds and equities data from January 2002 to August 2006. This older data used a different collection methodology compared to data from September 2006 and onward, and is no longer offered by S&P. WRDS acknowledged this issue after our inquiries: https://wrds-www.wharton.upenn.edu/pages/support/support-articles/markit/msf-analytics-2002-2005-is-legacy-version-1/. We also spotted sparse and incomplete data for the variable "*IndicativeFee*" in July 2007; however, WRDS and S&P cannot fix this issue.

[4]The eMAXX version we have subscribed to covers fixed income holdings data for North America.

[5]These observations are identified when the entry for FIRMID is CO-MANAGED.

[6]We classify an investor as an insurer if its FUNDCLASS is in (INS, LIN, PIN, RIN). We define an investor as a mutual fund if the FUNDCLASS is in (AMM, ANN, BAL, MMM, MUT, END, QUI, FOF, UIT). Thus, our broad category of mutual funds also includes money market funds, balanced funds, unit investment trusts, funds of funds, and variable annuity funds.

passive funds following three steps.

We first manually link the mutual funds in eMAXX to the CRSP Mutual Fund database (MFDB) by matching funds based on their names. Among 9,263 mutual funds in eMAXX, we could merge 1,508 bond mutual funds to MFDB. Then, we use MFDB's identifiers (index fund and ETF/ETN flags) to identify passive funds. This procedure identifies 297 (bond) funds as passive.

For those funds that MFDB does not identify as passive (bond) funds or funds that are not matched with MFDB bond funds, we further search index- or ETF-related words in fund names and classify them as passive funds if their name contains the keywords.[7] This second process adds additional 550 funds as passive funds.[8]

Finally, we manually verify each passive fund through searching and checking the managing firm official websites, Morningstar, EDGAR fund prospectus, etc, whenever necessary. After this step, we end up with a total of 847 passive funds.[9]

We create our bond ownership variables by aggregating a bond's ownership among investors by different types in each quarter. We exclude observations if the total investor holdings are larger than the amount outstanding. Then, we divide the holdings by the bond's amount outstanding for active funds, passive funds and insurance firms. Before merging with Markit bond lending data, the sample contains 1,086,821 bond-quarter observations for 90,711 bonds over 66 quarters from 2006 Q3 to 2022 Q4.

2.3 Summary Statistics

We merge the quarterly bond lending data and holdings data to create our baseline dataset. Our final sample includes 296,211 bond-quarter observations for 17,235 bonds issued by 1,706 firms from 2006 Q3 to 2022 Q4. We winsorize continuous variables at 1% and 99% by each quarter to mitigate the effects of outliers while avoiding look-ahead bias.

Table 1 presents the summary statistics of quarterly panel data on bond lending outcomes, investor ownership, and other characteristics. An average bond has loan quantity of 1.45%, lendable supply ratio of 23.72%, utilization rate of 6.73%, loan tenure of 75 days, borrowing fee of 44 basis points (bps), and credit spread of 213 bps. In terms of ownership structure,

[7] The list of keywords includes (1) words related to ETFs and index fund names (e.g., INDEX, INDX, ETF, ETN, EXCHANGE); (2) words related to bond index providers (e.g., BLOOMBERG, FTSE, BOXX, ISHARES%BOND%).

[8] These include index funds or ETFs that hold US corporate bonds and are in the eMAXX database but are not identified as bond mutual funds in MFDB.

[9] Table A2 in the Internet Appendix shows a sample list of passive funds in eMAXX. The full list of passive funds will be posted on the authors' website.

the average institutional ownership is 45.70%, of which insurance firms hold 31.41%, and mutual funds hold 13.94% on average in the sample period. In particular, the passive fund ownership is 3.63%, and the active fund ownership is 13.94%. A typical bond in our sample has a credit rating of BBB (which corresponds to a numerical value of 8.45), age of 4.94 years, time to maturity of 9.96 years, amount outstanding of $676 million, and zero trading day ratio of 35%.

Figure 2 shows the time series of ownership shares averaged across bonds within a year. In our merged sample, the ownership share of insurance companies is higher than that of other types. However, their share declines from 36.74% in 2006 to 27.69% in 2022. In contrast, the ownership share of passive mutual funds is small but increasing. It is close to zero in 2006, but increases to 6.77% in 2022.

In Figure 3, we plot the average of the lending outcome variables using all corporate bonds in our sample as well as the subsample of investment-grade and high-yield bonds. Panel A plots the average lendable supply. The supply is more than 30% of the amount outstanding in 2007 and 2008. Thereafter, its size declines steadily and remains around 20% of the bond market.

Figure 3 Panels B and C report the quantity on loan and the short loan quantity, which is the ratio of the bond lending used to short the bonds to the bond's amount outstanding.[10] Consistent with Hendershott, Kozhan, and Raman (2020), we observe a decline in quantity on loan and short loan quantity in 2009. Before the financial crisis, the amount lent represents about 4% of the amount outstanding. After the crisis, it drops to about 1% and remains stable thereafter. Comparing investment-grade bonds with high-yield bonds, high-yield bonds have a higher quantity on loan than investment-grade bonds.

Comparing Panels B and C, between 2006 and 2008, the values of the short loan quantity tend to be smaller than the quantity on loan, especially among investment grade bonds, because they are more likely to be used as collateral in financing trades. However, after 2009, the two variables are almost identical. Therefore, these data suggest that the role of financing transactions is limited, and that a large portion of the borrowed bonds are sold short.

Finally, Panel D reports the average borrowing fee. For all bonds, the fee ranges from 0.31% to 0.58% with no discernible pattern. Consistent with Asquith, Au, Covert, and Pathak (2013), the level of the borrowing fee is similar to or even slightly lower than the equity borrowing fee.[11] However, high-yield bonds have higher borrowing fees, ranging from 0.42%

[10]The variable "*Short Loan Quantity*" in Markit represents the number of securities on loan with dividend trading and financing trades removed. Markit uses a proprietary algorithm to strip out these trades.

[11]The level of the fee in our sample is higher than some of the previous research that uses the sell-side

to 0.81%. Because the cross-section of fees is skewed to the right, the fee for typical bonds is lower: the median borrowing fee, plotted in Panel E, remains lower than the averages, ranging from 0.24% to 0.43% for all bonds.

In the Internet Appendix B and C, we provide further details on the construction of daily and monthly data used in the paper. Table A1 shows the descriptive statistics of daily and monthly panel data.

3 Passive Ownership and Bond Lending Activities

3.1 Overall Sample

Empirical Method. In this section, we explore the relationship between passive bond ownership and bond lending activities, including lending supply, quantity on loan, and borrowing fees. Specifically, we run a panel regression of lending activity variable Y of bond i issued by firm k in quarter q on contemporaneous passive ownership shares,

$$Y_{i,k,q} = \beta PassiveFund_{i,k,q} + \gamma X_{i,k,q} + \alpha_{k,q} + \theta_i + \varepsilon_{i,k,q}, \tag{1}$$

where a set of control variables $X_{i,k,q}$ includes the log value of the amount outstanding, numerical rating, time to maturity, and the percentage of zero-trading days. Standard errors are double clustered at the bond and quarter levels.

Our primary variable of interest is $PassiveFund$ defined as the sum of the amount held by all passive funds divided by the bond's amount outstanding and expressed as a percentage. The slope coefficient β allows us to infer the influence of a one-percentage-point increase in passive ownership on lending activities. We also create the scaled ownership by insurance firms, $Insurer$, and that by active mutual funds $ActiveFund$ and compare the effects of passive ownership with them.

We aim to identify an exogenous variation in ownership that is orthogonal to bond issuers' characteristics that might influence lending activities. For example, if a firm has a higher default risk, then this might increase the speculative demand to borrow its bonds, while passive funds investing in high-yield bonds increase their ownership at the same time. To eliminate those unmodelled forces driving both the ownership and the outcome variables, we include the firm-quarter fixed effect in the panel regression. This procedure identifies the

database. Our database measures the borrowing fee from the perspective of ultimate borrowers. The sell-side data takes the perspective of ultimate lenders and, thus, their fee level is lower because intermediating dealers charge a higher fee to lend than to borrow.

coefficients by taking advantage of the variation across bonds with different maturity in the same quarter, issued by the same firm.

It is still possible that variation in maturity may create a mechanical correlation between the dependent variable and $PassiveFund$. For example, passive ownership may increase for short-maturity bonds while shorter maturity may reduce lending fees. In addition, a bond-level variable such as covenants and seniority may simultaneously move ownership and lending outcomes. Thus, we further control for bond fixed effects and bonds' maturity as an additional control variable to eliminate the bond- and maturity-specific shocks driving ownership and lending activities. This rich set of controls rules out potential bias in the estimated relationship between ownership and lending activities.

Main Results. We report β estimates, number of observations, and adjusted R^2 in Panel A of Table 2. We find that a one-percentage-point increase in passive ownership causes the loan quantity to fall 0.009 percentage points (pp), the lendable supply to rise 0.075 pp, and the borrowing fee to fall 0.004 pp. Since the standard deviation of $PassiveFund$ is 4.40%, a one-standard-deviation increase in passive ownership causes the loan quantity to fall by 0.039 percentage points (pp), the lendable supply to rise by 0.333 pp, and the borrowing fee to fall by 0.018 pp. The magnitudes of the reactions of these three outcome variables correspond to 2.9%, 2.5%, and 14.4% of their inter-quartile range, reported in Table 1, respectively. While the effect on the lending fee is substantial compared to its typical variation, the effect on quantity appears to be small. This estimate, however, hides something very interesting, where the shifts in the supply and demand curves almost cancel each other out.

We can infer the underlying shifts in the supply and demand curves by examining the signs of the changes in quantity and price variables. Column (2) of Table 2 shows that lendable supply increases as passive ownership increases. However, columns (1) and (3) show that the equilibrium loan quantity and fees fall. To make sense of these changes, in Panel A of Figure 1, we visualize the effect of increased passive ownership. With increased passive ownership, the increase in lendable supply indicates that the supply curve shifts outward. However, there is a decrease in the demand for bond lending that more than offsets the increased supply, resulting in even lower lending fees and a slightly lower equilibrium loan quantity. The effect on the equilibrium quantity is small because the increase in supply is offset by the decrease in demand.

The response of borrowing demand in the corporate bond market is opposite to that documented in the stock market. Specifically, Sikorskaya (2023) shows that a one-standard-deviation increase in benchmark intensity, another proxy for passive ownership, leads to a

0.348 pp and 0.032 pp *increase* in the quantity on loan and borrowing fees.[12] Thus, in the stock market, the demand for security lending appears to increase in response to increases in passive ownership. We explain below the apparent discrepancy between bonds and stocks.

Panel B of Table 2 reports the multivariate regressions including *PassiveFund*, *ActiveFund*, and *Insurer*. When the left-hand-side variable is lendable supply, the coefficients on *PassiveFund*, *ActiveFund*, and *Insurer* are 0.072 pp, 0.096 pp, and 0.100 pp, respectively. For borrowing fees, the corresponding coefficients are −0.004 pp, −0.0001 pp, and −0.001 pp, respectively. Thus, an increase in institutional ownership generally leads to an increase in bond supply and lower fees.

The difference between passive funds and other institutions arises when the left-hand variable is loan quantity. Here, a one-percentage-point increase in passive ownership reduces the loan quantity by 0.010 pp. In contrast, a one-percentage-point increase in active ownership and insurers increases the loan quantity by 0.030 and 0.019 pp, respectively. Thus, the demand for loan responds differently to ownership by different institutional types. When active fund or insurance ownership increases, borrowing demand may increase or decrease. The magnitude of the demand response, however, is dominated by changes in supply, and thus we observe price and quantity moving in opposite directions. However, in response to an increase in passive ownership, demand falls enough to dominate the increase in supply, so that price and quantity move in the same direction. We visualize these findings in Panel B of Figure 1, explaining the impact of increased ownership by insurance firms on bond lending.

These coefficients can be used to assess the impact of changing landscape of corporate bond ownership on bond lending. From 2006 to 2022, the share of passive funds, active funds, and insurer changes by 6.4 pp, 2.2 pp, and −9.1 pp. By multiplying these changes by the coefficient estimated in Table 2, we estimate that the ownership changes over the past 17 years have led to a 0.17 pp decline in loan quantity (12.6% of the inter-quartile range) and a 0.016 pp decline in lending fee (13.1% of the inter-quartile). Our estimates suggest that this structural change eases borrowing constraints, allowing dealers to engage in market-making activities more smoothly.

Mechanism. Insurance firms are known to buy and hold and their portfolio turnover rate is generally low. In eMAXX data, the average portfolio turnover rate for passive funds, active funds, and insurance firms are 3.6%, 4.9% and 1.7% per quarter, respectively. Based on this metric, insurers appear to be more inactive than passive funds. What then makes passive

[12]To obtain these values, we multiply the standard deviation of benchmark intensity, 2.56% (her Table 1), by the coefficients in Table 2. Prado, Saffi, and Sturgess (2016) examine the effect of total institutional ownership (rather than passive ownership) and find that a one standard deviation increase in ownership leads to a 0.056 pp decrease in fees.

funds different from insurers? The key to understanding this dichotomy is that passive funds follow the bond index and must trade to track the index, which includes and excludes bonds based on predetermined criteria. This generates mechanical transactions and inflates the portfolio turnover rate while pushing bond prices up in the index (Dick-Nielsen and Rossi 2018). In contrast, insurance firms are known to reach for yield (Becker and Ivashina 2015), implying that the bonds that they hold tend to be cheaper than those held by their peers.

The relation of bond price to ownership is the key to understanding why the demand to borrow a bond responds to changes in its ownership. A lower bond price motivates opportunistic investors such as hedge funds to send aggressive buy orders and dealers to sell short the bond to cater to this demand. Passive ownership alleviates this pressure by inflating bond prices. To see this, we next examine the response of bond prices.

Column (5) of Panel B, Table 2 reports the relationship between various types of institutional ownership and bonds' credit spreads, which is the difference between the corporate bond yield and the maturity-matched Treasury bond yield. Consistent with the findings of Dannhauser (2017) and Bretscher, Schmid, and Ye (2024b), higher passive ownership is associated with lower credit spreads. In our estimates, a one-percentage-point increase in passive ownership leads to a 0.004 pp decline in credit spreads. This is in contrast to insurance ownership: their ownership leads to a 0.007 pp increase in spreads, confirming their reaching-for-yield behavior. Despite their small ownership share, passive ownership reduces credit spreads, which attenuates the buying pressure of other investors and reduces dealers' demand for short bonds to cater to their trading needs.

The bonds held by passive funds are more expensive, but there may be several mechanisms behind this. For example, Reilly (2022) notes that dealers tend to include overvalued bonds in a creation basket of ETFs, which make up the majority of our passive funds. Thus, while passive ownership may or may not cause the bond price to rise, passive funds end up holding overpriced bonds with lower credit spreads due to the strategic behavior of dealers.

3.2 Subsample of Special Bonds

The effect of passive ownership on bond lending may differ depending on the reason for the lending. If a bond is special, lending is driven by increased demand to borrow it. On the other hand, non-special bonds may be lent to raise cash, which is driven by increased supply.

To understand the potential difference between bonds, we split the sample into special and non-special (GC) bonds. In the equity literature, a cutoff such as 1% of the lending fee is often used to define specialness (e.g., Sikorskaya 2023). Since the lending fee for bonds is somewhat lower than that for stocks, we do not use the same cutoff. Rather, each quarter

we define bonds in quarter q to be special if their average lending fee in quarter $q-1$ is in the top ten percentile of the corporate bond cross-section (Palia and Sokolinski 2024 also follow this rule to define special stocks). We use the lagged loan fee to define specialness because the fee in quarter q is the target we want to explain.

Using the subsample of special and GC bonds, we estimate the panel regression in Eq. (1). Table 3 Panel A reports the impact of a one-standard-deviation increase in passive ownership on the same five outcome variables in Table 2, separately for special and GC bonds. An increased passive ownership increases the lending supply and decreases borrowing fee, consistent with our full-sample results in Table 2. As expected, the magnitude of the coefficients is greater for special bonds. A one-standard-deviation increase in passive ownership increases lendable supply by 1.495 pp (9.4% of the sample average) and decreases the fee by 0.187 pp (1.1% of the sample average).

However, the effect of passive ownership on loan quantity is different for special bonds than for GC bonds. A one-standard-deviation increase in passive ownership increases the loan quantity for special bonds by an insignificant 0.086 pp, while it decreases the loan quantity for GC bonds by 0.034 pp ($t = -2.78$). Thus, the reduction in lending activity observed in the main sample is driven by GC bonds, not special bonds. This finding suggests that the impact of passive ownership is spread across a wide range of corporate bonds.

Why does the quantity on loan for special bonds increase albeit insignificantly? This is because the motivation to short special bonds is different from that for GC bonds. Anderson, Henderson, and Pearson (2018) show that informed trading occurs mainly among special bonds with high fees. Specifically, among bonds with high fees, bonds with high quantity on loan earn lower returns than those with low quantity on loan. Therefore, for special bonds, the decrease in credit spreads does not prompt the demand to fall.

3.3 Subsample of High Yield Bonds

Table 4 reports the estimation results of Eq. (1) using the subsample of investment grade and high yield bonds. We define high yield bonds if the numerical rating at the end of quarter $q-1$ is below BBB and investment grade bonds otherwise. We find that an increase in passive fund ownership has qualitatively the same effects on both investment grade and high yield bonds. In both cases, passive ownership increases the supply of bonds available for loan and reduces loan fees, indicating an increase in the supply of bonds available for loan. For the loaned quantity, a one-standard-deviation increase in passive ownership reduces the loaned quantity by 0.035 percentage point for investment grade bonds and by 0.055 percentage point for high-yield bonds. Due to the smaller sample size, the effect on high-yield bonds is

statistically indistinguishable from zero. Nevertheless, the key finding is that the demand for credit falls for both investment grade and high yield bonds in response to increased passive ownership.

4 Why Do Market Participants Borrow Corporate Bonds?

4.1 Univariate Analysis

In this section, we analyze the motivation for borrowing corporate bonds. As in Asquith, Au, Covert, and Pathak (2013), there are at least three reasons why market participants borrow bonds: 1) investors speculate on the potential decline in bond prices by borrowing bonds and selling them short, 2) dealers respond to clients' immediate buy orders by borrowing bonds and selling them short, 3) bond owners seek to finance their holdings by lending them out and receiving cash collateral. The first and second motivation generates the demand to borrow bonds, but the borrowers (and thus the short sellers) are different: in the first, customers such as hedge funds are the borrowers and short sellers of the bond; in the second, dealers are the borrowers and sellers. The last motivation generates the lending supply, which reflects the funding needs of bond owners.

To dissect these motivations, we start with a simple "smell" test using univariate regressions and then in the next section check for robustness by adding a number of control variables. As a starter, we run a panel regression of daily customer buy and sell volume scaled by the amount of bonds outstanding on day $d+h$, $Vol_{i,d+h,\xi}$, on daily changes in the amount of credit, also scaled by the amount of bonds outstanding, $dQ_{i,d}$,

$$Vol_{i,d+h,\xi} = a_{h,\xi} + b_{h,\xi} \cdot dQ_{i,d} + \varepsilon_{i,d+h,\xi}, \quad \text{where } \xi \in \{\text{'Buy','Sell'}\}, \tag{2}$$

for $h = -5, ..., 5$. We use daily changes in quantity on loan to capture the flow of activities because trading volume is also a flow variable (as opposed to a stock variable).

The slope coefficient $b_{h,\xi}$ measures the sensitivity of customer trades to changes in loaned quantity. This estimate can be used to distinguish whether customers or dealers are shorting the bonds. Suppose there is an increase in the quantity on loan driven purely by customer selling short, then we expect $b_{h,Sell} = 1$: that is, a one percentage point increase in the quantity on loan corresponds to a one percentage point increase in customer selling. If there is no trading of bonds other than those borrowed, then the R-squared of the regression will be one as well.

It is also possible that the increase in quantity on loan reflects a decrease in the number of

customers returning previously borrowed bonds. Then the increase in lending corresponds to a decrease in customer purchases. That is, customer short activity is decreasing. If this is the only driver of dQ, then we expect $b_{h,Buy} = -1$. More realistically, if the increase in borrowing is driven by both an increase in newly established short positions and a decrease in previously established short positions, we expect $0 < b_{h,Sell} < 1$, $-1 < b_{h,Buy} < 0$ and $b_{h,Sell} - b_{h,Buy} > 0$. To the extent that there are bond transactions unrelated to borrowing/lending, the regression R-squared may be lower than one.

If, on the other hand, it is dealers who borrow bonds and short them for market-making activities, then the prediction for the coefficients is the opposite. An increase in borrowing should correspond to an increase in customer *buy*, implying a positive coefficient, $0 < b_{h,Buy} < 1$. It may also correspond to a decrease in customer selling (as dealers' short covering activity decreases), implying a negative coefficient $-1 < b_{h,Sell} < 0$. If dealer short and short covering fully explains lending activities, $b_{h,Sell} - b_{h,Buy} < 0$ holds.

Finally, if bond lending is motivated by financing reasons, then lending is not associated with buying or selling the bond. Therefore, we expect the slope coefficients to be zero for both customer purchases and sales.

We estimate the regression in Eq. (2) using the daily subsample before and after September 4, 2017. We choose the cutoff date as the date when the SEC implemented a new rule for the settlement cycle of securities transactions. Before September 4, transactions are generally settled three business days after the trade date, but on September 4, this gap is reduced to two business days.[13]

Panel A of Figure 4 plots the coefficient estimates $b_{h,\xi}$ of the regression in Eq. (2) using the first subsample before September 4, 2017, along with two standard error bars. We compute standard errors by double-clustering at the bond and date level.

The plot shows a striking pattern for the coefficients on day $d-3$, which reflects the correlation between day $d-3$ volume and day d changes in borrowing quantity. We find that customer buying is strongly positively correlated with quantity on loan, while customer selling is negatively correlated. Using bonds with all credit ratings, a one percentage point increase in changes in quantity on loan corresponds to a 0.19 percentage point increase in customer buys and a 0.11 percentage point decrease in customer sells. Because $b_{h,Sell} - b_{h,Buy} = -0.3 < 0$ holds, dealers' market-making activities are an important driver of bond lending.

Panel B of Figure 4 plots the coefficient estimates $b_{h,\xi}$ for the subperiod after September 5, 2017. The figure looks similar to Panel A, except that the peak of the increase in customer

[13] In 2024, the settlement period is further reduced to one business day.

buying now shifts from $d-3$ to $d-2$, reflecting the fact that the settlement period is shortened from three to two days.

The fact that the sum of $|b_{d-3,Sell} - b_{d-3,Buy}|$ is less than one suggests that dealer short selling is an important but not the only driver of bond lending. The insensitivity of bond volume to lending may reflect the existence of financing transactions in which borrowed bonds are not sold. In addition, it is possible that customers speculate and sell borrowed bonds short, but their activity is dominated by dealers' short selling, which attenuates the magnitude of the slope coefficients. At any rate, the evidence we have so far suggests that dealer short selling is not negligible and on average greater than customer short selling.

Informed trading is more prevalent in the HY bond market as these bonds are more sensitive to issuers' default risk. Thus, we may observe more speculative short selling in HY bonds than IG bonds. In Figure A1, we show the univariate regression in Eq. (2) using the subsample of bonds based on the credit rating on day d. We find that the figures are virtually identical for IG and HY bonds, suggesting that the determinants of bond lending are similar across bonds with various credit ratings.

The effect of changing the settlement period is important for the daily data analysis using the securities lending database. On the settlement day, market participants typically do not make decisions to borrow or lend the securities. These decisions are likely to be made on trade dates that are 2 or 3 business days before settlement. There are exceptions because the settlement of security lending does not follow exactly the rule for settling outright purchase and sales transactions. In an emergency situation of failed delivery, market participants may tend to security lending transactions with very short settlement period, even on the same day. Still, these exceptions are rare. Therefore, typically, if one wants to understand the relationship between securities return and lending, then the "contemporaneous" relationship can be obtained by regressing the quantity lent on day d on the return on day $d-2$ or $d-3$.

4.2 Multivariate Analysis

To quantify the contributions of the three drivers of bond lending, we follow Diether, Lee, and Werner (2009) and regress changes in the quantity on loans on a set of explanatory variables. Specifically, the panel regression model is as follows,

$$\begin{aligned} dQ_{i,d} = {} & b_0 r_{i,d-s} + b_1 \bar{r}_{i,d-s-5,d-s-1} + b_2 Vol_{i,d-s,Buy} + b_3 Vol_{i,d-s,Sell} + b_4 \overline{dQ}_{i,d-5,d-1} \\ & + b_5 \overline{Vol}_{i,d-s-5,d-s-1} + b_6 \overline{Vol}_{i,d-s-5,d-s-1} + b_7 \bar{h}_{i,d-s-5,d-s-1,Buy} + b_8 \bar{h}_{i,d-s-5,d-s-1,Sell} \\ & + b_9 \sigma_{i,d-s-5,d-s-1} + \gamma_d + \alpha_i + \varepsilon_{i,d}, \end{aligned} \tag{3}$$

where the subscript s is 3 if d is on or before September 4, 2017 and 2 thereafter. Thus, day $d-s$ is the date when participants make the decision to borrow and sell a bond. The set of explanatory variables includes $r_{i,d-s}$, the daily return on bond i on day $d-s$; $Vol_{i,d-s,\xi}$, the daily volume with a trading side ξ scaled by amount outstanding; $h_{i,d-s,\xi}$, the half spread with a trade side ξ; $\sigma_{i,d-s-5,d-s-1}$, the bond return volatility computed over the five-day period from day $d-s-5$ to $d-s-1$. Variables with an upper bar refer to the average of the daily values over the period. To compare the economic significance of the slope estimates across variables, in this regression, all explanatory variables are standardized to have a mean of zero and a standard deviation of one. Standard errors are double clustered at the bond and day level.

Column (1) of Table 5 reports the regression estimates using contemporaneous and past bond returns as explanatory variables. Consistent with the existence of opportunistic customer short selling, the slope coefficients are positive. A one-standard-deviation increase in the contemporaneous (i.e. day $d-s$) return is associated with a 0.21 bps increase in lending, while the increase in lagged returns is associated with a 0.03 bps increase. Since the standard deviation of daily changes in the quantity on loan is 19.80 bps, the estimates are economically small, indicating that speculative short selling by customers is likely to play a minor role in explaining bond lending.

Column (2) of Table 5 adds lagged changes in quantity on loan, but the coefficients on the contemporaneous and past returns remain small.

In Columns (3) to (5), we examine the role of customer buying and selling volume. In Column (3), we use the contemporaneous buy and sell volume and the averages of the lagged volume. In Column (3), the point estimates for the coefficient on contemporaneous buys and sells are 4.54 bps and -3.90 bps, respectively. Consistent with the univariate analysis in the previous section, an increase in securities lending is strongly positively associated with a contemporaneous increase in customer buys and negatively associated with customer sells. The point estimates are economically significant when compared to the standard deviation of the right-hand-side variables. The magnitude of the estimates remains unchanged when we add other control variables such as half spreads (buys and sells separately) or the volatility of bond returns.

In Column (6), we add all the variables in the panel regression. The point estimates remain similar for all variables. The magnitude of the coefficient on the contemporaneous customer buying and selling dwarfs that on all other variables. For example, the coefficient on customer buying is about 30 times as large as that of the return, and that on the customer selling is 25 times as large as that on the return. The third and fourth largest coefficients are those on lagged customer sales (-1.64 bps) and lagged average changes in quantity on loans

(-1.67 bps). Therefore, dealers' market-making activities, in which they sell short bonds to customers, dominate other variables in explaining the variation in bond lending activity.

5 Identification Based on Maturity Cutoffs

Our main results assess the effect of increased passive ownership using within-firm variation in lending outcomes. While this is a valid approach for identifying ownership shocks, it is not the only one.

Bretscher, Schmid, and Ye (2024b) propose that one can use maturity cutoffs as a valid instrument for changing passive ownership. Specifically, they show that when the remaining maturity of a bond shrinks beyond a certain threshold, such as three or ten years, passive ownership increases. This happens because there are more short-term index funds than long-term index funds. This provides another clean identification of shocks to passive ownership, because the fundamental values of a bond remain very similar when its maturity changes from (say) 10.1 years to 9.9 years. Since Bretscher, Schmid, and Ye (2024b) study the effect of ownership on bond pricing and liquidity, we revisit their results focusing on bond lending outcomes.[14]

To assess the impact of switching ownership, we define a dummy variable that takes on a value of one if a bond's remaining time to maturity crosses the three, five, and ten year cutoffs on any day in month t and zero otherwise, denoted $Switch_{i,t}$. We then regress changes in lending outcome variables for bond i, including lending supply, quantity on loan, and lending fees. In addition, we use passive ownership as another outcome variable to verify that crossing maturity increases ownership. In this analysis, we use the monthly data constructed as described in the Internet Appendix C.

Specifically, we estimate a panel regression,

$$\Delta Outcome_i^{t-1\to t+h} = \beta^h Switch_{i,t} + Controls_{i,t-1} + \alpha_i + \lambda_t + e_{i,t}^h, \tag{4}$$

where $\Delta Outcome_i^{t-1\to t+h}$ is the change of the bond lending and ownership variables for bond i from $t-1$ to $t+h$. We set $h = -4, \ldots, 24$ to study the pre-trends, short- and medium-term impacts. $Controls_{i,t-1}$ includes the log of amount outstanding of the bond, numerical credit rating, and the fraction of zero trading days in a month. Each regression includes bond and year-month fixed effects. For this regression, we restrict to the sample that $\Delta Outcome_i^{t-1\to t+h}$

[14]Internet Appendix of Bretscher, Schmid, and Ye (2024b) also study several bond lending outcomes. Our results are very similar to theirs, but we extend the horizon for the outcome variables to examine the medium-term effect of increased passive ownership.

are all available across h for comparability. Standard errors are double-clustered at the bond and year-month levels.

Table 6 Panel A reports the coefficient estimates for passive ownership and the corresponding panel in Figure 5 plots the estimated coefficients with two-standard-error bars to visualize them. Consistent with Bretscher, Schmid, and Ye (2024b), we find that when a bond crosses the maturity cutoff, its passive ownership increases significantly. Specifically, the ownership increases 0.021 pp in the month when the bond maturity becomes less than the cutoff ($h = 0$) from a month before. While the initial reaction is statistically insignificant, the ownership gradually increases for the following nine months, with β^9 being estimated at 0.214 pp ($t = 4.54$). This increase is permanent, as the increase in ownership 24 months after crossing the cutoff is still high at 0.239 pp ($t = 3.79$). Thus, we confirm that our instrument is valid and generates non-trivial variation in passive ownership when compared with its sample average (3.63 pp) and inter-quartile range (4.91 pp).

Panels D to F in Table 6 and Figure 5 report the regression estimates in Eq. (4) for changes in quantity on loan, lendable supply, and lending fees. The response of the loan quality three, nine, 18, and 24 months after the bond crosses the cutoff is 0.06 pp, −0.02 pp, −0.11 pp, and −0.12 pp, respectively. That is, in the first three months, the loan quantity increases by a small amount, reflecting the buying pressure created by passive funds who must buy those bonds to track a bond index. However, over the medium term, the initial reaction reverses, and the quantity on loan declines. This happens because the increased passive ownership reduces the bonds' credit spreads and reduces the buying pressure from other speculative investors. As a result, dealers have to sell short bonds less than before, leading to a lower quantity on loan.

The decrease of quantity on loan identified using maturity cutoff as an instrument is qualitatively consistent with our main results based on the quarterly panel regressions with firm-quarter fixed effects. However, quantitatively, the point estimate is economically more significant. In our main result, a one-percentage-point increase in passive ownership reduces the quantity on loan by 0.087 pp. In the maturity cutoff analysis, for $h = 24$. the reaction of quantity on loan to the one-percentage-point increase in passive ownership generates a 0.500 pp (=0.119/0.239) decline in quantity on loan. This reaction is substantial given the average and inter-quartile range of quantity on loan (1.45 pp and 1.35 pp, respectively). In addition, in Panel E, lendable supply declines substantially after a bond crosses the maturity cutoff. The estimated change from $h = -1$ to $h = 24$ is −0.344 pp, which is 3.43 standard errors below zero. This is in contrast to our main results, where an increase in passive ownership raises the lendable supply.

To reconcile the apparent discrepancy in estimated reactions between two types of in-

struments, one must understand the nature of the maturity cutoff event. That is, when a bond crosses the maturity cutoff, different types of investors react *simultaneously*. To see this, in Panels B and C of Table 6, we report the changes in ownership share of insurance firms and active mutual funds. The corresponding panels in Figure 5 show the regression coefficient estimates.

When the bond crosses the cutoff, insurance firms gradually reduce their ownership share. While the changes in ownership in the month of crossing the maturity cutoff are close to zero, the cumulative changes become more negative as the horizon h increases. For $h = 24$, insurance firms' ownership declines 0.488 pp ($t = -5.22$). In contrast, active mutual funds initially reduce their share, but the effect eventually disappears over the medium term. For example, the estimated change from $h = -1$ to $h = 3$ is -0.287 pp ($t = -2.93$), but the cumulative change from $h = -1$ to $h = 24$ is insignificant 0.054 pp.

In summary, over the medium term, crossing the maturity cutoff significantly increases passive ownership and decreases insurance ownership. The decrease in insurance ownership reduces the lendable supply and dominates the increase in passive funds. Changes in insurance ownership dominate because the magnitude of the change is larger (-0.488 pp) than that of passive ownership (0.243 pp), and a one percentage point increase in insurance ownership has a larger impact on lendable supply (0.100 pp, see the unscaled coefficient in Table 2) than the same change in passive ownership (0.072 pp). As a result, the maturity cutoff event significantly reduces lendable supply, as shown in Panel E, Figure 5. This reduction in supply leads to a more pronounced decline in quantity on loan (Panel B) than that in our main results.

In contrast, the borrowing fee (Panel F) reacts little when a bond crosses the maturity cutoff. This is because the increase in passive ownership decreases the fee, while the decreased insurance ownership increases it. Since the two forces cancel each other out, the resulting reactions in the lending fee are insignificant for all horizons.

In this section, we use the event study approach to understand the impact of changing bond ownership on bond lending. The findings support our main results in Section 3, where all else equal, an increase in passive ownership reduces the quantity on loan and borrowing fees.

6 Conclusion

In this paper, we investigate the mechanism through which increased passive ownership impacts the lending activity of corporate bonds. To understand the mechanism, it is essential

to clarify why market participants borrow or lend corporate bonds, which have order-of-magnitude higher bid-ask spreads than stocks and thus their trading is costly for investors. We show that bond lending occurs mainly for dealers' market making activities and the role of speculative short sales by investors is limited. Therefore, short sale is positively related with an increase in buying pressure of investors, which propels dealers to sell short to cater to the customer demands. Thus, when a bond is more expensive, its price reduces buying pressure from speculative investors, reducing the demand to borrow corporate bonds. This is interesting because it is exactly the opposite of what would happen in the equity market, where it is less costly to sell short the security. In a low-cost environment, speculators will try to take advantage of overvalued securities by selling them short, thereby increasing the demand to borrow the security.

Because the motivation to sell short is to provide liquidity in bond trading, an increase in passive ownership reduces the demand to borrow bonds. At the same time, since passive funds are the natural lenders of the security, it increases the lendable supply, resulting in a reduction in borrowing fees. Our analysis based on the large panel data reveals that the decline in demand dominates the increase in supply, resulting in a small reduction in the equilibrium quantity of bonds borrowed.

References

Anderson, Mike, Brian J. Henderson, and Neil D. Pearson, 2018, Bond lending and bond returns, Working Paper, University of Illinois at Urbana-Champaign.

Asquith, Paul, Andrea S. Au, Thomas R. Covert, and Parag A. Pathak, 2013, The market for borrowing corporate bonds, *Journal of Financial Economics* 107, 155–182.

Becker, Bo, and Victoria Ivashina, 2015, Reaching for yield in the bond market, *Journal of Finance* 70, 1863–1902.

Bessembinder, Hendrik, Kathleen M. Kahle, William F. Maxwell, and Danielle Xu, 2008, Measuring abnormal bond performance, *Review of Financial Studies* 22, 4219–4258.

Blocher, Jesse, Adam V. Reed, and Edward D. Van Wesep, 2013, Connecting two markets: An equilibrium framework for shorts, longs, and stock loans, *Journal of Financial Economics* 108, 302–322.

Boehmer, Ekkehart, Charles M. Jones, and Xiaoyan Zhang, 2008, Which shorts are informed? *Journal of Finance* 63, 491–527.

Boehmer, Ekkehart, Charles M. Jones, and Xiaoyan Zhang, 2013, Shackling short sellers: The 2008 shorting ban, *Review of Financial Studies* 26, 1363–1400.

Boehmer, Ekkehart, and Juan Wu, 2013, Short selling and the price discovery process, *Review of Financial Studies* 26, 287–322.

Bretscher, Lorenzo, Lukas Schmid, Ishita Sen, and Varun Sharma, 2024a, Institutional corporate bond pricing, *Review of Financial Studies* forthcoming.

Bretscher, Lorenzo, Lukas Schmid, and Tiange Ye, 2024b, Passive demand and active supply: Evidence from maturity-mandated corporate bond funds, Working Paper, University of Lausanne.

Chen, Hui, Scott Joslin, and Sophie Xiaoyan Ni, 2018, Demand for crash insurance, intermediary constraints, and risk premia in financial markets, *Review of Financial Studies* 32, 228–265.

Choi, Jaewon, and Mathias Kronlund, 2017, Reaching for yield in corporate bond mutual funds, *Review of Financial Studies* 31, 1930–1965.

Cohen, Lauren, Karl B. Diether, and Christopher J. Malloy, 2007, Supply and demand shifts in the shorting market, *Journal of Finance* 62, 2061–2096.

Coles, Jeffrey L., Davidson Heath, and Matthew C. Ringgenberg, 2022, On index investing, *Journal of Financial Economics* 145, 665–683.

Dannhauser, Caitlin D, 2017, The impact of innovation: Evidence from corporate bond exchange-traded funds (etfs), *Journal of Financial Economics* 125, 537–560.

Dannhauser, Caitlin D, and Michele Dathan, 2023, Passive investors in primary bond markets, *Available at SSRN 4673698* .

Dannhauser, Caitlin D., and Saeid Hoseinzade, 2022, The unintended consequences of corporate bond etfs: Evidence from the taper tantrum, *Review of Financial Studies* 35, 51–90.

Dannhauser, Caitlin D, and Egle Karmaziene, 2023, The dealer warehouse–corporate bond etfs, *Available at SSRN 4660537* .

Dick-Nielsen, Jens, 2014, How to clean enhanced trace data, *Available at SSRN 2337908* .

Dick-Nielsen, Jens, Peter Feldhütter, Lasse Heje Pedersen, and Christian Stolborg, 2023, Corporate bond factors: Replication failures and a new framework, Copenhargen Business School Working Paper.

Dick-Nielsen, Jens, and Marco Rossi, 2018, The cost of immediacy for corporate bonds, *Review of Financial Studies* 32, 1–41.

Dickerson, Alex, Philippe Mueller, and Cesare Robotti, 2023, Priced risk in corporate bonds, *Journal of Financial Economics* 150, 103707.

Diether, Karl B., Kuan-Hui Lee, and Ingrid M. Werner, 2009, Short-sale strategies and return predictability, *Review of Financial Studies* 22, 575–607.

D'Avolio, Gene, 2002, The market for borrowing stock, *Journal of Financial Economics* 66, 271–306, Limits on Arbitrage.

Engelberg, Joseph E., Adam V. Reed, and Matthew C. Ringgenberg, 2018, Short-selling risk, *Journal of Finance* 73, 755–786.

Hendershott, Terrence, Roman Kozhan, and Vikas Raman, 2020, Short selling and price discovery in corporate bonds, *Journal of Financial and Quantitative Analysis* 55, 77–115.

Kolasinski, Adam C., Adam V. Reed, and Matthew C. Ringgenberg, 2013, A multiple lender approach to understanding supply and search in the equity lending market, *Journal of Finance* 68, 559–595.

Koont, Naz, Yiming Ma, Ľuboš Pástor, and Yao Zeng, 2024, Steering a ship in illiquid waters: Active management of passive funds, *Review of Financial Studies* .

Ma, Yiming, Kairong Xiao, and Yao Zeng, 2022, Mutual fund liquidity transformation and reverse flight to liquidity, *Review of Financial Studies* 35, 4674–4711.

Miller, Edward M., 1977, Risk, uncertainty, and divergence of opinion, *The Journal of Finance* 32, 1151–1168.

Muravyev, Dmitriy, Neil D. Pearson, and Joshua M. Pollet, 2022, Is there a risk premium in the stock lending market? Evidence from equity options, *Journal of Finance* 77, 1787–1828.

Muravyev, Dmitriy, Neil D. Pearson, and Joshua M. Pollet, 2023a, Anomalies and their short-sale costs, *Available at SSRN 4266059* .

Muravyev, Dmitriy, Neil D. Pearson, and Joshua M. Pollet, 2023b, Why does options market information predict stock returns? *Available at SSRN 2851560* .

O'Hara, Maureen, and Xing Alex Zhou, 2021, Anatomy of a liquidity crisis: Corporate bonds in the covid-19 crisis, *Journal of Financial Economics* 142, 46–68.

Palia, Darius, and Stanislav Sokolinski, 2024, Strategic borrowing from passive investors, *Review of Finance* 28, 1537–1573.

Pan, Kevin, and Yao Zeng, 2019, Etf arbitrage under liquidity mismatch, Working Paper, University of Pensylvania.

Prado, Melissa Porras, Pedro A.C. Saffi, and Jason Sturgess, 2016, Ownership structure, limits to arbitrage, and stock returns: Evidence from equity lending markets, *Review of Financial Studies* 29, 3211–3244.

Reilly, Chris, 2022, The hidden cost of corporate bond etfs, Working Paper.

Saffi, Pedro A.C., and Kari Sigurdsson, 2011, Price efficiency and short selling, *Review of Financial Studies* 24, 821–852.

Sikorskaya, Taisiya, 2023, Institutional investors, securities lending, and short-selling constraints, Working Paper, University of Chicago.

Von Beschwitz, Bastian, Pekka Honkanen, and Daniel Schmidt, 2023, Passive ownership and short selling, *Available at SSRN 4438781* .

Figure 1: Security Lending Supply and Demand

This figure illustrates the supply and demand curves for security lending markets. In Panel A, we consider an increase in passive ownership, which leads to a decreased quantity on loan and lower fees. In Panel B, we consider an increase in insurance ownership, which leads to an increase in quantity on loan and lower fees.

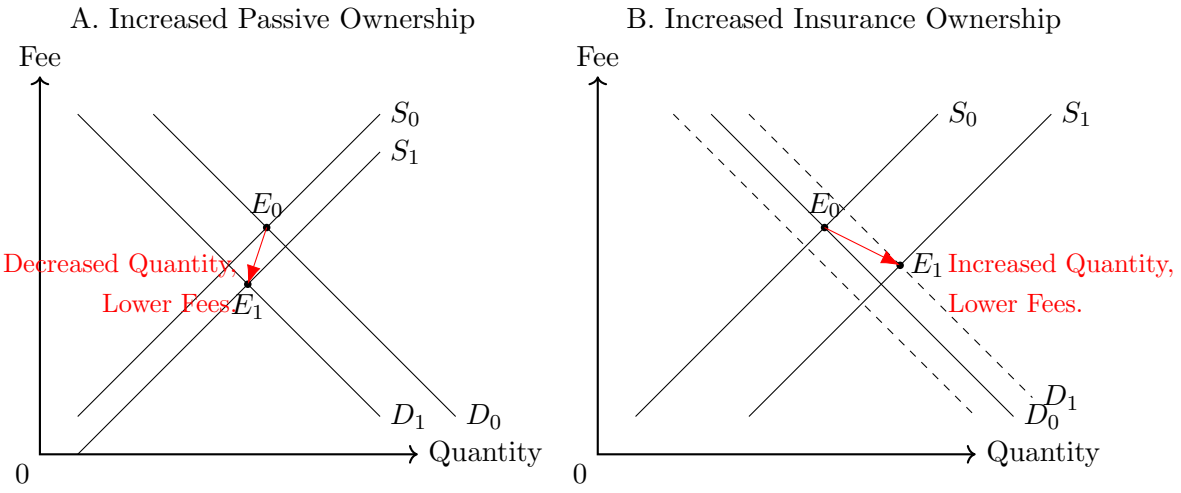


Figure 2: Time Series Plots of Bond Ownership

This figure plots the four-quarter moving average percentage ownership of corporate bonds included in our baseline quarterly panel data for each investor type from 2006 Q3 to 2022 Q4. The dotted green line plots holdings by insurance companies. The dashed blue line plots the share of bonds held by active mutual funds. The solid red line plots the share of bonds held by passive mutual funds, including index funds and ETFs. The holdings data are from eMAXX, and the amount outstanding data are from Mergent FISD. The details on sample construction can be found in Section 2.

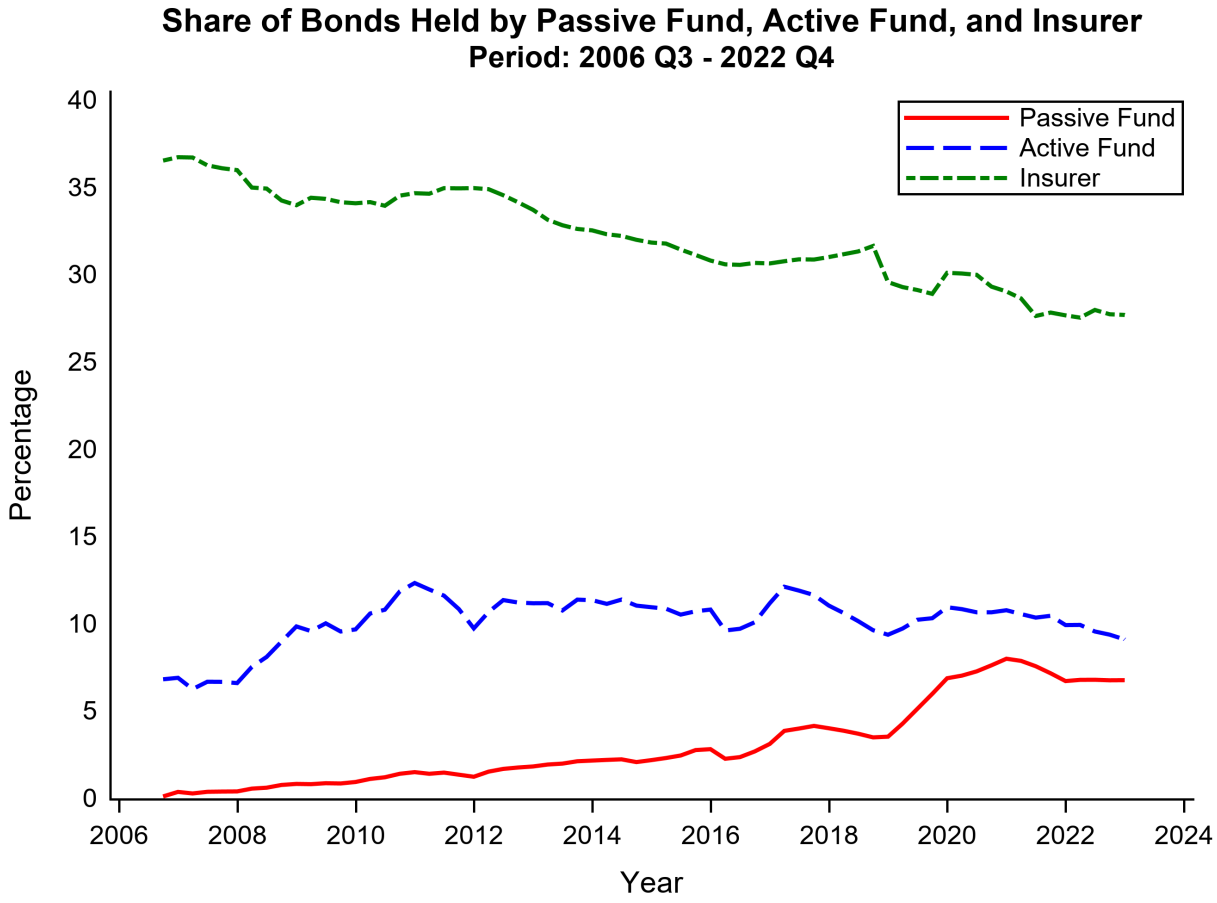
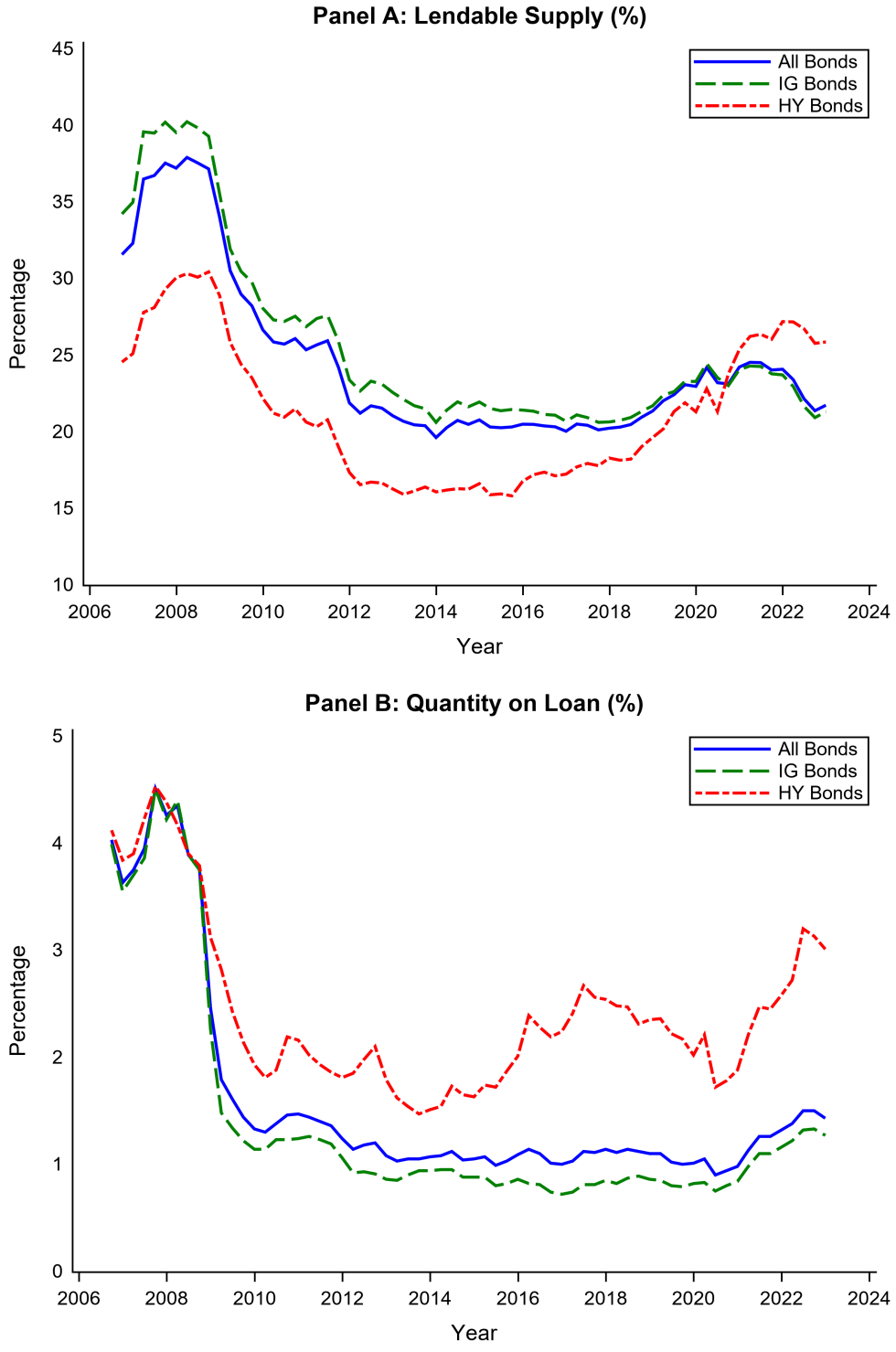
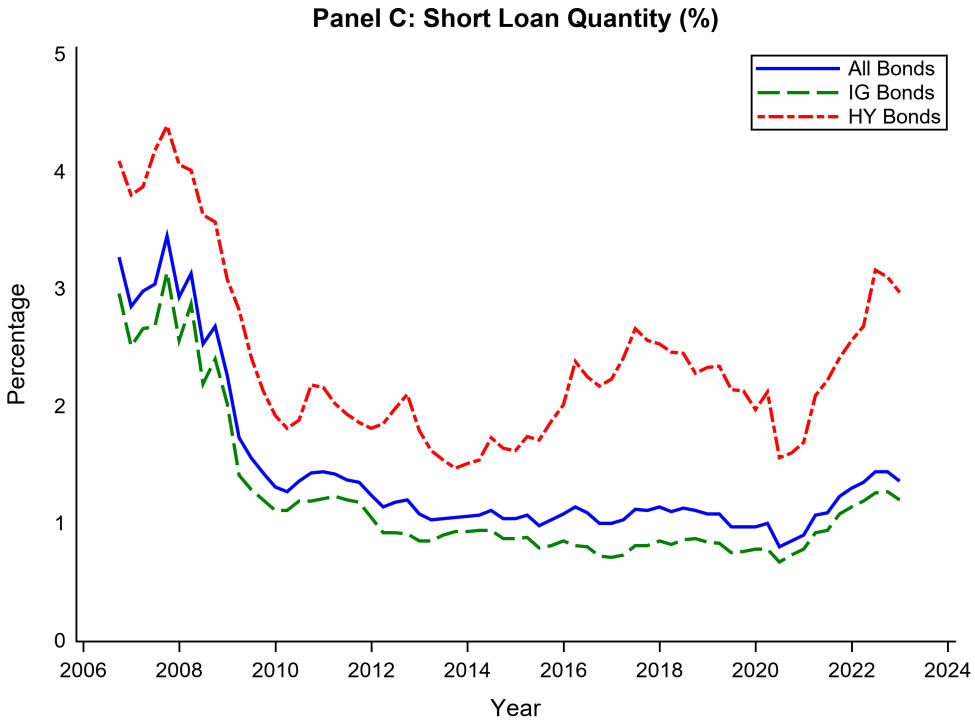
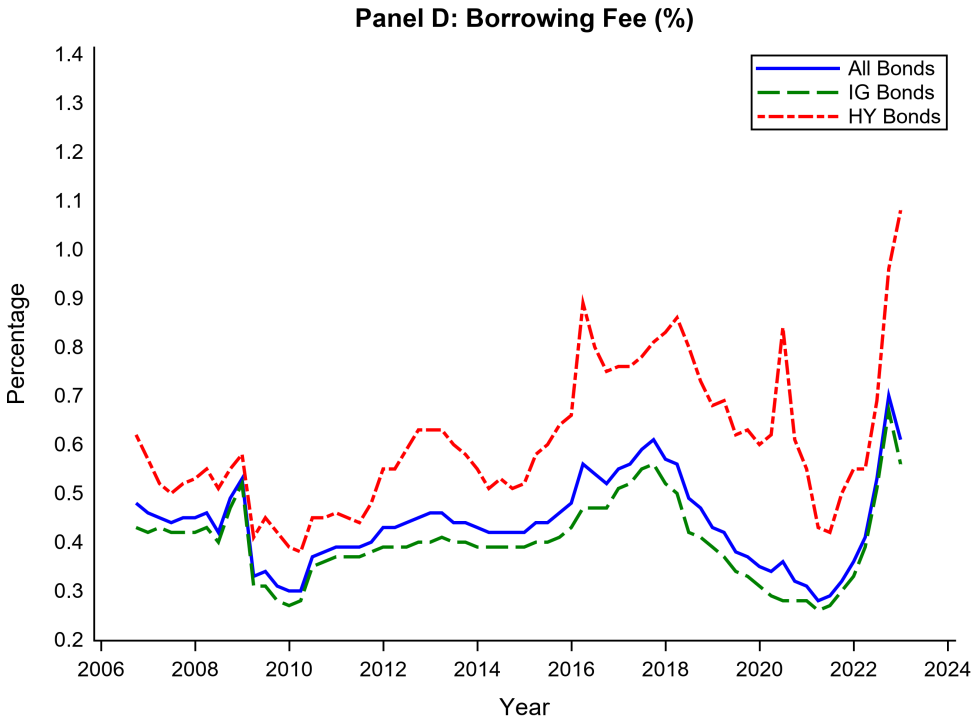


Figure 3: Time Series Plots of Bond Lending Activities

This figure plots the average lending market outcomes of corporate bonds included in our baseline quarterly panel data from 2006 Q3 to 2022 Q4. The solid blue line represents the whole sample, while the dashed green and dot red lines display investment grade and high yield bonds, respectively.




Panel C: Short Loan Quantity (%)
All Bonds
IG Bonds
HY Bonds
Percentage
0
1
2
3
4
5
2006
2008
2010
2012
2014
2016
2018
2020
2022
2024
Year


Panel D: Borrowing Fee (%)
All Bonds
IG Bonds
HY Bonds
Percentage
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
2006
2008
2010
2012
2014
2016
2018
2020
2022
2024
Year

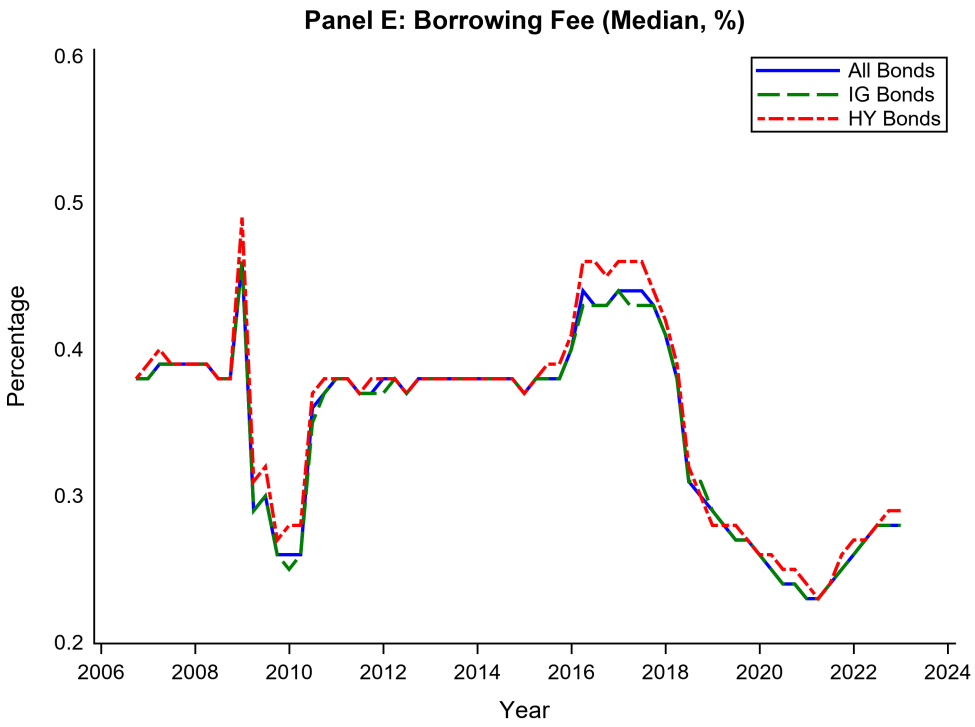

Panel E: Borrowing Fee (Median, %)
All Bonds
IG Bonds
HY Bonds
Percentage
0.6
0.5
0.4
0.3
0.2
2006
2008
2010
2012
2014
2016
2018
2020
2022
2024
Year

Figure 4: Panel Regression of Dollar Trading Volume on Changes in Quantity on Loan

This figure plots the slope coefficients of the panel regression of dealer-customer trading volume on the day $d + h$ on the day d changes in quantity on loan. Trading volume and quantity on loan are scaled by the amount outstanding. The y-axis represents a change in the percentage of the scaled dollar trading volume as a result of a one percentage change in the scaled quantity on loan. Panel A is for all bonds up to Sep 4, 2017, and Panel B is for all bonds after Sep 5, 2017.

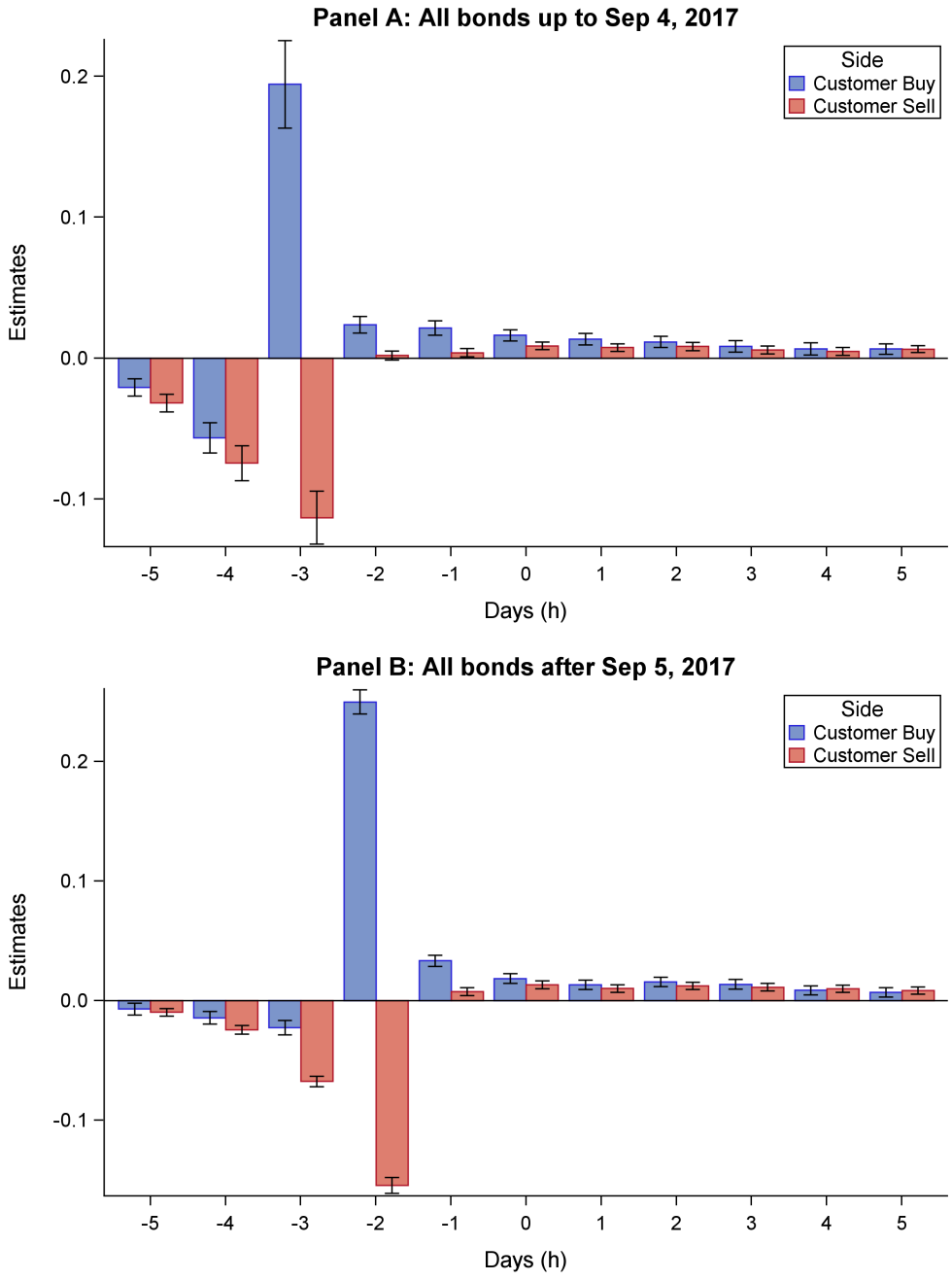


Figure 5: Investor Ownership and Bond Lending around Maturity Cutoffs

The figure plots the slope coefficients β^h from the following regression for $h \in [-4, 24]$

$$\Delta Outcome_i^{t-1 \to t+h} = \beta^h Switch_{i,t} + Controls_{i,t-1} + \alpha_t + \gamma_i + \varepsilon_{i,t}^h,$$

where $\Delta Outcome_i^{t-1 \to t+h}$ is the change of investor ownership and lending variables for bond i from $t-1$ to $t+h$. $Switch_{i,t}$ is an indicator variable equal to one if bond i crosses any one of the maturity cutoffs (i.e., 10 years, 5 years, and 3 years) in month t, and 0 otherwise. Thus, the y-axis represents the change of outcome variables relative to the pre-crossing level after a bond crosses the maturity cutoffs. Control variables include the log of the amount outstanding, credit rating, time to maturity, and the percentage of zero trading days. Each regression includes bond and year-month fixed effects. Error bars represent the two-standard-error confidence intervals, where standard errors are clustered at both the bond and year-month levels.

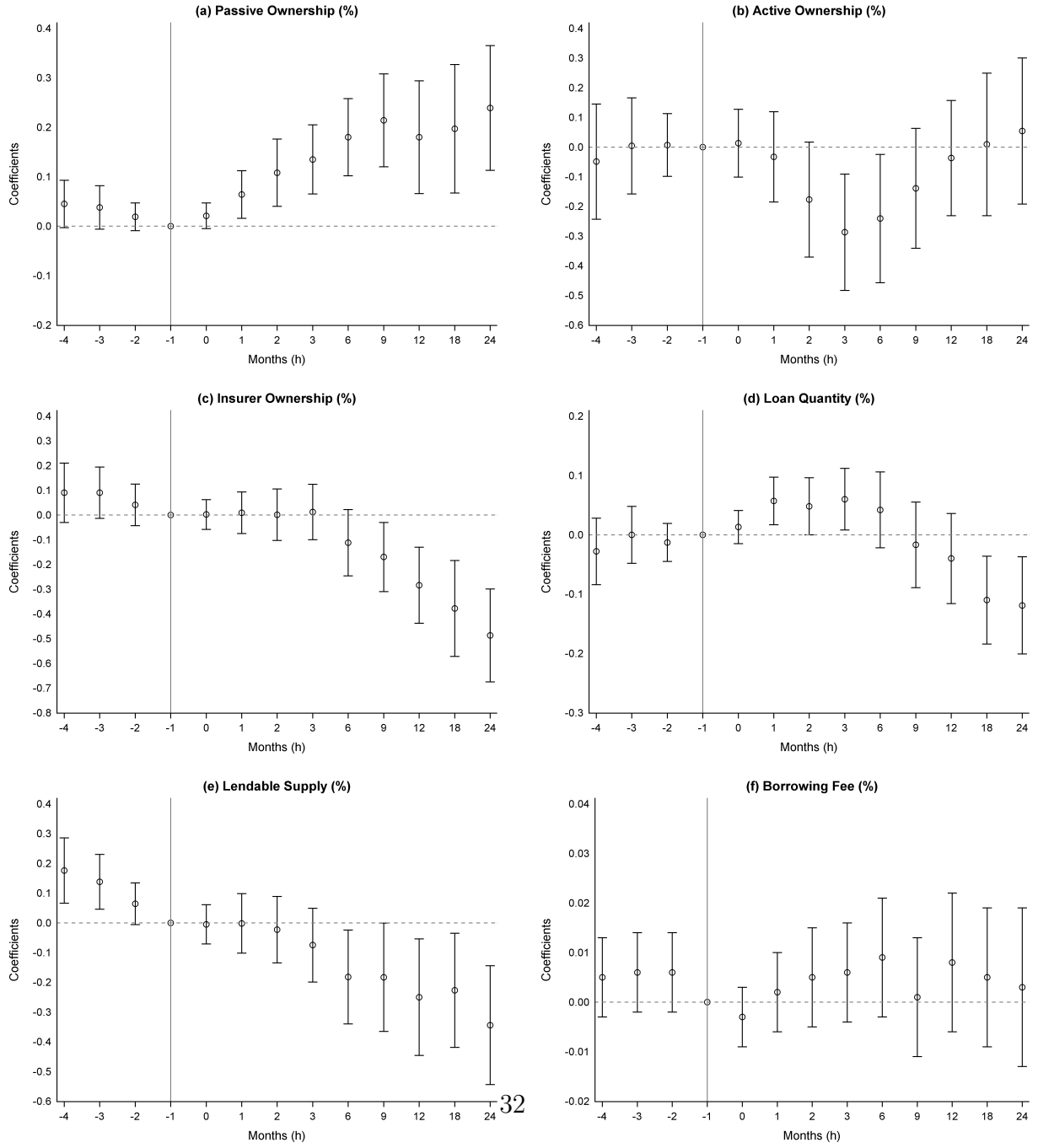


Table 1: Descriptive Statistics

This table reports the summary statistics of the main variables at the bond-quarter level. We compute quarterly averages of daily bond lending variables for each bond unless mentioned otherwise. *Loan Quantity* is defined as the quantity on loan from Markit divided by amount outstanding from Mergent FISD. *Lendable Supply* is the active lendable quantity from Markit divided by amount outstanding. *Utilization Rate* is defined as the ratio of the quantity on loan to the lendable quantity. *Loan Tenure* is the average number of days that bond loans have been open. *Borrowing Fee* is the buy-side fee paid by the ultimate borrower ("*IndicativeFee*" in Markit). *Rebate Rate* is the "*IndicativeRebate*" in Markit. *DCBS* is the cost of borrow score provided by Markit, ranging from 1 (low cost) to 10 (high cost). *Fee Risk* is the natural logarithm of the standard deviation of borrowing fees within a calendar quarter. *Recall Risk* is the natural logarithm of the standard deviation of utilization rate in a given quarter. *Lender Concentration* is a Herfindahl-Index-like measure at the bond level provided by Markit that describes the concentration of lenders. *Special1* is a dummy variable that equals one if at least one day with borrowing fee exceeds or equal to 1% in a quarter, and zero otherwise. *Special2* is a dummy variable that equals one in a given quarter if its borrowing fee is in the top decile of the fee distribution across bonds, and zero otherwise. *Credit Spread* is calculated as the average difference between the corporate bond yield and the yield of a matching Treasury bond within a quarter. h^{Buy} (h^{Sell}) is half spread from the customer buy (sell) side, defined as the quarterly average of the log price differences between customer buy (sell) trades and inter-dealer trades following O'Hara and Zhou (2021). We match customer buy (sell) trades with the closest in-time inter-dealer trade over the past five trading days with replacement. *Total Ownership* is the share of bonds held by all the investors in eMAXX. *Insurer* and *Mutual Fund* are the shares of bonds held by insurance firms and mutual funds, respectively. We use the investor type classification code provided by eMAXX to group investors into insurance firms and mutual funds. We manually link eMAXX to the CRSP Mutual Fund database by matching funds based on their names and use index fund and ETF flags (as well as search keywords in fund names) to further decompose mutual fund ownership into *Passive Fund* (i.e., index funds and ETFs) and *Active Fund*. *Amount* is the amount of bonds outstanding in millions of dollars. *Rating* is the numerical rating score, where 1 refers to a AAA rating by S&P and Aaa by Moody's, 21 refers to a C rating for both S&P and Moody's. *Age* is the age of a bond in years. *Maturity* is the time to maturity in years. *ZTD* is the percentage of zero trading days in a given quarter. To mitigate the influence of outliers, we winsorize variables at 1% and 99% for each quarter. The combined bond data are from Mergent FISD, TRACE, Markit, and eMAXX. The sample includes 17,235 bonds across 1,706 firms from 2006 Q3 to 2022 Q4.

Variable	Mean	SD	P1	P25	P50	P75	P99	IQR	Obs
Loan Quantity (%)	1.45	2.59	0.01	0.16	0.51	1.51	12.06	1.35	296,211
Lendable Supply (%)	23.72	10.96	1.68	16.36	22.69	29.71	56.91	13.35	296,211
Utilization Rate (%)	6.73	12.03	0.03	0.78	2.46	7.00	65.44	6.22	296,211
Loan Tenure (days)	74.78	87.80	1.00	23.29	44.19	89.28	462.32	65.99	296,211
Borrowing Fee (%)	0.44	0.49	0.18	0.28	0.37	0.40	2.96	0.13	296,211
Rebate Rate (%)	0.56	1.45	−2.12	−0.25	−0.12	1.08	4.90	1.33	296,211
DCBS	1.06	0.29	1.00	1.00	1.00	1.00	2.78	0.00	296,211
Fee Risk	−2.85	1.01	−5.26	−3.45	−2.90	−2.48	0.14	0.97	251,697
Recall Risk	−0.29	1.75	−6.66	−1.08	0.01	0.86	2.68	1.93	289,670
Lender Concentration	0.49	0.32	0.00	0.29	0.49	0.73	1.00	0.44	296,211
Special1 (fee $\geq$ 1%)	0.14	0.34	0.00	0.00	0.00	0.00	1.00	0.00	296,211
Special2 (top decile)	0.10	0.30	0.00	0.00	0.00	0.00	1.00	0.00	296,211
Credit Spread (%)	2.13	2.67	0.23	0.88	1.41	2.40	11.54	1.52	291,571
h^{Buy} *(%)*	0.28	0.88	−1.59	0.01	0.15	0.43	3.15	0.43	284,345
h^{Sell} *(%)*	0.27	0.97	−2.11	0.00	0.16	0.45	3.09	0.45	282,873
Total Ownership (%)	45.70	17.58	8.82	33.19	45.18	57.71	86.73	24.52	296,211
Insurer (%)	31.41	20.72	0.40	14.04	28.40	46.03	82.19	32.00	296,211
Mutual Fund (%)	13.94	12.33	0.00	4.84	10.52	19.47	53.48	14.63	296,211
Passive Fund (%)	3.63	4.40	0.00	0.49	2.60	5.40	16.51	4.91	296,211
Active Fund (%)	10.28	11.41	0.00	1.96	6.19	14.48	49.12	12.52	296,211
Amount ($ mil)	676	568	100	300	500	775	3,000	475	296,211
Rating	8.45	3.10	1.50	6.50	8.00	10.00	17.00	3.50	296,211
Age (years)	4.94	4.48	0.32	1.80	3.64	6.61	21.45	4.82	296,211
Maturity (years)	9.96	8.68	1.13	3.71	6.55	14.27	29.69	10.56	296,211
ZTD (%)	34.88	29.86	0.00	6.35	28.57	59.38	96.83	53.03	296,211

Table 2: Passive Ownership and Bond Lending Activities

This table presents the results from regressing bond lending outcomes and credit spreads on ownership of institutional investors. The dependent variables are quarterly averages of loan quantity, lendable supply, borrowing fee, DCBS, and credit spread. *Passive Fund*, *Active Fund*, and *Insurer* represent factions of bond par amount held by passive mutual funds, actively managed mutual funds, and insurance firms, respectively. Bond control variables include the log value of amount outstanding, rating, time to maturity, and the fraction of zero-trading days. The variable definitions can be found in Table 1. We include bond and firm × quarter effects in each regression. We double cluster standard errors by firm and year-quarter, and t-statistics are in parentheses. *, **, and *** indicate the significance at the 10%, 5%, and 1% levels, respectively. The sample period is from 2006 Q3 to 2022 Q4.

	Loan Quatity	Lendable Supply	Borrowing Fee	DCBS	Credit Spread
	(1)	(2)	(3)	(4)	(5)
Panel A: Passive Funds Only					
Passive Fund	-0.0087^{***}	0.0749^{***}	-0.0041^{***}	-0.0022^{***}	-0.0042^{***}
	(-2.94)	(5.53)	(-3.44)	(-3.20)	(-2.76)
Bond Controls	Yes	Yes	Yes	Yes	Yes
Firm×Qtr FE	Yes	Yes	Yes	Yes	Yes
Bond FE	Yes	Yes	Yes	Yes	Yes
Observations	280,330	280,330	280,330	280,330	275,701
Adjusted R^2	0.598	0.811	0.465	0.485	0.952
Panel B: Passive Funds Plus Other Investors					
Passive Fund	-0.0096^{***}	0.0719^{***}	-0.0041^{***}	-0.0022^{***}	-0.0042^{***}
	(-3.23)	(5.10)	(-3.41)	(-3.18)	(-2.86)
Active Fund	0.0295^{***}	0.0962^{***}	-0.0001	-0.0000	0.0001
	(11.28)	(8.12)	(-0.18)	(-0.18)	(0.08)
Insurer	0.0192^{***}	0.1004^{***}	-0.0011^{**}	-0.0005^{**}	0.0070^{***}
	(5.13)	(9.22)	(-2.48)	(-2.16)	(8.37)
Bond Controls	Yes	Yes	Yes	Yes	Yes
Firm×Qtr FE	Yes	Yes	Yes	Yes	Yes
Bond FE	Yes	Yes	Yes	Yes	Yes
Observations	280,330	280,330	280,330	280,330	275,701
Adjusted R^2	0.602	0.814	0.465	0.486	0.953

Table 3: Passive Ownership and Bond Lending Activities, Subsample Results by Specialness

This table presents the results from regressing bond lending outcomes and credit spreads on ownership of institutional investors. The results are separately reported for special bonds and general collateral (GC) bonds. A bond is defined as special in a given quarter if its lagged borrowing fee is in the top decile of the fee distribution across bonds, and as GC, otherwise. Bond control variables include the log value of amount outstanding, rating, time to maturity, and the fraction of zero-trading days. All continuous independent variables are standardized to have a mean of zero and a standard deviation of one. We include bond and firm × quarter effects in each regression. We double cluster standard errors by firm and year-quarter, and t-statistics are in parentheses. *, **, and *** indicate the significance at the 10%, 5%, and 1% levels, respectively. The sample period is from 2006 Q3 to 2022 Q4.

	Special					GC				
	Loan Quantity	Lendable Supply	Borrowing Fee	DCBS	Credit Spread	Loan Quantity	Lendable Supply	Borrowing Fee	DCBS	Credit Spread
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Passive Funds Only										
Passive Fund	0.086	1.495***	−0.187***	−0.114***	−0.036	−0.034***	0.243***	−0.004***	−0.001**	−0.014**
	(0.62)	(3.99)	(−3.50)	(−3.21)	(−0.63)	(−2.78)	(5.30)	(−3.07)	(−2.03)	(−2.13)
Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm×Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bond FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,463	12,463	12,463	12,463	11,987	243,516	243,516	243,516	243,516	240,026
Adjusted R^2	0.769	0.756	0.609	0.616	0.959	0.572	0.832	0.238	0.157	0.953
Panel B: Passive Funds Plus Other Investors										
Passive Fund	0.043	1.407***	−0.185***	−0.113***	−0.039	−0.037***	0.233***	−0.004***	−0.001**	−0.014**
	(0.30)	(3.99)	(−3.46)	(−3.17)	(−0.69)	(−3.00)	(4.88)	(−3.05)	(−2.04)	(−2.24)
Active Fund	0.501***	0.927***	−0.027	−0.010	0.021	0.275***	1.023***	0.001	0.001	0.017
	(4.25)	(4.65)	(−0.77)	(−0.48)	(0.33)	(9.72)	(7.54)	(0.38)	(0.73)	(1.64)
Insurer	0.3757*	1.796***	−0.053	−0.004	0.190	0.364***	2.020***	−0.003	−0.000	0.158***
	(1.82)	(3.81)	(−0.64)	(−0.07)	(1.30)	(4.56)	(9.46)	(−0.83)	(−0.10)	(9.36)
Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm×Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bond FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,463	12,463	12,463	12,463	11,987	243,516	243,516	243,516	243,516	240,026
Adjusted R^2	0.774	0.759	0.609	0.616	0.959	0.575	0.835	0.238	0.157	0.953

Table 4: Passive Ownership and Bond Lending Activities, Subsample Results by Credit Rating

This table presents the results from regressing bond lending outcomes and credit spreads on ownership of institutional investors. The results are separately reported for investment grade (IG) and high yield (HY) bonds. A bond is defined as high yiled if its credit rating at the end of last quarter is below BBB, and as investment grade, otherwise. Bond control variables include the log value of amount outstanding, rating, time to maturity, and the fraction of zero-trading days. All continuous independent variables are standardized to have a mean of zero and a standard deviation of one. We include bond and firm × quarter effects in each regression. We double cluster standard errors by firm and year-quarter, and t-statistics are in parentheses. *, **, and *** indicate the significance at the 10%, 5%, and 1% levels, respectively. The sample period is from 2006 Q3 to 2022 Q4.

	IG					HY				
	Loan Quantity	Lendable Supply	Borrowing Fee	DCBS	Credit Spread	Loan Quantity	Lendable Supply	Borrowing Fee	DCBS	Credit Spread
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Passive Funds Only										
Passive Fund	-0.035^{**}	0.286^{***}	-0.015^{***}	-0.008^{***}	-0.010^{*}	-0.055	0.814^{***}	-0.052^{***}	-0.031^{***}	-0.076^{**}
	(-2.60)	(5.27)	(-3.20)	(-2.86)	(-1.80)	(-0.97)	(3.66)	(-3.66)	(-3.75)	(-2.55)
Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm×Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bond FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	233,432	233,432	233,432	233,432	229,301	46,107	46,107	46,107	46,107	45,604
Adjusted R^2	0.569	0.819	0.313	0.324	0.930	0.665	0.802	0.705	0.692	0.950
Panel B: Passive Funds Plus Other Investors										
Passive Fund	-0.039^{***}	0.268^{***}	-0.015^{***}	-0.008^{***}	-0.011^{*}	-0.068	0.800^{***}	-0.052^{***}	-0.031^{***}	-0.076^{**}
	(-2.86)	(4.84)	(-3.21)	(-2.88)	(-1.95)	(-1.22)	(3.63)	(-3.65)	(-3.73)	(-2.52)
Active Fund	0.256^{***}	1.178^{***}	0.012^{*}	0.007^{**}	0.012	0.397^{***}	0.962^{***}	-0.012	-0.008	0.018
	(7.34)	(7.25)	(1.96)	(2.10)	(1.33)	(10.21)	(4.90)	(-1.44)	(-1.40)	(0.88)
Insurer	0.363^{***}	1.836^{***}	-0.015	-0.006	0.147^{***}	0.370^{**}	2.924^{***}	-0.026	-0.020	0.165^{**}
	(4.42)	(8.48)	(-1.66)	(-1.19)	(10.40)	(2.46)	(6.08)	(-0.94)	(-1.07)	(2.28)
Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm×Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bond FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	233,432	233,432	233,432	233,432	229,301	46,107	46,107	46,107	46,107	45,604
Adjusted R^2	0.572	0.821	0.314	0.324	0.931	0.670	0.807	0.705	0.692	0.950

Table 5: Panel Regression of Daily Changes in Quantity on Loan

This table reports the estimates from the panel regression of changes in the quantity on loan for all bonds:

$$dQ_{i,t} = \beta X_{i,t} + \rho dQ_{i,t-5,t-1} + \alpha_t + \gamma_i + \varepsilon_{i,t},$$

where a set of explanatory variables includes the daily return r_t, the average of the past 5 days' returns $r_{t-5,t-1}$, the turnover rate of customer buy ($Turn^{Buy}$) and sell ($Turn^{Buy}$) on day t and averages from $t-5, t-1$, half spreads for buy trades (h^{buy}) and sell trades (h^{sell}) averaged from day $t-5$ to $t-1$, and standard deviation of returns over the last five days $\sigma(r)_{t-5,t-1}$. Daily changes in the amount outstanding are scaled by the amount outstanding of the bond on day t and expressed as a percentage. Bond controls include the natural logarithm of the amount outstanding, credit ratings, and time to maturity. The variables on the right-hand side are standardized so that they have a mean of zero and a standard deviation of one. We include bond and date fixed effects in each regression specification. We double cluster standard errors by bond and date, and t-statistic are given in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively. We require each bond to have at least 252 daily observations in the regression. The sample includes 11,411 bonds across 1,259 firms from September 13, 2006 to December 30, 2022.

	(1)	(2)	(3)	(4)	(5)	(6)
r_t	0.0021^{***}	0.0022^{***}				0.0016^{***}
	(9.86)	(10.18)				(8.07)
$r_{t-5,t-1}$	0.0003^{*}	0.0007^{***}				0.0013^{***}
	(1.81)	(3.79)				(7.23)
$dQ_{t-5,t-1}$		-0.0139^{***}		-0.0167^{***}	-0.0143^{***}	-0.0167^{***}
		(−22.92)		(−25.93)	(−23.75)	(−25.97)
$Turn_t^{Buy}$			0.0454^{***}	0.0458^{***}	0.0441^{***}	0.0458^{***}
			(111.84)	(112.22)	(109.61)	(112.07)
$Turn_t^{Sell}$			-0.0390^{***}	-0.0389^{***}	-0.0394^{***}	-0.0389^{***}
			(−121.86)	(−121.68)	(−122.44)	(−121.69)
$Turn_{t-5,t-1}^{Buy}$			0.0003	0.0060^{***}		0.0060^{***}
			(1.48)	(20.20)		(20.18)
$Turn_{t-5,t-1}^{Sell}$			-0.0107^{***}	-0.0165^{***}		-0.0164^{***}
			(−61.98)	(−55.52)		(−55.28)
$h_{t-5,t-1}^{Buy}$			−0.0001	−0.0001		-0.0002^{**}
			(−0.85)	(−1.09)		(−1.89)
$h_{t-5,t-1}^{Sell}$			−0.0000	−0.0001		0.0003^{***}
			(−0.03)	(−0.60)		(3.21)
$\sigma(r)_{t-5,t-1}$					-0.0021^{***}	-0.0011^{***}
					(−9.09)	(−5.54)
Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes
Bond FE	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,693,324	10,693,324	10,693,324	10,693,324	10,693,324	10,693,324
Adjusted R^2	0.010	0.015	0.052	0.059	0.055	0.059

Table 6: Investor Ownership and Bond Lending Activities around Maturity Cutoffs

The figure plots the slope coefficients β^h from the following regression for $h \in [-4, 24]$

$$\Delta Outcome_i^{t-1 \to t+h} = \beta^h Switch_{i,t} + Controls_{i,t-1} + \alpha_t + \gamma_i + \varepsilon_{i,t}^h,$$

where $\Delta Outcome_i^{t-1 \to t+h}$ is the change of investor ownership and lending variables for bond i from $t-1$ to $t+h$. We require the outcome variable changes to be available for all h. $Switch_{i,t}$ is an indicator variable equal to one if bond i crosses any one of the maturity cutoffs (i.e., 10 years, 5 years, and 3 years) in month t, and 0 otherwise. Control variables include the log of the amount outstanding, credit rating, time to maturity, and the percentage of zero trading days. Each regression includes bond and year-month fixed effects. We double cluster standard errors by firm and year-month, and t-statistics are in parentheses. *, **, and *** indicate the significance at the 10%, 5%, and 1% levels, respectively. The sample includes 311,112 bond-month observations for 9,668 corporate bonds issued by 1,181 firms from February 2007 to December 2022.

h	-4	-3	-2	0	1	2	3	6	9	12	18	24
						Panel A: LHV = Δ*Passive Fund*$^{t-1 \to t+h}$						
Switch	0.045*	0.038*	0.019	0.021	0.064***	0.108***	0.135***	0.180***	0.214***	0.180***	0.197***	0.239***
	(1.86)	(1.74)	(1.36)	(1.56)	(2.69)	(3.20)	(3.87)	(4.57)	(4.54)	(3.17)	(3.01)	(3.79)
						Panel B: LHV = Δ*Active Fund*$^{t-1 \to t+h}$						
Switch	-0.049	0.004	0.007	0.013	-0.033	-0.177^{*}	-0.287^{***}	-0.241^{**}	-0.139	-0.037	0.009	0.054
	(-0.50)	(0.04)	(0.14)	(0.24)	(-0.43)	(-1.82)	(-2.93)	(-2.22)	(-1.38)	(-0.39)	(0.08)	(0.44)
						Panel C: LHV = Δ*Insurer*$^{t-1 \to t+h}$						
Switch	0.090	0.090*	0.041	0.002	0.009	0.001	0.012	-0.112^{*}	-0.170^{**}	-0.284^{***}	-0.378^{***}	-0.487^{***}
	(1.49)	(1.72)	(0.97)	(0.06)	(0.22)	(0.02)	(0.21)	(-1.67)	(-2.45)	(-3.69)	(-3.91)	(-5.21)
						Panel D: LHV = Δ*Loan Quantity*$^{t-1 \to t+h}$						
Switch	-0.028	-0.000	-0.013	0.013	0.057***	0.048**	0.060**	0.042	-0.017	-0.040	-0.110^{***}	-0.119^{***}
	(-1.01)	(-0.00)	(-0.79)	(0.97)	(2.88)	(2.01)	(2.31)	(1.34)	(-0.47)	(-1.04)	(-2.98)	(-2.92)
						Panel E: LHV = Δ*Lendable Supply*$^{t-1 \to t+h}$						
Switch	0.176***	0.138***	0.064*	-0.005	-0.002	-0.023	-0.075	-0.182^{**}	-0.183^{**}	-0.250^{**}	-0.227^{**}	-0.344^{***}
	(3.22)	(2.99)	(1.80)	(-0.16)	(-0.04)	(-0.42)	(-1.21)	(-2.32)	(-2.01)	(-2.56)	(-2.35)	(-3.43)
						Panel F: LHV = Δ*Borrowing Fee*$^{t-1 \to t+h}$						
Switch	0.005	0.006*	0.006*	-0.003	0.002	0.005	0.006	0.009	0.001	0.008	0.005	0.003
	(1.17)	(1.77)	(1.80)	(-1.01)	(0.58)	(1.07)	(1.16)	(1.44)	(0.10)	(1.07)	(0.70)	(0.36)

Internet Appendix
"Passive Ownership and Corporate Bond Lending"

A Corporate Bond Filters

In this section, we describe our procedure to filter corporate bonds based on the Mergent Fixed Income Securities Database (FISD) database and the Enhanced Trade Reporting and Compliance Engine (TRACE) database from WRDS.

TRACE data contains transaction prices and volume, trade direction, and the exact date and time of each trade. Following Dick-Nielsen (2014), we clean the TRACE data, remove canceled transaction records, and adjust records that are subsequently corrected or reversed. We also follow Bessembinder, Kahle, Maxwell, and Xu (2008) to correct potential data errors and remove observations in enhanced TRACE data with large return reversals, defined as a 20% or greater return followed by a 20% or greater return of the opposite sign. We merge the TRACE database with Mergent FISD to collect information on bond characteristics such as amount outstanding, credit rating, and time to maturity.

Following the recent literature (e.g., Dickerson, Mueller, and Robotti 2023; Dick-Nielsen, Feldhütter, Pedersen, and Stolborg 2023), we apply additional filters to eliminate (1) bonds that are not listed or traded in the U.S. public market; (2) bonds that are U.S. Government, private placements, mortgage-backed, asset-backed, agency-backed, or equity-linked;[15] (3) convertible bonds or bonds with a floating coupon rate or an odd frequency of coupon payments; (4) bonds that have less than one year to maturity; (5) bond transactions that are labeled as when-issued, locked-in, have special sales conditions, or have more than a two-day settlement period; (6) transaction records with trade size larger than issue size or trade size is not a integer; (7) bonds that do not have a principal value of $1,000; (8) bonds with incomplete issuance information (offering date, amount, and maturity) or non-positive historical amount outstanding (e.g., bonds are called); and (9) bonds that are not issued by public firms (i.e., with a valid PERMNO from CRSP).

B Daily Bond Sample Construction

In this section, we provide further details on the construction of the daily bond panel data used in Table 5.

After matching the daily Markit security lending data to the merged Mergent FISD-TRACE bond sample as specified in Section A, we obtain 18,458,551 observations for 18,085 corporate bonds issued by 1,755 firms from September 11, 2006 to December 30, 2022.

Next, we move to construct customer buy/sell volume. Enhanced TRACE records the direction of trades from the reporting dealers' perspective. Thus, for each customer-dealer

[15] Following Dick-Nielsen, Feldhütter, Pedersen, and Stolborg (2023), we define equity-linked bonds, as bonds whose field "issue name" contains any of the strings "EQUITYLINKED", "EQUITY LINKED", and "INDEX-LINKED".

trade, we treat dealer-buy trades as customer sales and dealer-sell trades as customer buys. We treat missing trading volume and customer buy/sell trade observations in TRACE as zero volume when computing bonds' transaction volume. To distinguish zero volume from missing observations, we first create empty panel data by setting the beginning and ending dates for an initial list of bonds in the sample, which is determined by the intersection of the three databases (TRACE, Mergent FISD, and Markit) we use. For the list of trading days, we use those in CRSP and exclude bond trades recorded on the days when stock markets are closed.[16] The beginning and ending dates for each bond are set by its issuance date and maturity date or the last call date. We then merge TRACE volume to the empty panel to determine which days have zero volume.

To obtain an estimate of transaction cost, we follow O'Hara and Zhou (2021) and match each customer buy/sell trade with the closest in time inter-dealer trade in that bond over the past five trading days. We construct half spreads for both the customer buy side and sell side as the volume-weighted average of the log price differences between customer buy/sell trades and inter-dealer trades. Moreover, to mitigate the microstructure noise, we compute daily bond returns and volatility using quote prices from the Bank of America Merrill Lynch (BAML) database provided by the Intercontinental Exchange (ICE).

The SEC announced on March 22, 2017, that the settlement cycle (i.e., the time between the transaction date and the settlement date) for most broker-dealer securities transactions will change from three business days (i.e., T+3) to two business days (i.e., T+2) on September 5, 2017. Thus, to account for these settlement gaps, we adjust trading volume, bond returns, and half spreads on day d by day $d-3$ values for the sample up to September 4, 2017 and by day $d-2$ values after September 5, 2017. Then, we compute return volatility and five-day moving averages based on the "adjusted" variables.

We require each bond to have at least 252 daily observations after merging daily bond lending data, trading volume, half spreads, and bond returns data. The final sample includes 10,693,324 bond-day observations for 11,411 corporate bonds across 1,259 firms from September 13, 2006 to December 30, 2022. We winsorize continuous variables at 1% and 99% by month to mitigate the effects of outliers while avoiding look-ahead bias. Table A1, Panel A reports the descriptive statistics for the daily bond panel data.

C Monthly Bond Sample Construction

In this section, we provide further details on the construction of the monthly bond panel data used in Figure 5.

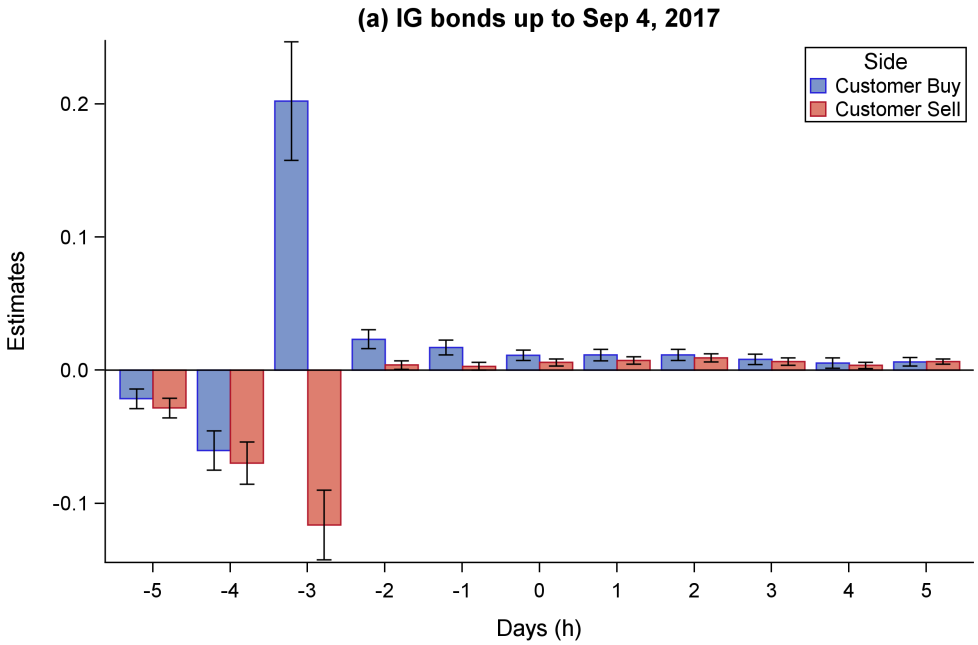
We start with the daily Markit bond lending data (after matching with the merged Mergent FISD-TRACE bond sample) and compute the monthly average of lending outcome variables by averaging the daily Markit data within each bond-month observation. Following Bretscher, Schmid, and Ye (2024b), we further exclude bonds that were issued less than 6 months ago. We obtain 815,719 observations for 17,214 corporate bonds issued by 1,718 firms over 196 months from September 2006 to December 2022.

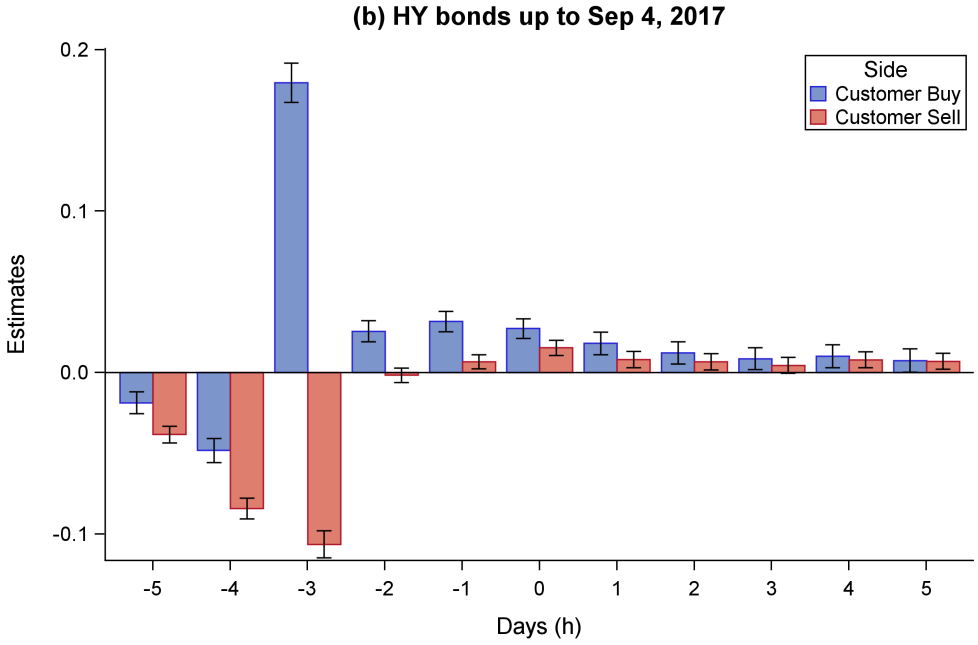
[16] This choice excludes some sparse trades on weekends but includes more trading days than Treasury market data.

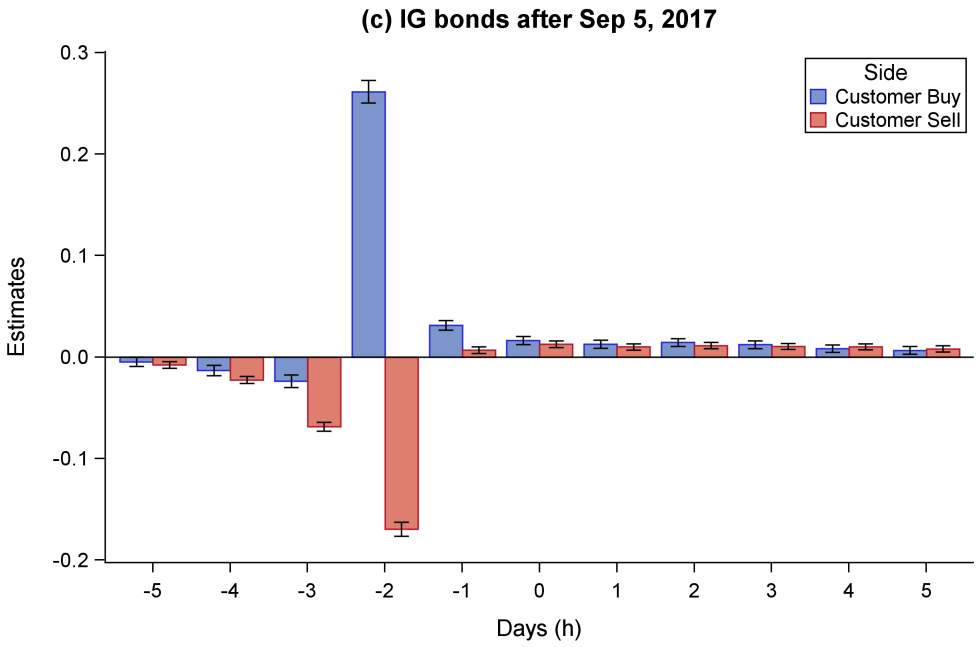
Next, we match the monthly bond lending data to the quarterly holdings data from eMAXX based on the bond CUSIPs and calendar quarters. We define a *Switch* indicator that equals one if a bond crosses one of the there cutoffs: 10-, 5-, and 3-year time to maturity. We compute the change of passive ownership and lending outcome variables from month $t-1$ to month $t-h$ and require all the outcome variable changes to be available for $h \in [-4, 24]$. These filters lead to a final sample of 311,378 bond-month observations for 9,668 corporate bonds across 1,181 firms from February 2007 to December 2022. We winsorize continuous variables at 1% and 99% by month. Table A1, Panel B reports the descriptive statistics for the monthly bond panel data.

Figure A1: Panel Regression of Dollar Trading Volume on Changes in Quantity on Loan, IG vs HY

The figure plots the slope coefficients of the panel regression of dealer-customer trading volume on day $d + h$ on day d changes in quantity on loan. Trading volume and quantity on loan are scaled by the amount outstanding. The y-axis represents a change in the percentage of the scaled dollar trading volume as a result of a one percentage change in the scaled quantity on loan. Subfigures (a) to (b) are for investment grade and high yield bonds up to Sep 4, 2017. Subfigures (c) to (d) are for investment grade and high yield bonds after Sep 5, 2017.






(c) IG bonds after Sep 5, 2017
Side
Customer Buy
Customer Sell
Estimates
0.3
0.2
0.1
0.0
-0.1
-0.2
-5
-4
-3
-2
-1
0
1
2
3
4
5
Days (h)

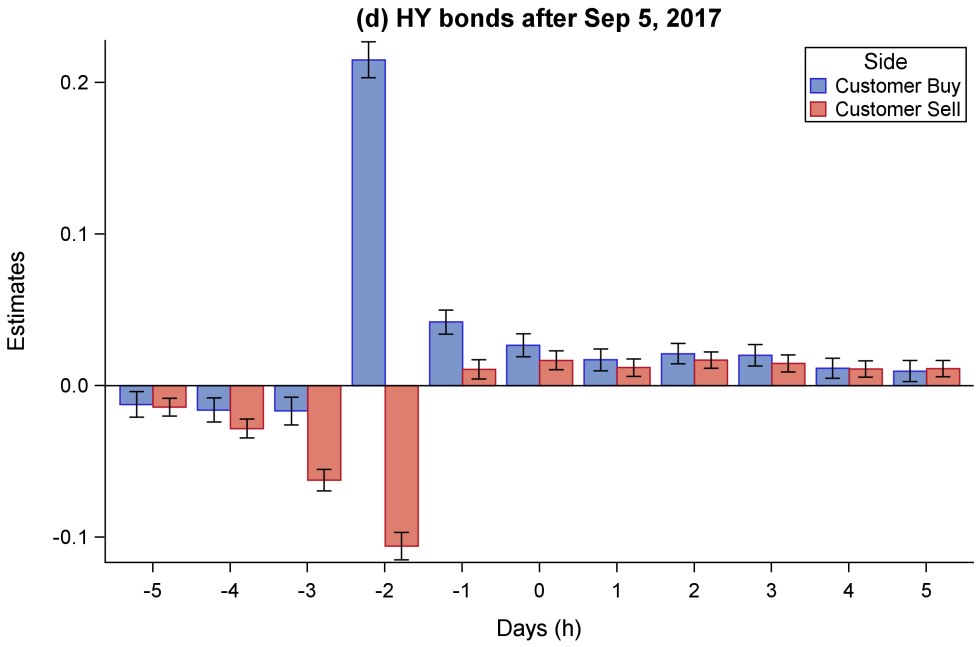

(d) HY bonds after Sep 5, 2017
Side
Customer Buy
Customer Sell
Estimates
0.2
0.1
0.0
-0.1
-5
-4
-3
-2
-1
0
1
2
3
4
5
Days (h)

Table A1: Descriptive Statistics of Daily and Monthly Panels

This table reports the summary statistics of the main variables at the bond-day and bond-month levels. The combined bond data are from Mergent FISD, TRACE, Markit, and eMAXX. The definitions of common bond characteristics are the same: *Amount* is the amount of bonds outstanding in millions of dollars; *Rating* is the numerical rating score, where 1 refers to a AAA rating by S&P and Aaa by Moody's, 21 refers to a C rating for both S&P and Moody's; *Age* is the age of a bond in years; *Maturity* is the time to maturity in years. Panel A shows the descriptive statistics of the sample used in Table 5, which includes 11,411 bonds across 1,259 firms from September 13, 2006 to December 30, 2022. dQ is the changes in quantity on loan scaled by the amount outstanding. The explanatory variables includes the daily return r_t, the average of the past 5 days' returns $r_{t-5,t-1}$, the turnover rate of customer buy ($Turn^{Buy}$) and sell ($Turn^{Buy}$) on day t and averages from $t-5, t-1$, half spreads for buy trades (h^{buy}) and sell trades (h^{sell}) averaged from day $t-5$ to $t-1$, and standard deviation of returns over the last five days $\sigma(r)_{t-5,t-1}$. Panel B shows the summary statistics of the sample used in Figure 5, which includes 9,668 bonds across 1,181 firms from February 2007 to December 2022. We compute monthly averages of daily bond lending variables for each bond unless mentioned otherwise. *Loan Quantity* is defined as the quantity on loan from Markit divided by amount outstanding. *Lendable Supply* is the active lendable quantity from Markit divided by amount outstanding. *Utilization Rate* is defined as the ratio of the quantity on loan to the lendable quantity. *Loan Tenure* is the average number of days that bond loans have been open. *Borrowing Fee* is the buy-side fee paid by the ultimate borrower ("*IndicativeFee*" in Markit). *Rebate Rate* is the "*IndicativeRebate*" in Markit. *DCBS* is the cost of borrow score provided by Markit, ranging from 1 (low cost) to 10 (high cost). *Fee Risk* is the natural logarithm of the standard deviation of borrowing fees within a month. *Recall Risk* is the natural logarithm of the standard deviation of utilization rate in a given month. *Lender Concentration* is a Herfindahl-Index-like measure at the bond level provided by Markit that describes the concentration of lenders. *Special1* is a dummy variable that equals one if at least one day with borrowing fee exceeds or equal to 1% in a month, and zero otherwise. *Special2* is a dummy variable that equals one in a given month if its borrowing fee is in the top decile of the fee distribution across bonds, and zero otherwise. *Credit Spread* is calculated as the average difference between the corporate bond yield and the yield of a matching Treasury bond within a month. h^{Buy} (h^{Sell}) is the monthly average of the log price differences between customer buy (sell) trades and inter-dealer trades following O'Hara and Zhou (2021). We match customer buy (sell) trades with the closest in time inter-dealer trade over the past five trading days with replacement. *Total Ownership* is the share of bonds held by all the investors in eMAXX. *Passive Fund*, *Active Fund*, and *Insurer* represent factions of bond par amount held by passive mutual funds, actively managed mutual funds, and insurance firms, respectively. We use the investor type classification code provided by eMAXX to group investors into insurance firms and mutual funds. We manually link eMAXX to the CRSP Mutual Fund database by matching funds based on their names and use index fund and ETF flags (as well as search keywords in fund names) to further decompose mutual fund ownership into *Passive Fund* (i.e., index funds and ETFs) and *Active Fund*. To mitigate the influence of outliers, we winsorize variables at 1% and 99% for each period.

Variable	Mean	SD	P1	P25	P50	P75	P99	Obs
Panel A: Daily Bond Panel								
dQ	−0.002	0.198	−0.719	−0.011	0.000	0.007	0.732	10,693,324
$dQ_{t-5,t-1}$	−0.001	0.093	−0.313	−0.016	0.000	0.013	0.322	10,693,324
r_t	0.016	0.765	−1.732	−0.143	0.017	0.181	1.759	10,693,324
$r_{t-5,t-1}$	0.017	0.351	−0.847	−0.056	0.016	0.099	0.838	10,693,324
$\sigma(r)$	0.421	0.645	0.018	0.138	0.274	0.513	2.399	10,693,324
Amount ($ mil)	875	640	250	500	700	1,000	3,250	10,693,324
Rating	8.464	3.205	1.000	6.000	8.000	10.000	17.000	10,693,324
Age (years)	4.003	3.476	0.173	1.521	3.115	5.540	17.674	10,693,324
Maturity (years)	8.667	7.824	1.151	3.515	5.926	9.219	29.425	10,693,324
$Turn^{Buy}$	0.189	0.442	0.000	0.001	0.026	0.145	2.320	10,693,324
$Turn^{Sell}$	0.117	0.330	0.000	0.000	0.004	0.049	1.818	10,693,324
$Turn^{Buy}_{t-5,t-1}$	0.208	0.297	0.002	0.033	0.099	0.255	1.468	10,693,324
$Turn^{Sell}_{t-5,t-1}$	0.133	0.215	0.000	0.011	0.046	0.161	1.068	10,693,324
$h^{Buy}_{t-5,t-1}$	0.390	0.770	−1.088	0.058	0.226	0.570	2.925	10,693,324
$h^{Sell}_{t-5,t-1}$	0.357	0.804	−1.190	0.048	0.211	0.520	3.013	10,693,324
Panel B: Monthly Bond Panel								
Loan Quantity (%)	1.555	2.599	0.008	0.192	0.593	1.691	12.376	311,378
Lendable Supply (%)	24.767	9.904	4.738	18.134	23.834	30.180	54.620	311,378
Utilization Rate (%)	6.879	11.950	0.035	0.850	2.624	7.281	66.203	311,378
Loan Tenure (days)	81.966	91.301	4.238	25.427	51.259	101.851	485.169	311,378
Borrowing Fee (%)	0.402	0.323	0.162	0.287	0.375	0.392	2.085	311,378
Rebate Rate (%)	0.495	1.214	−1.412	−0.249	−0.105	1.203	4.901	311,378
DCBS	1.035	0.207	1.000	1.000	1.000	1.000	2.000	311,378
Fee Risk	−3.016	0.959	−6.220	−3.552	−2.985	−2.645	−0.363	221,285
Recall Risk	−0.971	1.879	−8.357	−1.800	−0.682	0.233	2.106	307,892
Lender Concentration	0.451	0.301	0.000	0.264	0.444	0.659	1.000	311,378
$Turn^{Buy}$ *(%)*	0.144	0.170	0.000	0.034	0.090	0.189	0.846	311,378
$Turn^{Sell}$ *(%)*	0.095	0.119	0.000	0.018	0.055	0.126	0.586	311,378
h^{Buy}*(%)*	0.334	0.917	−1.631	0.013	0.201	0.540	3.258	287,492
h^{Sell}*(%)*	0.283	0.989	−2.069	−0.026	0.188	0.510	3.353	287,540
Total Ownership (%)	46.599	15.784	13.432	35.302	46.250	57.326	84.883	311,378
Insurer (%)	32.629	18.584	1.588	17.356	31.005	45.673	78.953	311,378
Passive Fund (%)	3.727	4.648	0.000	1.099	2.837	5.044	23.387	311,378
Active Fund (%)	9.938	11.040	0.000	2.197	5.914	13.479	48.597	311,378
Amount ($ mil)	788	608	150	400	563	1,000	3,000	311,378
Rating	8.195	2.946	2.000	6.000	8.000	9.500	16.500	311,378
Age (years)	4.476	3.656	0.962	1.956	3.488	5.633	19.115	311,378
Maturity (years)	11.313	8.537	3.082	5.110	7.384	18.345	29.099	311,378

Table A2: Sample List of Passive Funds in eMAXX

This table lists the top 25 passive funds in eMAXX as of 2022 in terms of the number of distinct corporate bonds held. We select passive funds based on index fund and ETF/ETN flags from the CRSP Mutual Fund database after manually matching eMAXX FUNDID to funds in CRSP by fund names. We further identify passive funds by searching keywords in fund names related to ETFs/index funds/bond index providers. We restrict corporate bond holdings to dollar bonds issued by US firms that have trade records in the Enhanced TRACE database.

Obs	FUNDID	FUNDNAME
1	170047	iShares Core Total USD Bond Market ETF
2	126137	iShares Broad USD Investment Grade Corporate Bond ETF
3	81464	iShares Core US Aggregate Bond ETF
4	189653	Vanguard USD Corporate Bond UCITS ETF
5	29844	Vanguard Total Bond Market Index Fund
6	136739	Vanguard Total Bond Market II Index Fund
7	156760	Schwab US Aggregate Bond ETF
8	29969	Vanguard Balanced Index Fund
9	191887	Schwab US Aggregate Bond Index Fund
10	142023	TIAACREF Bond Index Fund
11	29242	Vanguard Total Bond Market Index Portfolio
12	44906	U.S. Total Bond Index Master Portfolio
13	136745	SPDR Barclays Intermediate Term Corporate Bond ETF
14	126141	iShares Intermediate Government/Credit Bond ETF
15	153656	SPDR Barclays Issuer Scored Corporate Bond ETF
16	133663	LVIP SSgA Bond Index Fund
17	191555	iShares ESG Aware USD Corporate Bond ETF
18	126144	iShares Government/Credit Bond ETF
19	191556	iShares ESG Aware US Aggregate Bond ETF
20	195989	Vanguard Global Aggregate Bond UCITS ETF
21	174293	State Street Aggregate Bond Index Portfolio
22	144043	iShares 10+ Year Investment Grade Corporate Bond ETF
23	32593	EQ/Core Bond Index Portfolio
24	136743	SPDR Barclays Long Term Corporate Bond ETF
25	123766	iShares 1-3 Year Credit Bond ETF

Table A3: Passive Ownership and Other Lending Outcomes

This table presents the results from regressing other bond lending outcomes on ownership of institutional investors. The dependent variables are quarterly averages of utilization rate, loan tenure, fee risk, recall risk, and half spreads. *Passive Fund*, *Active Fund*, and *Insurer* represent factions of bond par amount held by passive mutual funds, actively managed mutual funds, and insurance firms, respectively. Bond control variables include the log value of amount outstanding, rating, time to maturity, and the fraction of zero-trading days. The variable definitions can be found in Table 1. All continuous independent variables are standardized to have a mean of zero and a standard deviation of one. We include bond and firm × quarter effects in each regression. We double cluster standard errors by firm and year-quarter, and t-statistics are in parentheses. *, **, and *** indicate the significance at the 10%, 5%, and 1% levels, respectively. The sample period is from 2006 Q3 to 2022 Q4.

	Utilization	Loan Tenure	Fee Risk	Recall Risk	h^{Buy}	h^{Sell}
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Passive Funds Only						
Passive Fund	-0.257^{***}	-0.214	0.025^{***}	-0.026^{***}	0.011^{**}	-0.007
	(-3.95)	(-0.74)	(4.12)	(-3.37)	(2.59)	(-1.53)
Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes
Firm×Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes
Bond FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	280,330	280,330	236,392	274,079	268,866	267,494
Adjusted R^2	0.629	0.381	0.353	0.502	0.259	0.272
Panel B: Passive Funds Plus Other Investors						
Passive Fund	-0.270^{***}	-0.221	0.025^{***}	-0.026^{***}	0.011^{**}	-0.008
	(-4.12)	(-0.74)	(4.10)	(-3.35)	(2.63)	(-1.58)
Active Fund	0.985^{***}	0.575	-0.010	0.041^{***}	-0.026^{***}	0.016^{**}
	(7.11)	(0.86)	(-0.94)	(2.81)	(-4.22)	(2.31)
Insurer	0.638^{***}	10.778^{***}	-0.029^{*}	-0.129^{***}	0.065^{***}	0.038^{***}
	(3.32)	(6.10)	(-1.75)	(-4.78)	(6.38)	(3.25)
Bond Controls	Yes	Yes	Yes	Yes	Yes	Yes
Firm×Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes
Bond FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	280,330	280,330	236,392	274,079	268,866	267,494
Adjusted R^2	0.631	0.381	0.353	0.502	0.259	0.272

Specialization in Financial Markets*

Milena Wittwer[⋆], and Andreas Uthemann[+]

May 20, 2025

Abstract

Intermediary asset pricing models typically assume integrated markets with homogeneous assets, akin to Walrasian auctions. In practice, however, markets are fragmented and financial products are diverse. Using a unique dataset linking intermediaries' trades across Canadian stock, bond, and derivative markets, we examine where, what, and at what prices intermediaries trade to assess whether market fragmentation, product diversity, and the resulting specialization are empirically relevant features that asset pricing models should incorporate. We document substantial specialization: intermediaries concentrate their trading activity unevenly across markets and products, with product specialization more pronounced than market specialization. Furthermore, more specialized intermediaries consistently obtain better prices. These findings highlight the need for asset pricing models that account for specialization—especially across products.

Keywords: Market segmentation, specialization, financial intermediaries, market design, asset prices
JEL: G00, G10, G12, G19, D40

*The views presented are those of the authors and not necessarily the Bank of Canada. We thank Mirela Sandulescu for an excellent discussion. We also thank Jason Allen, Vincent Vincent Bogousslavsky, Adi Sunderam, Jonathan Wallen, and the experts at the Bank of Canada and the Montreal Stock Exchange—in particular, Corey Garriott and the Triton project team. We also thank Connor Breck, Carel Chok, Costanza Didonna, Matthew Hagerty, Harrison Lynch, Brendan McLaughlin, Daniel Smith, and Joseph Wagner for excellent research assistance. Any errors are our own. Correspondence to: [§]Milena Wittwer (Boston College): wittwer@bc.edu, and [+]Andreas Uthemann (Bank of Canada): authemann@bankofcanada.ca

1 Introduction

Intermediary asset pricing suggests that frictions faced by financial intermediaries can constrain arbitrage and influence asset prices (e.g., Shleifer and Vishny (1997); Gromb and Vayanos (2002); Brunnermeier and Pedersen (2009); He and Krishnamurthy (2013); Brunnermeier and Sannikov (2014)). Most models assume frictionless markets, akin to a Walrasian auction. Consequently, the related empirical literature largely overlooks frictions arising from market fragmentation (e.g., Pasquariello (2014); Adrian et al. (2014); Du et al. (2018); He et al. (2017); Siriwardane et al. (2022)). This contrasts with the fragmented nature of financial markets, where different asset classes trade in distinct venues (Malamud and Rostek (2017); Weill (2020); Chen and Duffie (2021); Budish et al. (2024)). Moreover, since financial assets are treated as homogenous, the literature tends to overlook the diversity of financial products and, consequently, the role of product specialization (Babus and Hachem (2023); Babus et al. (2024); Mota and Siani (2024)).

We introduce a unique dataset to study the role of market and product 'segmentation'—or equivalently, 'specialization'—in financial markets to provide novel stylized facts that inform future asset pricing models. By linking trades across Canadian stock, bond, and derivative markets, we analyze cross-market and cross-product specialization by examining where, what and at what prices brokers and dealers trade, offering insights into the returns to specialization.[1] Cross-market specialization may arise from differences in market clearing rules or entry costs, while cross-product specialization within a market—where such frictions are absent—may instead reflect differences in trading expertise, relationships, or client preferences.

Our dataset covers all trades executed on Canada's fixed-income market and all exchanges owned by the Toronto Stock Exchange Group (TMX) from 2019 to 2022. TMX owns three stock exchanges, which account for roughly 60 percent of equity trade volume in Canada, and the country's only derivatives exchange. A key feature of this dataset is the ability to track dealers over time and across markets using legal entity identifiers (LEIs)—an attribute rarely available in trade-level datasets, particularly for stocks and derivatives. We link this information to public data in order to classify securities into products, for instance, corporate bonds, large-cap stocks, Exchange Traded Funds (ETFs), and Treasury futures. Additionally, we manually assign dealers to their parent institutions, i.e., the LEIs of their holding companies, and categorize them by

[1] An alternative label for the institutions analyzed in this paper is 'market-maker.' We avoid this term because the small set of financial institutions that dominate trading—the set we study in this paper—engage in activities beyond market-making, including executing trades for clients and responding to client needs. For instance, on exchanges, only a small set of firms are formally designated as market makers with an obligation to provide liquidity.

type, such as primary dealers or hedge funds.

Using these data, we establish three stylized facts about specialization. The first demonstrates its existence and quantifies its extent: dealers allocate their trading activity unevenly across markets and products—they specialize.

To quantify market specialization, we construct dealer-specific market specialization scores, which range from zero (no trades in a market) to one (exclusive trading within it). Banks tend to concentrate in bonds, high-frequency traders in derivatives, and primary dealers in government debt are the most active across markets. A similar pattern emerges within markets: trading is unevenly distributed across product segments, reflecting product specialization. We capture this using a dealer-specific product specialization score, defined as the ratio of a dealer's trade share in a product segment (relative to all dealers) to the sum of these trade shares across products in a market. Like the market score, it ranges from zero to one.

Product specialization appears shaped by both market structure—centralized versus decentralized—and product complexity. On centralized stock exchanges with standardized products, dealers typically trade broadly. In contrast, specialization is stronger in the decentralized fixed-income over-the-counter (OTC) market, where search and relationship frictions may push dealers to focus on specific bond types. Yet these frictions cannot fully explain specialization: it also arises in the centralized derivatives market, which, like stock exchanges, operates via an anonymous limit order book, but features less standarized products.

Across markets, product specialization is more pronounce than market specialization—this is our third fact. To derive it we introduce a specialization index that integrates market and product specialization scores, and allows us to decompose within-market from cross-market specialization for each dealer. For this, we adapt the Theil (1967) index to account for the fact that not all dealers participate in every market segment. While commonly used to measure inequality in socio-economic contexts (e.g., Anand and Segal (2015)), the Theil index has not, to our knowledge, been applied to trade settings.

The decomposition shows that, for most dealers, product specialization within a market is greater than market specialization. This finding suggests that, for large financial institutions, barriers to market entry are less restrictive than factors that limit trading across products within the same market. As a result, policies aimed at moderately changing entry costs or membership fees—such as the recently revised fee schedules for registered broker-dealers in the U.S. and Canada (FINRA (2024); CIRO (2024))—may have limited impact on market participation.

Next, we examine whether market and product specialization affect transaction prices, aiming to establish our third, and final, stylized fact that specialized dealers trade at better prices. We focus on relative prices across dealers, not on how specialization affects aggregate price

levels. Dealer specialization could influence transaction prices by improving inventory management, or shaping beliefs about fundamentals. However, in a frictionless and competitive market, any price effects from specialization would be arbitraged away.

We, therefore, begin by showing that none of the markets is sufficiently frictionless to prevent some dealers to outperform others. We measure a trade's margin as its price advantage relative to the average price at which the same security trades on that day.[2] A margin of 1% indicates that the dealer pays 1% less than the daily average when buying (and sells at 1% more when selling). We show that dealers systematically differ in the prices they obtain, across all markets, even after controlling for trade size, security-time fixed effects, and other observables. High-frequency traders tend to outperform others on exchanges, while retail-facing brokers underperform in the bond and derivatives markets.

Having established that there is scope for price effects, we investigate whether the successful dealers trade across products or markets or whether they specialize. We show that some dealers who trade exclusively within a single market outperform those who trade across markets in the bond and derivatives market, but not the stock market. This is in line with the idea that a decentralized market structure and product complexity promote specialization. Across markets, dealers who consistently secure better prices for bonds do not achieve better prices in stocks or derivatives, and vice versa, suggesting limited trading synergies across markets and products, and reinforcing the role of specialization.[3]

Finally, we exploit cross-sectional variation in dealer specialization to show that more specialized dealers obtain better prices. To address concerns about reverse causality—where more successful dealers become more specialized—and omitted variables, such as dealer sophistication or efficiency, we implement two strategies. First, we relate lagged specialization scores to current trade margins. The idea is that last year's specialization is less likely to be influenced by, or directly affect, current prices. Second, we use an instrumental variables (IV) approach. While identifying exogenous variation in trading is notoriously difficult, our data offer a unique opportunity: we can distinguish whether a stock market trade is for the dealer's own account or for a client. Client orders serve as plausibly exogenous shocks for dealer own-account trades,

[2]Our approach follows the market microstructure literature, which commonly defines transaction costs as the trade price relative to a benchmark (e.g., Hendershott and Madhavan (2015); Hau et al. (2021); O'Hara and Zhou (2021); Pinter et al. (2024)). Ideally, one would compare the trade price to a mid-price or fundamental value, but data constraints prevent this. Instead, we use the average daily price, which we can construct consistently across markets. Our regressions control for security-week fixed effects to account for differences across securities and over time.

[3]This may reflect the tendency of individual traders or desks within institutions to focus on a narrow set of assets, optimizing trading within their domain (Lu and Wallen (2024)).

once observable trade characteristics and a rich set of fixed effects are controlled for.

While neither approach fully eliminates endogeneity concerns, they reveal a consistent pattern when taken together: more specialized dealers trade at better prices. The price effect is moderate at the trade level but becomes economically meaningful when aggregated over time. For example, on the stock exchange, moving from no market (product) specialization to full specialization increases margins by 4–39 (28–41) basis points per trade.

Taken together, our findings underscore the importance of market and product specialization in intermediary asset pricing—two dimensions largely overlooked in existing models. They call for asset pricing frameworks that incorporate specialization, particularly across asset classes, and point to several promising directions for future theoretical and empirical work.

One direction for future research is to analyze whether, and if so how, specialization shapes market-outcomes, including aggregate price levels. Doing so requires either large exogenous variation in specialization across several dealers or a structural framework—for example, by extending Vayanos and Vila (2021) to accommodate richer market structures and product complexity. Such analysis could lay the groundwork for a broader research agenda examining whether—and through what mechanisms—granular frictions observed in micro-level data aggregate into distortions in market equilibrium outcomes. It would also complement the empirical literature asset pricing literature (e.g., Pasquariello (2014); Adrian et al. (2014); Du et al. (2018); He et al. (2017); Siriwardane et al. (2022)), which typically relies on market-level data to capture broad effects but offers limited visibility into underlying mechanisms—aside from a few exceptions (e.g., Siriwardane (2019); Wittwer and Allen (2023)).

In line with the goal of unpacking mechanisms, another direction for future research is to examine why and how product complexity and market structure shape product specialization—as suggested by the contrast between product specialization on the stock market versus the OTC market and the derivatives exchange. This would require models that account for different market structures and product characteristics, shifting the focus beyond a single market or product, as seen in much of the existing literature. With few exceptions—such as Dougast et al. (2022) and studies on derivatives and their underlying assets dating back to Kumar and Seppi (1992)—existing models predominantly focus on a single market structure and asset class.

Another aspect for future research would be to explore how trading interconnectedness evolves during periods of distress. Our data shows that large banks dominate trading across markets, which raises concerns about financial stability. From a policy perspective, this implies that regulatory changes affecting dealer bank balance sheets—such as adjustments to the supplementary leverage ratio to accommodate increased government debt issuance—will impact

all markets. Investigating which types of institutions amplify negative spillovers and which help mitigate them would contribute to the extensive literature on contagion, following Allen and Gale (2000).

Finally, our cross-market and multi-asset perspective highlights the need for empirical market microstructure studies (contributing to a large literature, including Hasbrouck and Sofianos (1993); O'Hara (2015); Menkveld (2016); Bessembinder et al. (2020)), and the growing literature on demand estimation (following Koijen and Yogo (2019)) to move beyond isolated markets or individual products within a market. Most existing studies in both literatures focus on a narrow set of assets, such as a single bond type or common equity, and largely constrain substitution across asset classes, limiting the potential for spillover effects. Allen et al. (2020), Chaudhary et al. (2022), Üslü and Pintér (2023), Allen and Wittwer (2024), and Dix and Wittwer (2025) take initial steps in this direction, but given the empirical patterns documented in this study, much remains to be explored.

Similarly, though more distantly related, the extensive asset pricing literature on factor structures, including common stochastic discount factors, has traditionally focused on asset-class-specific factors, but has begun shifting toward identifying joint factors that span multiple asset classes (e.g., Sandulescu (2020); Chen et al. (2024)). Sandulescu (2020), for example, documents significant integration between U.S. corporate bonds and equities, consistent with the empirical patterns we observe for Canada.

2 Institutional environment

Before detailing the construction of the dataset we use to examine dealer specialization, it is useful to review the key market features. The structure of Canadian financial markets closely mirrors that of other developed nations, including the United States. The three primary asset classes—bonds, stocks, and derivatives—each operate in separate markets.

Fixed-income market. Fixed-income instruments are issued in primary markets and traded in decentralized over-the-counter (OTC) markets. In traditional OTC markets, buyers must contact sellers individually to conduct bilateral trades. Consequently, these markets largely depend on large financial institutions—dealers—to intermediate between investors, such as firms, public entities, and individuals. Although not all trade occurs bilaterally today, the market remains fragmented.

Firms seeking to become fixed-income dealers must apply to the Canadian Investment Regulatory Organization (CIRO). CIRO membership is available to Canadian entities registered to

Table 1: Products in the fixed-income market

Product	Trade share
Government Bonds and Bills	63.44
Provincial, Municipal Bonds and Bills	9.33
Bankers' Acceptances	8.81
Bank, Agency Papers	7.96
Corporate Bonds	6.14
ABS, MBS, CMB	4.88
Strips	0.27

Notes: Table 1 shows the daily average share of total trade-volume, computed as the total amount of bonds (in terms of par value) traded on a day, in the bond market per product. ABS are Asset-Backed Securities, MBS are Mortgage-Backed Securities, and CMB are Canada Mortgage Bonds. Appendix Table A1 describes each product category.

operate as dealers or advisors in any province or jurisdiction in Canada. CIRO members must satisfy CIRO's financial and operations compliance, business conduct compliance and registration requirements, including minimal capital requirements (typically C$250,000), and pay annual membership fees (CIRO website).

Fixed-income securities range from long-term bonds to short-term money-market instruments. We classify all securities into product categories, as explained in Appendix Table A1. Government bills and bonds are traded the most, as shown in Table 1. Then we have provincial and municipal debt, and Bankers' Acceptance (which is a money market instrument that is issued by a business and guaranteed by a bank), bank or agency papers (money market instruments issued by banks or agencies), and corporate debt. Mortgage- or asset-backed securities (ABS, MBS, CMB), and strip bonds (which are debt instruments in which both the principal and regular coupon payments, that have been removed, are sold separately) are relatively small.

Equity market. Equity products are in most countries traded on centralized exchanges. Exchanges differ from OTC markets in that the market clears centrally on a limit order book.

In Canada there are nine exchanges.[4] Our focus lies on exchanges that are owned by the TMX group, given our data. The TMX group owns three exchanges: Toronto Stock Exchange (TSX), TSX Venture Exchange (TSXV), and TSX Alpha Exchange (Alpha). In our sample period, 2019-2022, about 58% of the total volume traded, and 63% of the total dollar value traded in an average month on any of the Canadian equity markets was traded on a TMX exchange.[5]

[4]The nine exchanges are: NEO Exchange Inc., Canadian Securities Exchange (CSE), Instinet Canada Cross Limited (ICX), Liquidnet Canada Inc. (Liquidnet), Nasdaq CXC Limited (Nasdaq Canada), Tradelogiq Markets Inc. (TMI), TMX Group (TSX, TSXV, TSX Alpha), TriAct Canada Marketplace (Match Now).

[5]These numbers are computed with data from CIRO, accessible here: https://

Table 2: Stock market products

Product	Trade share
Small Stock	53.58
Large Stock	32.33
Uncommon Shares	7.41
Exchange Traded Funds	5.98
Other or Missing	0.68

Notes: Table 2 shows the daily average share of total trade-volume on the stock market, computed as the total amount of stocks (i.e., the total number of shares) traded on a day, per product. Appendix Table A2 describes each product category.

Only exchange members can place orders for their own account, or on behalf of non-exchange members, i.e., their clients. To become a TMX exchange member, a firm must be a member of a self-regulatory organization (CIRO in Canada), have a CDS clearing agreement, and establish electronic access to TSX and/or TSX Venture Trading Engine.[6] In addition, a firm must pay an entry cost, which is relatively high for members who seek to be eligible to trade, roughly C$65,000. To keep the membership status, each exchange member must pay a monthly membership fee (in 2023 $1,500), in addition to trading fees, which are explained on TMX's website.

Exchange members can trade a variety of products, ranging from common stocks, and ETFs to more specialized products, such Exchange Traded Receipts (which let investors own gold bullion stored in the Royal Canadian Mint Gold Reserves). We group products in five categories, as explained in Appendix Table A2: large and small common company stocks, ETFs, non-common shares, and other/missing.

Derivatives market. Derivatives are traded over-the-counter, or on exchanges. We focus on exchange-traded derivatives. In Canada, there is a single derivative exchange, the Montreal Exchange (MX). It is owned by TMX group, and operates similarly to the other TMX exchanges.

To trade on the MX, a firm must become an MX exchange member. The requirements are similar to those for TMX. In particular, each MX member must be a CIRO member if the firm is Canadian and a member of the analogue regulatory entity of their nationality otherwise.[7]

`www.iiroc.ca/sections/markets/reports-statistics-and-other-information/reports-market-share-marketplace`, accessed on 08/10/2023.

[6] If the firm does not have a CDS clearing agreement, it must have a relationship with a clearing facilitator. For more details, see TMX's website.

[7] More specifically, requirements are different for Canadian and foreign firms. Canadian firms must be member of a Canadian self-regulatory organization (Investment Industry Regulatory Organization of Canada); must be a member of the Canadian Derivatives Clearing Corporation or conclude a clearing agreement with one of its members (MX). Foreign firms must be located in one of the following juris-

Table 3: Derivative products

Product	Trade share
Treasury Futures	37.44
Equity Options and Share Futures	27.77
Short Rate Derivatives	20.75
Bundles and Spreads	7.85
Index Options and Futures	6.51
Currency Options	0.04

Notes: Table 3 shows the daily average share of total trade-volume on the derivatives market, computed as the total amount of derivative contracts traded on a day, per product. Note that the amount of contracts does not reflect the value of the underlying assets. Appendix Table A3 describes each product category.

Members also have to pay MX-specific monthly membership fees and trading fees.

We group the derivative products into categories, closely following the MX website, as explain in Appendix Table A3. The largest category in terms of trade volume are Treasury futures, followed by equity options and share futures, and short rate derivatives (as shown Table 3). Trading activity in currency options is negligible.

Some derivative products, like Treasury futures or index futures, are highly standardized, while others are more complex. One example are 'user-defined-strategies' (UDS), which allow participants to create customized option strategies based on their individual risk management needs. We classify them under bundles and spreads since UDS tend to combine multiple derivative contracts. Even equity options and share futures are more complex than common stocks, because traders can specify the maturity and strike price, in addition to the underlying asset.

3 Data

We combine different data sources on five market segments, TSX, TSXV, Alpha, MX and the fixed-income market. These five market segments represent three markets: stock market (TSX, TSXV, Alpha), derivatives market, and the fixed-income market. The main data sources that allow us to observe trade information are proprietary to the TMX group and CIRO. We hand-collect publicly available information on CIRO and exchange members, financial products, and market conditions to enrich the data.

diction: United States, United Kingdom, Republic of Ireland, Israel, Jersey, the Netherlands and France; must be duly formed pursuant to the relevant laws of the country; must be registered with a securities or derivative instruments regulator, or a recognized self-regulatory organization, unless it is exempted from such registration in its jurisdiction and subject to all other applicable restriction; must have entered into a clearing agreement with a member of the Canadian Derivatives Clearing Corporation; must have a designated agent for service of process residing in Quebec (MX).

Fixed-income market. Our main data source for the fixed-income market is the Debt Securities Transaction Reporting System, MTRS2.0. This data base stores trades that involve at least one CIRO Member (who have an obligation to report all of their trades) since November 2015.[8] Our sample covers trades with all Canadian fixed-income products from 2019 until 2022. Trades between two institutions or individuals who are not CIRO Members are not reported. According to market experts, however, these trades are rare.

For each transaction we see which security is traded, and a series of security-characteristics which allows us to classify securities into product categories and assign industry sectors. We also observe the quantity and price of the trade, the time at which the trade is reported, and the side of the trade (buy/sell).

A rare feature of the MTRS2.0 data relative to most of the existing datasets that cover OTC markets is that most firms carry a unique identifier. In this study, we focus on CIRO dealers. Traders who act as dealers in the primary market have to report their own trades with their legal identifiers (LEIs). Other CIRO dealers are allowed (but not obligated) to mask their identity when they are reporting their own trades, but not when they are reported as counterparty (with LEIs). Given that most trades occur with at least one party acting as a primary dealer, masked identifiers are infrequent—roughly 5% of trades and 1% of trade volume.[9] Since we are interested in how much dealers buy and sell, we stack buy- and sell-side trades, and remove sales or purchases by non-dealers.

Equity and derivatives market. We observe trade-level data for all exchanges that are owned by the TMX group between 2019 and 2022. For each trade, we observe the time of the trade (up to milliseconds), the security (i.e., the TMX symbol), the amount, the price, and trading-firm IDs. For equities, we also see the best national bid and ask offer for each symbol that was valid right before each trade executes. Moreover, we know whether the trade is for the exchange member's own account or a client account. More specifically, for stock market trades, we can distinguish between an inventory account (IN), a client account (CL), and an account that members who are designated market makers use for their market making active (ST). Although there are a few other types of accounts, these are negligible. For trades in the derivatives

[8] A small group of Bank of Canada staff have access to the raw data, and this is anonymized before it can be shared, subject to a non-disclosure agreement, to external researchers.

[9] Whenever a masked dealer trades with a primary dealers or government distributor, it is possible to back out the identify of the masked dealer by relying on the fact that both dealers need to report the trade. The remaining trades by masked dealers are those between two masked dealers. In these cases, we cannot rule out the possibility that our data sample includes both sides of the trade due to double reporting. In all other cases, Bank of Canada staff has carefully removed one of the trade sides, so that each trade appears only once in our data set.

market, we observe analogous account-types.

To account for both sides of each transaction, we stack buy- and sell-side trades. Moreover, in order to link dealers across markets, we link the market-specific trading-IDs to each company's LEI. We achieve this by downloading exchange member lists containing trading IDs and company names for all instances where the TMX website was archived by the Wayback Machine during our sample period. We then identify each company's LEI using https://www.lei-lookup.com. Doing so, we account for mergers, acquisitions, and name changes over time.

Lastly, to categorize securities into products, we merge the data on stock market trades with publicly available listing information for each listed symbol in December of each year in our sample, relying on the wayback machine. Finally, to validate data quality, we verify that we observe the same daily trade volume and nominal value as CIRO and the MX exchange report publicly for the stock markets and the derivatives exchange. See Appendix A for details.

Dealers. A distinctive feature of our data is the ability to track firms registered as CIRO dealer members in the fixed-income market or as exchange members across these markets. Throughout the paper, we refer to these traders as 'dealers.' It is important to note that we do not classify firms as 'dealers' based on their trading or market-making activities. Instead, our definition relies solely on firms' membership status, which grants them the ability to place trade orders on their own behalf on exchanges and to trade with clients in OTC markets, regardless of their specific role in the market. We adopt this definition because it is more exogenous—at least conditional on market entry—than classifications based on endogenous trading behavior.

For each dealer LEI, we identify the LEI of its holding company parent using information from gleif.org. Doing so, we manually track mergers, acquisitions, and name changes found through Google searches to the best of our ability.

Further, we classify all LEIs and their parents into institution types: Table 4 shows that brokers constitute the largest category at the parent level, representing financial entities primarily engaged in brokerage services. Following them are asset managers and high-frequency traders, which include hedge funds, proprietary trading firms, and private equity firms. Investment banks come next, followed by other, typically smaller, banks and credit unions. At the LEI level, the dataset also includes some mutual funds and retail branches of larger institutions, such as banks, that focus on retail investing.

Summary statistics. Appendix Table A5 summarizes our trade data for stocks (TSX, TSXV, Alpha), bonds (MTRS), and derivatives (MX) to provide an overview of a typical trading day and trade in each market. The bond market is the largest in terms of trade volume. The number

Table 4: Dealer types

Institution type	# of LEIs	# of Parents
Asset Manager	15	21
Bank	7	17
Investment Bank	12	15
Broker	115	65
High-Frequency Trader	17	21
Mutual Fund	8	0
Pension Fund and Insurance	4	3
Retail	13	3
Other	3	2

Notes: Table 4 shows the number of LEIs and parent-LEIs of each dealer type at the LEI and parent-level. For the type classification we follow the methodology of the Bank of Canada used to classify institutions into types for the MTRS 2.0 data (explained in Appendix A). Appendix Table A4 defines each type category we observe in our data.

of dealers who actively trades on an average day is similar across markets, ranging from 48 in the derivatives market to 60 in the stock market.[10] For derivatives, trade size and volume reflect the number of contracts, not the underlying asset value. Similarly, the trade price reflects the price payed to exchange the derivative, i.e., the option fee in case the contract is an option, not the strike price.

4 Dealer specialization

We examine two types of specialization: market specialization, which may arise from differences in market-clearing rules and entry costs, and product specialization within a market, which could stem from variations in trading expertise or client relationships.

To preview, we will establish our first stylized fact in what follows:

Fact 1 (Specialization)**.** *Dealer trading is uneven both across markets and across products within a market—dealers specialize.*

To quantify specialization, we assign each dealer to their holding company. This uniformly removes any type of in-house segmentation for all dealers, which we know to play a role (as shown by Siriwardane (2019), and Lu and Wallen (2024), among others). As a result, our specialization measures are conservative and likely underestimates

[10]When the level of aggregation for institutions changes, the set of players adjusts slightly. This occurs because an entity identified by its LEI may not be classified as a dealer, whereas its parent institution might be.

Additionally, we distinguish between markets rather than market segments, treating TSX, TSXV, and Alpha as a single market. These exchanges are highly integrated, likely due to common ownership and structural similarities. Nearly all trade volume is executed by dealers active across all three exchanges (see Appendix Table A6).

Market specialization. Market specialization might arise because of different market structures. Bonds trade in a decentralized market that is not directly connected to stock markets, or the derivatives exchange. Moreover, each market is characterized by different entry costs.

Supporting the idea that frictions, or market-specific preferences, might hinder universal market participation, Figure 1a highlights that not all parent LEIs (the y-axis) participate in all markets (the x-axis). If a dealer trades at least once in a given market, we plot a black line for that market, otherwise, the line is white. Thus, if all dealers participated in every market, the graph would be entirely black. Instead, the presence of both black and white indicates that dealers do not participate in all markets.

Market activity, measured in trade volume, is also uneven across markets. This is evident for the 20 largest dealers in Figure 1b. If dealers maintained similar market shares—defined as their fraction of total trade volume in each market—the horizontal bars would have consistent color shading, with lighter shades indicating larger market shares.

Including all dealers, Figure 2 visualizes the pairwise correlation of market shares across markets.[11] We see that most dealers concentrate their trading in specific markets; if market shares were uniform across markets, the points would align along the 45-degree line. Furthermore, smaller dealers tend to specialize more strongly, as indicated by points near the axes, suggesting they transact almost exclusively in one market. Among larger dealers, a subset focuses more heavily on the bond market, primarily banks that act as primary dealers (see Appendix Figure A1a).

To control for the overall market size of a dealer, we introduce a market specialization score, which ranges between 0 (the dealer does not trade in the market under consideration) and 1 (the dealer only trades in the market). Formally, the score divides dealer j's market share in market m and year y, s_{yjm}, by the sum of the dealer's market shares in all markets:

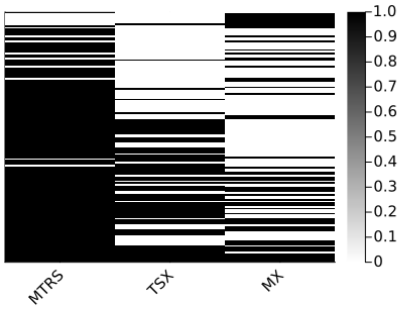
$$\text{specialization}_{yjm} = \frac{s_{yjm}}{\sum_m s_{yjm}} \in [0, 1]. \tag{1}$$

Figure 3 shows the market specialization scores for all dealers, averaged over the years. Many

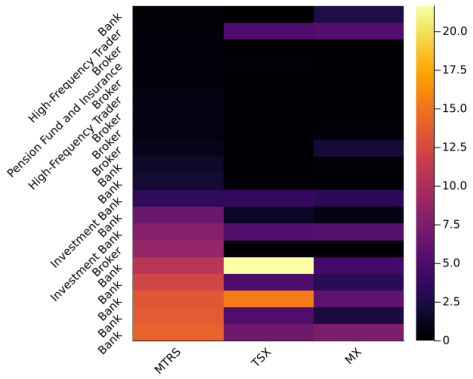
[11] Appendix Figures A2 and A3 show similar correlation patterns when considering dealers at the LEI-level, and when excluding trades for client accounts.

Figure 1: Dealer presence and market shares per market

(a) Dealer presence



(b) Market shares



Notes: Figure 1a shows dealer presence in each market—bonds (MTRS), stocks (TSX), and derivatives (MX)—with black indicating that a dealer has traded at least once in the respective market. Figure 1b displays the average annual market shares of the 20 dealers with the highest average annual trade volume in any market. In both figures, each row represents a dealer, sorted by total trade volume across all markets. Dealers with the highest overall trade volume appear at the bottom, while those with the lowest appear at the top. Since the fixed-income market is the largest by trade volume, this sorting places dealers with relatively low participation in fixed income—but significant presence in other markets—toward the top.

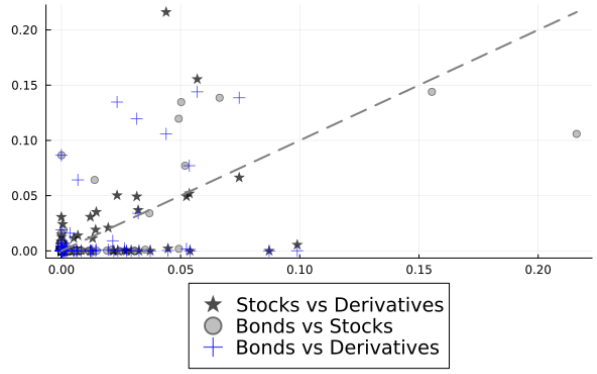
dealers do not trade in all markets—their scores align along the x-or y-axes and on the diagonal that connects the 1 on the x-axis with the 1 on the y-axis. Some only trade in one market and therefore have a specialization score of 1.

Among the three markets, the derivatives market stands out as the most detached, with some dealers—particularly high-frequency traders—engaging almost exclusively in derivatives, even when their overall trading volume exceeds 5% (see Appendix Figure A1b). In total, approximately 35% of MX's trade volume is attributed to high-frequency traders when including client trades, rising to 72% when excluding them, as shown in Appendix Table A7. Around 25% of traders operate exclusively on MX, even at the parent-company level. This is largely driven by hedge funds, proprietary trading firms, and private equity firms that specialize in MX trading.

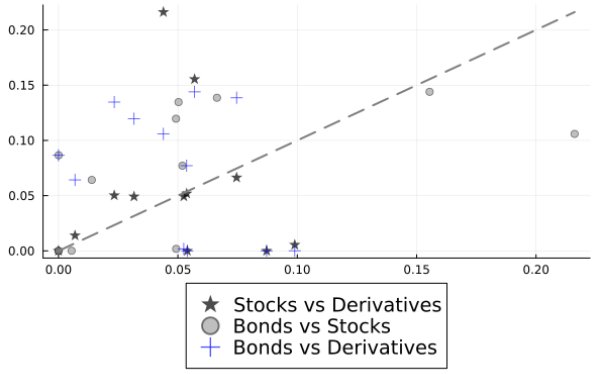
Banks acting as primary dealers are the most active dealers in linking markets (see Appendix Figures A4 and A5). In the bond market, nearly all dealers trading across all markets are primary dealers. On the stock market, primary dealers active in all markets account for approximately 68% of total trade volume. In the derivatives market, their share is lower at around 40%, yet no other dealer type plays a larger role in connecting derivatives with the other two markets. This underscores the significance of primary dealers beyond fixed income and highlights potential contagion risks during financial distress.

Figure 2: Dealer market shares, $s_{yjm} \in [0, 1]$, in an average year

(a) All dealers



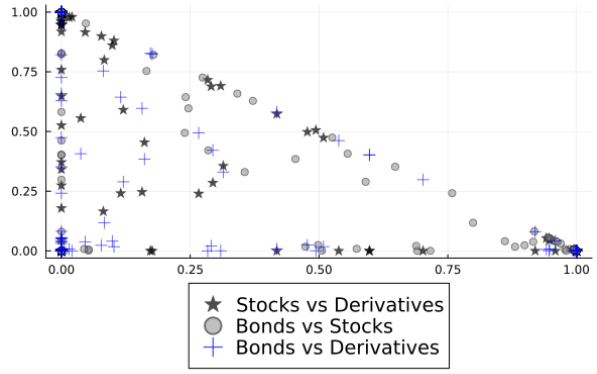
(b) Largest dealers



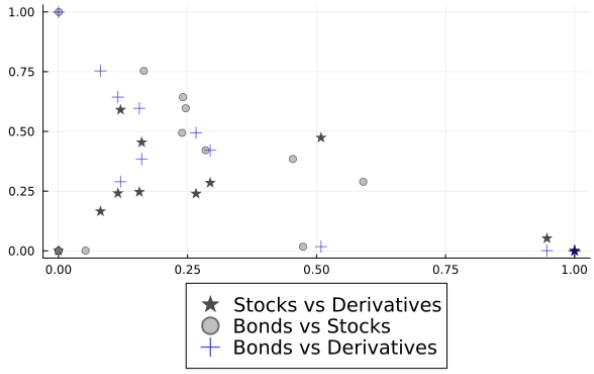
Notes: Figure 2a plots all dealer j's market shares, $s_{yjm} \in [0, 1]$, for each market m, averaged across years. The stars show each dealer's stock market share on the y-axis and their derivatives market share on the x-axis; the circles show the stock market shares versus bond market shares, and the crosses the bond versus derivative market shares, on the y-axis and x-axis respectively. Figure 2b zooms in on dealers who trade at least 5% of the market share in one of the three markets.

Figure 3: Market specialization

(a) All dealers



(b) Largest dealers



Notes: Figure 3a plots all dealer j's market specialization scores, $\text{specialization}_{yjm} = s_{yjm} / \sum_m s_{yjm} \in [0, 1]$, for each market m, averaged across years. The stars show each dealer's stock market score on the y-axis and their derivatives market score on the x-axis; the circles show the stock versus bond market shares, and the crosses the bond versus derivative market score, on the y-axis and x-axis respectively. Figure zooms in on dealers who trade at least 5% of the market share in one of the three markets.

Product specialization. Product specialization may be driven by various factors, including differences in trading expertise, more effective inventory management, or relationships to clients with preferred habitat. On exchanges, some specialization also arises mechanically: firms designated as market makers are assigned subsets of securities and are obligated to trade them, leading to rule-based specialization. However, we show that this mechanical effect does not drive our results. When we exclude trades for market-making accounts (which designated market makers have to use when they are trading in their capacity as market maker), the main findings remain unchanged (see, for example, Appendix Figure 4 compared to Figure 4, which we explain below).

We assess the degree of product specialization within each market analogously to our assessment of market specialization.[12] First, Figure 4—analogous to Figure 1—visualizes dealer participation and product market shares—the fraction of a product's trade volume handled by each dealer—among the largest dealers across products. Comparing across markets, product specialization is lowest in the stock market. Not only do all dealers trade all products (as we see from the black box on the RHS of Figure 4a), dealers also distribute their trading more evenly across products. In the stock market colors on the LHS of Figure 4a are more consistent within dealers (horizontally) than across dealers (vertically), indicating a more even distribution of market shares, than in the bond and derivatives market.

Second, Figure 5—the analogue of Figure 3—presents the pairwise correlation of product specialization scores for a subset of products, which we define analogously to market specialization scores (1),

$$\text{specialization}_{yjmp} = \frac{s_{yjmp}}{\sum_p s_{yjmp}} \in [0,1], \tag{2}$$

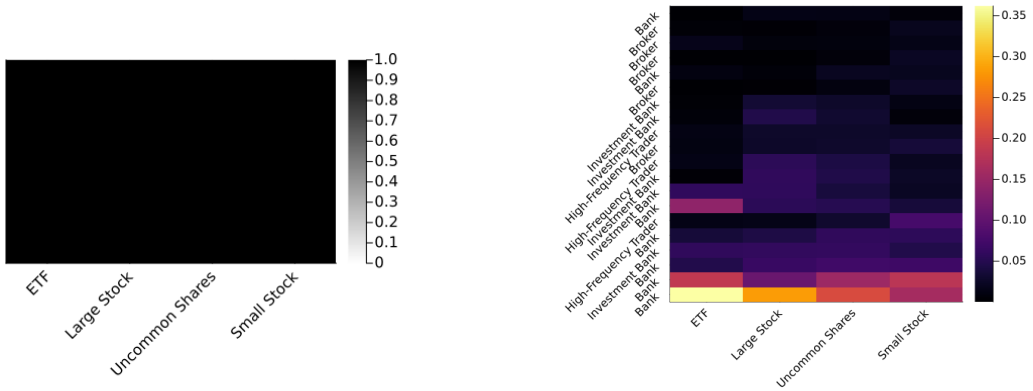
where s_{yjmp} represents the fraction of dealer j's trade volume in product p within market m, relative to all other dealers. Since each market contains more than three products, the figure is less intuitive than, and not directly comparable to, Figure 3a. However, as with market specialization, dealers with scores near the x- or y-axis—or at the extreme value of 1 which means that the dealer only trades one product—demonstrate higher degrees of product specialization. Comparing across markets, we note that product specialization scores in the stock market (blue crosses) tend to be more moderate, whereas scores in the bond and derivatives markets are more frequently close to 1 or 0, reflecting stronger specialization.

Our interpretation of these empirical patterns is that product specialization is influenced

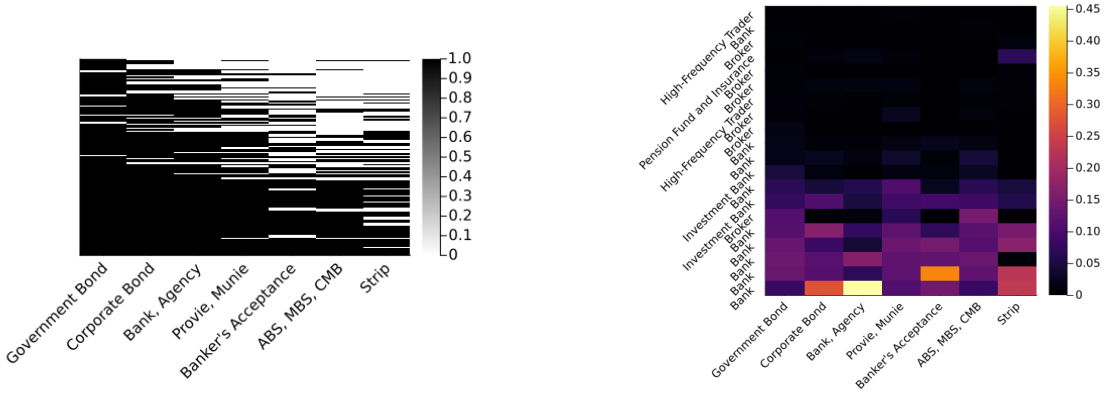
[12] In Appendix Figure A6 we further analyze the proportion of securities each dealer trades relative to the total number of securities in each market. For all dealers, this fraction is highest on the stock market, followed by the derivatives market, and then by the bond market.

Figure 4: Dealer presence and market shares across products in each market

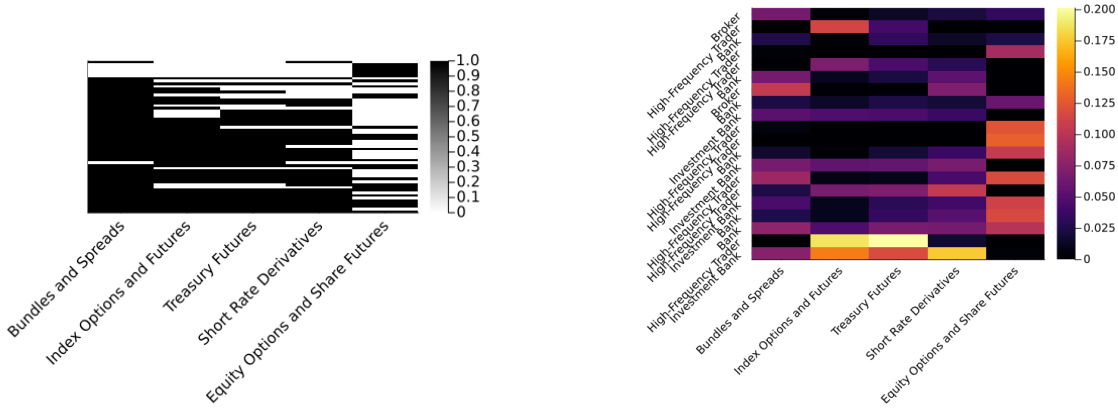
(a) Stock market



(b) Bond market



(c) Derivatives market

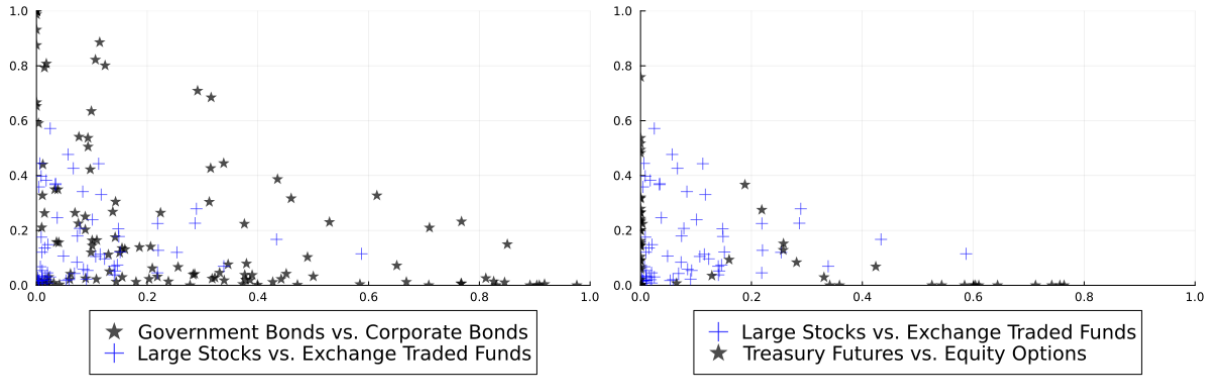


Notes: Figure 4 is similar to Figure 1 but replaces markets with products within each market, the stock market (a), bond market (b), and derivatives market (b), where we exclude currency options because they are so small. On the RHS we show whether dealers trade each product at least once in black. On the LHS we see the average annual product market shares of the largest dealers (on the LHS) for each market. In all figures, each row represents a dealer, sorted by total trade volume in the respective market. Dealers with the highest overall trade volume appear at the bottom, while those with the lowest appear at the top.

Figure 5: Product specialization

(a) Bond and stock market

(b) Stock and derivatives market



Notes: Figure 5 plots all dealer j's product specialization scores, $\text{specialization}_{yjmp} = s_{yjmp} / \sum_p s_{yjmp} \in [0,1]$, for a subset of products p for each market m, averaged across years. In 5a we compare bonds versus stocks and in 5b we stocks versus derivatives. The crosses show the specialization scores for large stocks on the y-axis and ETFs on the x-axis. The stars in 5a show the scores for government bonds on the y-axis and for corporate bonds on the x-axis; they show the scores for Treasury futures versus equity options and share futures in 5b.

by both market structure—OTC versus exchange—and product complexity. On the centralized stock exchange, where products are largely standardized, specialization is minimal.[13] In contrast, it is more pronounced in the decentralized bond market and the derivatives exchange, which includes both standardized and complex products.

One reason for decentralized markets to feature higher product specialization is that trading is more strongly dictated by the network structure among dealers and clients. We know from existing studies that dealers form long lasting relationships with clients, and that clients tend to have tastes for specific bonds (Di Maggio et al. (2017); Hendershott et al. (2020); Jurkatis et al. (2023); Allen and Wittwer (2024)). This could attribute to stronger product specialization in OTC markets compared to stock markets where network structures and relationships are less relevant.

Yet, search frictions and relationships in OTC markets alone cannot explain product specialization; otherwise, we would not observe specialization on the derivative exchange, which operates similarly to stock exchanges. Unlike stock exchanges, the derivatives exchange accommodates both standardized products, like Treasury futures, and more complex derivative contracts. Standardized products are likely to attract a broader set of dealers, similar to the universal dealer presence in stock markets, as they require minimal customization—dealers simply

[13]To further support our idea, we examine different types of uncommon shares (such as preferred shares and debentures). We find that more complex products are not traded by all dealers (see Appendix Figure A7 and Appendix Table A8 for the complete list of suffices).

select from a predefined menu. Consistent with this, many more dealers trade Treasury futures, short-rate derivatives, and index futures—the most standardized derivatives—compared to options or share futures, despite the latter accounting for around 28% of total average daily trade volume. However, this pattern does not hold universally. Many dealers trade bundles and spreads, which are more complex, suggesting that product complexity alone does not determine specialization.

Comparing market and product specialization. Dealer specialization by market and product aligns with the growing literature on segmentation in financial markets and confirms our prior. However, it is less clear which type of specialization is more pronounced. This distinction matters for both policy and modeling, as it highlights the more relevant dimension of segmentation. We will now gather evidence to establish our second main fact:

Fact 2 (Market versus product specialization)**.** *Across-product specialization within a market is for most dealers larger than across-market specialization.*

To compare cross-market and cross-product specialization, we introduce a specialization index—based on Theil (1967)'s index—, which integrates the specialization scores from Figures 3 and 5 in a way that allows for direct comparison. The standard Theil T-Index, T_j, captures the distribution of dealer j's market share across product-market segments. If the dealer trades the same fraction of total volume in each segment, $T_j = 0$. A positive T_j indicates specialization, with higher values reflecting greater variation in the dealer's activity across segments.

For our purposes, the original index is not suitable because it fails to account for the fact that non-participation by dealers in a market or market-segment increases specialization. To address this, we introduce a non-participation cost, ξ, which applies when a dealer does not trade in a market segment. Since this punishment term is chosen arbitrarily, the magnitude of our index by itself is not informative. However, it is valuable for comparing specialization within a market across products to specialization across different markets, which is our primary objective. To see this, note that the index decomposes into two components: one measuring within-market specialization, T_j^w, and another measuring across-market specialization, T_j^a:

$$T_j = T_j^a + T_j^w, \text{ with } T_j^w = \frac{1}{M}\sum_m T_{jm}^w.$$

T_{jm}^w measures the how dealer j's market share in a market-product segment is distributed across products in market m, with more uneven distributions meaning higher specialization; and T_j^a

Table 5: Specialization decomposition

	Across-market (T_j^a)	Across-product (T_j^w)
Dealers who participate in all markets	0.02–1.10	0.09–5.22
Dealers who participate in two markets	2.09–2.77	1.87–5.87
Conditional on being active in only two markets	0.42-1.10	0.20–3.96
Dealer who participate in one market	–	3.48–6.71
Conditional on being active in only one market	–	0.15–1.95

Notes: Table 5 reports the range of across-market specialization indices (T_j^a) and within-market, cross-product specialization indices (T_j^w) for a punishment term of $\xi = 5$. Dealers are grouped by activity in all markets, two markets, or one market. For dealers who are active in $K \in \{1, 2\}$ markets, ξ increases both indices by $(3-K)/3\xi$. Indices are computed with and without accounting for non-participation penalties. For single-market dealers, only cross-product indices are calculated, since the cross-market index is uninformative.

measures how the average of product market shares in each market are distributed across markets. Consult Appendix B for mathematical details.

Table 5 provides the range of indices across dealers for three dealer groups, those active in all markets, those active in two out of three markets, and those active in only one market; Appendix Figure A11a shows all indices for each dealer separately. For dealers who are not active in all markets, we compute both indices subject to punishment for non-participation in the market they are not active in. In addition, we also compute the indices conditioning on market participation to compare cross-product specialization within a market across these dealer groups.

For most dealers, within-market specialization is larger than across-market specialization, implying larger cross-product specialization compared to cross-market specialization. Notably, this is not driven by the fact that there are more products in a market than entire markets, like it would when considering other measures, such as the variance. Instead, the difference is driven by unequal participation across submarkets.[14]

Greater product than market specialization suggests that, for large financial institutions, barriers to market entry are less restrictive than factors that limit arbitrage across products

[14]The magnitudes of the indices depend on the punishment ξ-term. However, the conclusion that within-market specialization is larger than across market specialization does not depend on the choice of ξ. To show this, we shut off non-participation punishments by setting ξ to zero. In that case, the average (median) within product index (across all dealers) is 0.83 (0.73), and the average across-product index is 0.91 (1.10). The indices are relatively similar in size, but this is driven by dealers who only participate in one market. For those dealers both indices are identical. When restricting attention to dealers who participate in at least two markets (for which the within-market and across market measures differ), the average (median) within product index (across all dealers) is 0.53 (0.45), which is significantly smaller than the average across-product index is 0.75 (0.83).

within a given market. Given that product specialization appears to depend on whether a market is centralized (e.g., exchanges) or decentralized (e.g., OTC markets) and on the complexity of the products traded, our findings underscore the need for theories that account for different market structures or different degrees product complexities, unlike most existing models that focus on a single market structure with standardized assets. While our analysis does not assess whether trading specialization improves welfare, it highlights the role of market design and product complexity in shaping market fragmentation.

5 Dealer specialization and transaction prices

Thus far, our analysis has focused on market shares—that is, quantities. The second part of the paper turns to prices. We ask whether market and product specialization affect transaction prices. This could occur if specialization improves inventory management, shifts beliefs about fundamentals, or allows some dealers to extract better prices in the presence of limited competition. We focus on price differences relative to market averages, leaving effects on aggregate equilibrium price levels for future research. Concretely, the remaining of our paper serves to establish our third stylized fact:

Fact 3 (Specialization and prices)**.** *Dealers who are specialized trade at better prices relative to average market prices.*

To detect systematic differences in trade prices, we consider dealers at the LEI-level, but our findings are robust if we consider the dealers' parents instead. To reduce the sample size for our data from the exchanges, which is very large, we collapse the exchange data from the exchanges to the level of market segment (TSX/Alpha/TSXV/MX), day, security, dealer, trade direction (buy/sell), and trade type (active/passive). With slight abuse of terminology, we refer to each row in the collapsed dataset as a 'transaction' τ, and compute the total quantity traded, quantity_τ, and the average price, price_τ, for each τ.

Measuring price advantages. To detect systematic price differences across both sides of the trade and ensure comparability across securities with varying price levels, we follow the market microstructure literature and compare transaction prices relative to benchmark prices. Ideally, trading prices would be compared to fundamental values, but since these are rarely observable, equity studies use the prevailing mid-price, while bond market studies rely on inter-dealer prices, among other alternatives.

We seek a benchmark that is consistently available across markets and use each security's average daily price, which means our measure incorporates intra-day volatility. To account for

variation across securities and over time, all regressions include security-week fixed effects. Moreover, to ensure the average price is meaningful, we focus on sufficiently liquid securities traded at least three times in a day, with results remaining qualitative robust when restricting to more liquid securities (e.g., those traded at least five times daily).[15] We also exclude approximately 2% of derivatives transactions executed at negative prices—common in certain spread types—as these would complicate the interpretation of our findings.

With these data, we want to measure the relative price advantage compared to the market. The easiest approach would be to consider the percentage difference of a transaction τ for security s relative to the average price for that security on that day t:

$$\text{Margin}_\tau = \frac{\text{average price}_{ts} - \text{trade price}_\tau}{\text{average price}_{ts}} \times 100 \times \text{ trade sign}_\tau$$

Trade sign is one when the trader buys and -1 when they sell. A 1% margin says that the trade price is 1% below the securities' daily avg. price when buying, and above when selling. However, since there are occasional outliers—which is common for trade-level data—we follow Hendershott and Madhavan (2015)'s measure, which is identical to Margin_τ for prices that are sufficiently close to the average price, and trims outliers:[16]

$$\text{margin}_\tau = -\ln(\text{trade price}_\tau / \text{average price}_{ts}) \times 100 \times \text{ trade sign}_\tau \approx \text{Margin}_\tau. \tag{3}$$

Margins vary significantly more in the derivatives market, where price volatility within a day is highest, followed by the stock market, and finally the bond market, where price volatility is more moderate (see Appendix Figures A13–A16). Due to differences in trade sizes and prices across markets, a 1% margin difference results in different total payment magnitudes for the median trade: approximately C$160 in the stock market, C$12,500 in the bond market, and 10 cents in the derivatives market.

Sufficient condition for price effects. Before analyzing how specialization affects transaction prices, we first verify that no market is frictionless enough to prevent some dealers to systematically outperform others. In fully frictionless markets, such differences would be ar-

[15] This restriction is particularly stringent in the derivatives market, as symbols often include detailed and flexibly specified contract information. Our restricted dataset covers over 99% of stock market trades, approximately 88% of bond market trades, and about 54% of derivatives trades.

[16] Appendix Figure A12 shows the relationship between our main margin measure (3) and its linear approximation. Relative to the linear percentage difference, the log-measure attenuates the poor trades (which, for buyers, are those executed at higher than average prices), and amplifies the successful trades.

bitraged away, and specialization would not impact prices—even if it influenced underlying inventory costs or beliefs.

We regress our margin measure on dealer-indicator variables separately for each market:

$$\text{margin}_\tau = \alpha + \sum_j \beta_j \mathbb{I}(\text{dealer} = j) + \gamma \text{control}_\tau + \zeta_t + \zeta_{ws} + \epsilon_\tau. \tag{4}$$

If there is no systematic difference of dealers across markets, all dealer coefficients should be statistically insignificant from zero. We include a day fixed effects, ζ_t, to absorb time-varying shocks that affect the entire market, and year-week-security fixed effects, ζ_{ws}, to account for the average weekly margin of a given security.[17] This addresses two potential biases. The first arises because different dealers trade different securities, which naturally have varying margins; the second comes from the feature that the set of traded securities varies over time.

Additionally, we include control variables, though they do not significantly affect the overall pattern of the dealer coefficients. First, given prior evidence that trade size influences outcomes, we control for trade size. Second, for exchange trades, we account for the account type associated with the trade. For bond trades, we distinguish the trade type—whether it occurs between dealers (i.e., CIRO dealer members), between a dealer and an inter-dealer broker, or between a dealer and a non-dealer. Across markets, we use the same large primary dealer as the baseline for consistency.

In this and all other margin regressions, we cluster standard errors at the daily level to account for arbitrary intra-day correlations across dealers, securities, and trades. This is crucial when traders split orders throughout the day or react to price shocks that impact multiple securities.[18] Because some days feature many more trades trade than others, the day-clusters are highly uneven in size. This can result in conventionally computed standard errors being underestimated (MacKinnon et al. (2023)). One common solution is to compute standard errors via wild (WCR) bootstrapping following Cameron et al. (2008), and Roodman et al. (2019). Unlike standard methods, this approach does not rely on asymptotic approximations to the

[17] We do not include day-security symbol fixed effects because some symbols are not traded frequently enough throughout the day. However, for robustness, we have estimated all regressions with symbol-day fixed effects, and our main conclusions remain unchanged.

[18] An alternative approach is to cluster by dealers, accounting for correlations in a dealer's trades across days while ignoring intra-day correlations across dealers. However, with fewer than 100 dealers in the stock and derivatives markets and uneven cluster sizes, we are not confident this would yield reliable standard errors, even when boostrapping standard errors (MacKinnon et al. (2023)). Another option is to cluster at the symbol level to capture correlated shocks affecting the same symbol over time. We do not adopt this approach because many price shocks are likely correlated across symbols, making symbol-level clustering insufficient for addressing cross-symbol dependencies.

test statistic's distribution, which can be inaccurate when clusters are uneven or few in number. Instead, it constructs confidence intervals using bootstrap resampling, and therefore yields more reliable test statistics when clusters are small or uneven. To ensure robustness, we report coefficients that are statistically significant under both wild-bootstrapped and conventionally computed standard errors.

Our findings (Figure 6 and Appendix Figure A19) show that some dealers consistently secure better prices across all markets, both at the LEI- and parent-level. In the stock market, the best dealer (a large broker) achieves margins 0.64% better than the baseline (a primary dealer), while the worst (a smaller broker) lags by 0.31%, translating into an annual benefit of approximately C$266 million for the best dealer and a C$6 million loss for the worst.[19] Dealer differences are more pronounced in the derivative market due to price volatility, with the best (a large hedge fund) outperforming by 1.38% and the worst (a proprietary trading firm) underperforming by 1.30%, though total payment differences remain modest given contract prices and trade volumes. In the bond market, most dealers earn lower margins than the dominant primary dealer baseline, yet the best (a large insurance company) outperforms by 0.08%, gaining C$17 million annually, while the worst (serving retail clients) underperforms by 0.48%, losing C$278 thousand.

Table 6 examines which dealer types achieve better margins using cross-dealer variation by estimating regression (4) with dealer-type indicators, setting asset managers as the baseline.[20] High-frequency traders outperform other types in derivatives and perform well in stocks, though mutual funds dominate. Pension funds and insurance companies, the weakest performers in stocks, achieve the highest bond margins. Dealers specializing in retail clients earn the lowest margins in bonds and derivatives but perform comparably to asset managers in stocks.

Specialization affects prices. Having established that no market is sufficiently frictionless to eliminate systematic price differences, we next ask whether and how specialization affects transaction prices. In theory, the relationship between dealer specialization and prices is ambiguous. Greater specialization might enable dealers to trade at lower prices. However, it could also be that specialized dealers are more efficient than their less specialized counterparts and,

[19]To translate margin percentages into annual monetary losses or gains, we assume that each trade is executed at the median price, using the total amount (e.g., number of shares in the stock market) that the dealer under consideration trades in an average year.

[20]Consistent with prior studies (Bernhardt et al. (2005)), larger trades receive worse prices on exchanges, and execution prices are worse for client trades than for inventory or market-making accounts. In the bond market, we find no trade-size discounts, aligning with mixed literature (Pinter et al. (2024); Allen and Wittwer (2024)). Dealers earn higher margins trading with clients than with other dealers or brokers.

Table 6: Margin regression with dealer-types (bonds, stocks, derivatives)

Bond market	Margin
Trade size	+0.000
	(0.000)
	[0.000]
Counterparty is broker	+0.008**
	(0.003)
	[0.002]
Counterparty is client	+0.018***
	(0.002)
	[0.002]
Bank	0.021*
	(0.010)
	[0.004]
Broker	+0.003
	(0.007)
	[0.002]
High-Frequency Trader	0.027
	(0.032)
	[0.028]
Investment Bank	−0.049**
	(0.015)
	[0.011]
Mutual Fund	−0.013
	(0.008)
	[0.004]
Pension Fund and Insurance	+0.101***
	(0.008)
	[0.003]
Retail	−0.160***
	(0.009)
	[0.003]
Date-& symbol-year-week fes	Yes
N	6,757,118
R^2	0.017
Within-R^2	0.000

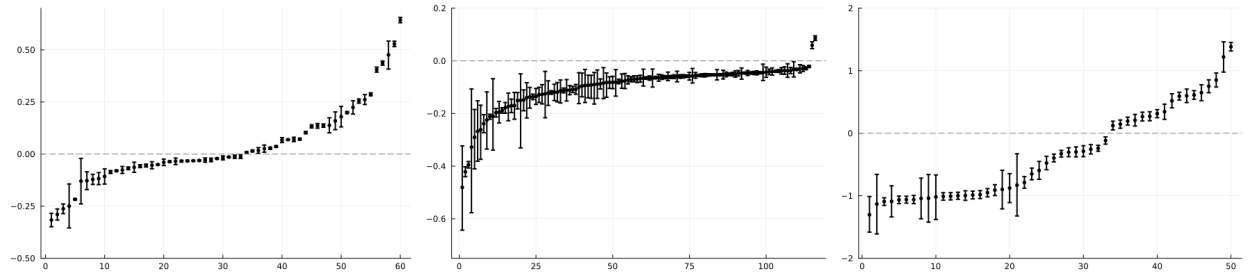
Stock market	Margin
Trade size	−0.162***
	(0.006)
	[0.006]
Client account	−0.069***
	(0.003)
	[0.001]
Inventory account	−0.000
	(0.003)
	[0.002]
Market-maker account	+0.070***
	(0.003)
	[0.002]
Bank	+0.071***
	(0.002)
	[0.002]
Broker	+0.045***
	(0.001)
	[0.001]
High-Frequency Trader	+0.115***
	(0.004)
	[0.004]
Investment Bank	+0.009***
	(0.002)
	[0.002]
Mutual Fund	+0.159***
	(0.019)
	[0.019]
Pension Fund and Insurance	−0.046***
	(0.003)
	[0.003]
Retail	+0.026***
	(0.002)
	[0.002]
Date-& symbol-year-week fes	Yes
N	111,051,211
R^2	0.008
Within-R^2	0.001

Derivative market	Margin
Trade size	−24.255***
	(2.249)
	[1.909]
Client account	−0.275***
	(0.066)
	[0.028]
Inventory account	+1.332***
	(0.072)
	[0.030]
Bank	−0.304**
	(0.097)
	[0.078]
Broker	+0.388***
	(0.034)
	[0.026]
High-Frequency Trader	+1.001***
	(0.041)
	[0.031]
Investment Bank	+0.597***
	(0.043)
	[0.036]
Other	+0.404
	(0.275)
	[0.261]
Retail	−0.394***
	(0.067)
	[0.051]
Date-& symbol-year-week fes	Yes
N	4,529,584
R^2	0.036
Within-R^2	0.006

Notes: Table 6 shows the estimation results from regressing margins (3) of stock market trades on trade-size, the account-type (client, inventory, market-market, or other—the baseline), dealer-types, and day- and security-year-week fixed effects on the LHS, for fixed-income trades in the middle, and for derivatives on the RHS. For bonds, we replace the account-type with the type of trade (dealer-broker, dealer-client, dealer-dealer—the baseline). Asset managers are the baseline. Margins are in percentage points, and quantities are in million C$. We cluster standard errors at the daily-level and report conventionally computed robust clustered standard errors in round brackets, and wild-bootstrapped standard errors in squared brackets. The stars reflect to the larger standard errors.

Figure 6: Dealer coefficients that are statistically different from zero at 5% significance level

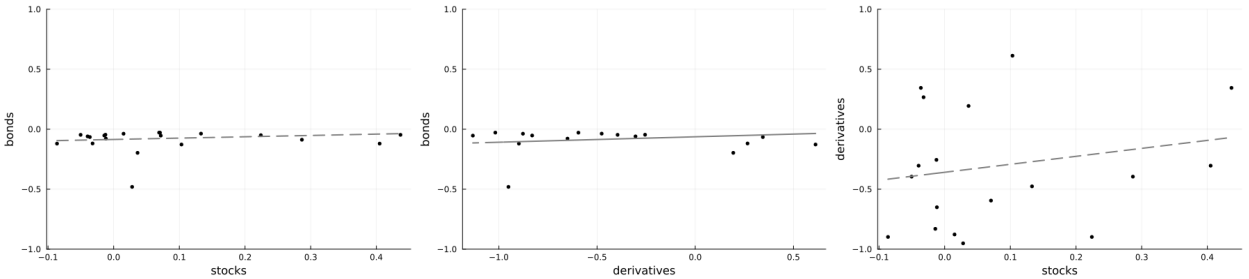
(a) Stocks (b) Bonds (c) Derivatives



Notes: Figure 6a shows the dealer coefficients, and 95% wild-bootstrapped confidence intervals (clustered at the daily-level), when regressing margins (3) of a stock market trade on dealer indicator variables and control variables (trade-size, the account-type, security-week-year and day fixed effects). Figures 6b and 6c show the analogue for the bond and derivatives market, respectively. For the bond market, we replace the account-type with a variable that indicates the type of trade (dealer-dealer, dealer-client, dealer-broker). In all graphs we exclude dealer coefficients that are not significantly different from zero at a significance level of 5% according bootstrapped and conventional inference to be conservative. Since we sort coefficients from small to large, the x-axis are not comparable across markets, as they do not reflect dealer identifiers. Appendix Figures A18 shows the analogous figures with conventionally computed confidence intervals; Appendix Figures A19 aggregates dealers to the parent-level.

Figure 7: Cross-market correlation between dealer coefficients of dealers active in all markets

(a) Bonds-stocks (b) Bonds-derivatives (c) Stocks-derivatives



Notes: Figure 7a shows the within-dealer correlation of coefficients in the bond (y-axis) versus stock market (x-axis), 7b and 7c show the correlation for the other two market pairs. We exclude dealer coefficients that are not significantly different from zero at a significance level of 5% according to bootstrapped and conventional inference to be conservative. Appendix Figure A21 shows the cross-market correlation between dealer coefficients when aggregating dealers to the parent-level.

as a result, are willing to trade at prices that less specialized dealers avoid. To find out we use two complementary strategies: examining dealers individually and analyzing patterns in the cross-section.

We begin by zooming in on the dealer level to assess whether dealers who trade across markets and products achieve better prices, or whether specialized dealers outperform. Specifically, we examine dealer fixed effects from regression (4) for dealers exclusively active in a single market—an extreme form of specialization. Appendix Figure A17 shows that many of these dealers earn worse-than-average prices relative to the baseline dealer, potentially due to lower trading volume. However, outcomes vary by market. In the bond market, the second-best dealer is a bond-only trader, suggesting successful specialization. In the derivatives market, specialization appears even more advantageous: a substantial share of high-performing dealers trade only derivatives, possibly reflecting the relative complexity of these products compared to equities and bonds.

Supporting the notion of market specialization, we find little evidence that dealers who perform well in one market also perform well in others. Figure 7 shows no significant cross-market correlations in dealer fixed effects from regression (4) for dealers active in multiple markets. Appendix C provides analogous evidence in favor of strong product specialization, showing that few dealers outperform across multiple products.

Next, we leverage cross-sectional variation across dealers in their product and market specialization scores (1) and (2). Ideally, we would want to know if specialization causes specialized dealers to trade at different prices compared to non-specialized ones.

A first naive approach would be to regress margins on specialization scores, in addition to the same control variables and fixed effects we have used above, for regression (4).

$$\text{margin}_{\tau} = \alpha + \beta \text{market specialization}_{yjm} + \gamma \text{controls}_{\tau} + \zeta_t + \zeta_{ws} + \epsilon_{\tau}, \tag{5}$$

and similarly with product specialization, which varies by year, dealer, market and product: product specialization$_{yjmp}$.[21]

A natural and pressing concern in interpreting these regressions is reverse causality: specialization may help dealers obtain better prices, but better prices may also lead dealers to specialize. As a result, the estimates may be biased, reflecting the broader endogeneity of prices and quantities in equilibrium.

We adopt two complementary approaches to mitigate endogeneity concerns. Our first approach is to lag specialization scores to examine whether dealers who were more specialized

[21] For completeness, we report estimation outputs from regression (5) in Appendix Table A9

last year obtain better prices today. Specifically, we estimation regression (5), but replace market specialization$_{yjm}$ by market specialization$_{y-1,jm}$ and similarly for the product specialization scores. This would provide the causal effect of specialization on margins if past specialization is exogenous to current pricing—that is, if dealers did not choose last year's specialization based on anticipated future margin opportunities, and if there are no unobserved time-varying dealer-level factors affecting both specialization and pricing.

Table 7 presents the regression results, showing that dealers with higher specialization scores in the previous year earn larger margins today. The relationship is stronger for product specialization than for market specialization. For example, on the stock exchange, moving from no market specialization (score of 0) to full specialization (score of 1) increases margins by 3.8 basis points; a full shift in product specialization raises margins by 41 basis points. Given the large trading volumes dealers manage over the year, even these seemingly modest per-trade gains translate into substantial monetary value.

Our second approach is to instrument for specialization scores in regression (5)—that is, to find an observable variable that affects margins only through its influence on specialization. Identifying valid instruments that allow us to disentangle quantity and price effects is notoriously difficult, as emphasized in the growing literature on demand estimation following Koijen and Yogo (2019).[22]

We use client orders on the stock exchanges—where we observe a sufficient volume of such orders—as an instrument for dealer specialization when dealers trade for their own accounts.[23] The idea is that dealers have limited discretion over client orders, which must be executed promptly. For example, a retail investor placing a stock order through a Fidelity brokerage account will have that order executed automatically by Fidelity. These client orders generate variation in dealer specialization that is plausibly unrelated to the margins dealers earn on their own-account trades. If this exclusion restriction holds—conditional on our standard control variables and fixed effects—the instrument allows us to identify the causal effect of specialization on stock market margin.

[22]Common instruments include stock index inclusions (e.g., Shleifer (1986); Chang et al. (2015); Pavlova and Sikorskaya (2023)); capital flows (e.g., Coval and Stafford (2007); Ben-David et al. (2022)); announcements of quantitative easing (e.g., Krishnamurthy and Vissing-Jørgensen (2011)); COVID-19 stimulus programs (e.g., Greenwood et al. (2022)); variation in government bond supply (e.g., Krishnamurthy and Vissing-Jørgensen (2012)); unexpected inventory shocks to dealers (e.g., Allen and Wittwer (2023)); and regulatory constraints such as investment mandates (e.g., Koijen and Yogo (2019)).

[23]We cannot apply the same strategy to the derivatives or bond markets. For derivatives, too few active traders receive enough client orders to construct a meaningful instrument. For bonds, client orders are not observed.

Table 7: Correlation between margins and last year's specialization scores

	Stocks		Bonds		Derivatives	
Lagged market specialization	0.038***		0.008		0.218***	
	(0.002)		(0.006)		(0.049)	
	[0.001]		[0.001]		[0.011]	
Lagged product specialization		0.413***		0.065***		0.290***
		(0.004)		(0.014)		(0.052)
		[0.001]		[0.004]		[0.009]
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	87,299,519	87,299,519	5,358,555	5,358,555	3,350,296	3,350,296
R^2	0.008	0.009	0.017	0.017	0.038	0.038
Within-R^2	0.000	0.001	0.000	0.000	0.000	0.000

Table 7 shows the estimation results from regressing margins (3) on our specialization measures (1) and (2) from last year, respectively, for each market separately, using all trades. In all regressions we include the same control variables and fixed effects as in regression (6): trade size, account-types for the exchange, trade-type for the bond market, dealer-types, date fixed effects and security-year-week fixed effects. We cluster standard errors at the daily-level and report conventionally computed robust clustered standard errors in round brackets, and wild-bootstrapped standard errors in squared brackets. The stars reflect to the larger standard errors.

The IV-estimates, reported in Tables 8, suggest that more specialized dealers obtain prices that are roughly 30 basis points better than average.[24] The OLS estimate for market specialization is smaller than the IV estimate, while the OLS estimate for product specialization is larger, though the difference is not statistically significant at the 5% level. A downward bias in the OLS estimate could arise from omitted variables that induce a negative correlation between the error term and specialization. For example, more specialized dealers may be more efficient and thus willing to accept lower margins. Conversely, an upward bias could result from positive correlations—for instance, if dealers choose to specialize in products or markets where they already enjoy favorable pricing conditions due to strong client relationships or reputational advantages.

Neither of our strategies to address endogeneity is without limitations. For example, if dealers are forward-looking and build specialization in anticipation of future pricing advantages, lagged specialization scores are not exogenous. Similarly, the IV approach hinges on an exclusion restriction that could be violated if clients systematically direct orders to dealers who secure better prices on their own-account trades. However, we view this risk as limited: clients do not observe transaction prices, which are private to exchange members and the exchange

[24]Appendix Table A10 shows the analogous results for the derivative exchange, where we include margins from all trades, not just trades for dealer-accounts to obtain sufficient power. The exclusion restriction is therefore more restrictive.

Table 8: IV regressions of margins on specialization scores for stocks

	(First Stage)	(OLS)	(IV)
s^c_{yjm}	-0.442***		
	(0.003)		
	[0.002]		
Market specialization		0.134***	0.385***
		(0.004)	(0.024)
		[0.001]	[0.000]
Controls	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes
N	27,300,881	30,356,466	27,300,881
R^2	0.349	0.124	0.126
Within-R^2	0.012	0.000	0.000

	(First Stage)	(OLS)	(IV)
s^c_{yjmp}	-0.206***		
	(0.002)		
	[0.001]		
Product specialization		0.335***	0.284***
		(0.008)	(0.037)
		[0.002]	[0.001]
Controls	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes
N	27,300,881	30,356,466	27,300,881
R^2	0.349	0.124	0.126
Within-R^2	0.012	0.000	0.000

Table 8 shows the IV estimation results for own-account stock market trades. Consider the LHS of 8. In column (First Stage), we show the first stage of the two stage least square estimator—regressing the market specialization score (1) on the fraction of all client-orders dealer j executes in market m in year y relative to other dealers, $s^c_{yjm} \in [0,1]$. In column (OLS) we present the OLS coefficient from regressing margins on market specialization, using trades for the dealer's own account for the stock market, and all trades for the derivatives market. In column (IV) we depict the corresponding IV estimate. The table on the RHS shows the analogous for product specialization, where the instrument is the fraction of all client-orders for product p dealer j executes in market m in year y relative to other dealers, $s^c_{yjmp} \in [0,1]$. In all regressions we include the same control variables and fixed effects as in regression (6): trade size, account-types for the exchange, dealer-types, date fixed effects and security-year-week fixed effects. We cluster standard errors at the daily-level and report conventionally computed robust clustered standard errors in round brackets, and wild-bootstrapped standard errors in squared brackets. The stars reflect to the larger standard errors.

operator, making such selection unlikely to be a first-order concern. More importantly, both approaches—despite their individual limitations—yield a consistent pattern: more specialized dealers obtain better prices.

6 Conclusion

We analyze dealer specialization across bond, stock, and derivative markets using a unique dataset that tracks trading activity across all major Canadian financial markets. Our findings show that product specialization within a market is stronger than market specialization, though not all dealers participate in every market. While no market is frictionless enough to prevent some dealers from consistently securing better prices, we find no evidence of cross-market or cross-product trading synergies. These results challenge the traditional view that financial intermediaries operate seamlessly across markets and products, and underscore the importance of market structure and product complexity in driving market fragmentation.

References

Adrian, T., E. Etula, and T. Muir (2014). Financial intermediaries and the cross-section of asset returns. *The Journal of Finance 69*(6), 2557–2596.

Allen, F. and D. Gale (2000). Financial contagion. *Journal of Political Economy 108*(1), 1–33.

Allen, J., J. Kastl, and M. Wittwer (2020). Estimating demand systems with bidding data. Working paper.

Allen, J. and M. Wittwer (2023). Centralizing over-the-counter markets? *Journal of Political Economy 131*(12), 3310–3351.

Allen, J. and M. Wittwer (2024). Bunlidng trades in over-the-counter markets. Working paper.

Anand, S. and P. Segal (2015). Chapter 11: The global distribution of income. In: Handbook of Income Distribution, A. B. Atkinson, Bourguignon, F. (eds.), 2: 937-979.

Babus, A. and K. Hachem (2023). Markets for financial innovation. *Journal of Economic Theory 208*, 1–39.

Babus, A., S. Moreira, and M. Marzani (2024). The rise of specialized financial products. Working paper.

Ben-David, I., J. Li, A. Rossi, and Y. Song (2022). Ratings-driven demand and systematic price fluctuations. *The Review of Financial Studies 35*, 2790–2838.

Bernhardt, D., V. Dvoracek, E. Hughson, and I. M. Werner (2005). Why do larger orders receive discounts on the london stock exchange? *The Review of Financial Studies 18*(4), 1343–1368.

Bessembinder, H., C. Spatt, and K. Venkataraman (2020). A survey of the microstructure of fixed-income markets. *Journal of Financial and Quantitative Analysis 55*(1), 1–45.

Brunnermeier, M. K. and L. H. Pedersen (2009). Market liquidity and funding liquidity. *The Review of Financial Studies 22*(6), 2201–2238.

Brunnermeier, M. K. and Y. Sannikov (2014). A macroeconomic model with a financial sector. *American Economic Review 104*(2), 379–421.

Budish, E., R. S. Lee, and J. J. Shim (2024). A theory of stock exchange competition and innovation: Will the market fix the market? *Journal of Political Economy 132*(4), 1209–1246.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economic Statistics 90*, 414–427.

Chang, Y.-C., H. Hong, and I. Liskovich (2015). Regression discontinuity and the price effects of stock market indexing. *The Review of Financial Studies 28*(1), 212–246.

Chaudhary, M., Z. Fu, and J. Li (2022). Corporate bond multipliers: Substitutes matter. Working paper.

Chen, D. and D. Duffie (2021). Market fragmentation. *American Economic Review 111*(7), 2247–2274.

Chen, Z., N. Roussanov, X. Wang, and D. Zou (2024). Common risk factors in the returns on stocks, bonds (and options), redux. Working paper.

CIRO (2024). Proposed integrated fee model. Public announcement, available at: https://www.ciro.ca/newsroom/publications/proposed-integrated-fee-model?utm_source=chatgpt.com.

Coval, J. and E. Stafford (2007). Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics 86*(2), 479–512.

Di Maggio, M., A. Kermani, and Z. Song (2017). The value of trading relations in turbulent times. *Journal of Financial Economics 124*(2), 266 – 284.

Dix, R. and M. Wittwer (2025). Estimating demand systems with trade data. Working Paper.

Dougast, J., S. Üslü, and P.-O. Weill (2022). The theory of participation in otc and centralized markets. *The Review of Economic Studies 89*(6), 3223–3266.

Du, W., A. Tepper, and A. Verdelhan (2018). Deviations from covered interest rate parity. *The Journal of Finance 73*(3), 915–957.

FINRA (2024). Self-regulatory organizations; financial industry regulatory authority, inc.; notice of filing and immediate effectiveness of a proposed rule change to adjust FINRA fees to provide sustainable funding for FINRA's regulatory mission. Public announcement, available at: https://www.federalregister.gov/documents/2024/11/27/2024-27764/self-regulatory-organizations-financial-industry-regulatory-authority-inc-notice utm_source=chatgpt.com.

Greenwood, R., T. Laarits, and J. Wurgler (2022). Stock market stimulus. Working paper.

Gromb, D. and D. Vayanos (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics 66*, 361–407.

Hasbrouck, J. and G. Sofianos (1993). The trades of market makers: An empirical analysis of NYSE specialists. *The Journal of Finance 48*(5), 1565–1593.

Hau, H., P. Hoffmann, S. Langfield, and Y. Timmer (2021). Discriminatory pricing of overthe-counter derivatives. *Management Science 67*(11), 6660–6677.

He, Z., B. Kelly, and A. Manela (2017). Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics 126*(1), 1–35.

He, Z. and A. Krishnamurthy (2013). Intermediary asset pricing. *American Economic Review 103*(2), 732–770.

Hendershott, T., D. Li, D. Livdan, and N. Schürhoff (2020). Relationship trading in over-the-counter markets. *The Journal of Finance 75*(2), 683–743.

Hendershott, T. and A. Madhavan (2015). Click or call? Auction versus search in the over-the-counter market. *Journal of Finance 70*(1), 419–447.

Jurkatis, S., A. Schrimpf, K. Todorov, and N. Vause (2023). Relationship discounts in corporate bond trading. *BIS Working Papers, No. 1140*, 1–46.

Koijen, R. S. J. and M. Yogo (2019). A demand system approach to asset pricing. *Journal of Political Economy 127*(4), 1475–1515.

Krishnamurthy, A. and A. Vissing-Jørgensen (2011, Fall). The effects of quantitative easing on interest rates: Channels and implications for policy. *Brookings Papers on Economic Activity*.

Krishnamurthy, A. and A. Vissing-Jørgensen (2012). The aggregate demand for Treasury debt. *Journal of Political Economy 120*(2), 233–267.

Kumar, P. and D. J. Seppi (1992). Futures manipulation with “cash settlement". *Journal of Finance 47*(4), 1485–1502.

Lu, L. and J. Wallen (2024). What do bank trading desks do? Working paper.

MacKinnon, J. G., M. O. Nielsen, and M. D. Webb (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics 232*, 272–299.

Malamud, S. and M. Rostek (2017). Decentralized exchange. *American Economic Review 107*(11), 3320–3362.

Menkveld, A. J. (2016). The economics of high-frequency trading: Taking stock. *Annual Reviewof Financial Economics 8*, 1–24.

Mota, L. and K. Siani (2024). Financially sophisticated firms. Working paper.

O’Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics 116*, 257–270.

O’Hara, M. and X. A. Zhou (2021). Anatomy of a liquidity crisis: Corporate bonds in the COVID-19 crisis. *Journal of Financial Economics 142*(1), 46–68.

Pasquariello, P. (2014). Financial market dislocations. *The Review of Financial Studies 27*(10), 1868–1914.

Pavlova, A. and T. Sikorskaya (2023). Benchmaring intensity. *The Review of Financial Studies 36*(3), 859–903.

Pinter, G., C. Wang, and J. Zou (2024). Size discount and size penalty: Trading costs in bond markets. *Review of Financial Studies 37*(7), 2156–2190.

Roodman, D., M. O. Nielsen, J. G. MacKinnon, and M. Webb (2019). Fast and wild: Bootstrap inference in stata using boottest. *The STATA Journal 19*(1), 4–60.

Sandulescu, M. (2020). How integrated are corporate bond and stock markets? Working paper.

Shleifer, A. (1986). Do demand curves for stocks slope down? *The Journal of Finance 41*(3), 579–590.

Shleifer, A. and R. W. Vishny (1997). The limits of arbitrage. *The Journal of Finance 52*(1), 35–55.

Siriwardane, E., A. Sunderam, and J. L. Wallen (2022). Segmented arbitrage. Working paper.

Siriwardane, E. N. (2019). Limited investment capital and credit spreads. *The Journal of Finance 74*(5), 2303–2347.

Theil, H. (1967). *Economics and Information Theory*. North-Holland Publishing Company.

Üslü, S. and G. Pintér (2023). Comparing search and intermediation frictions across markets. Working paper.

Vayanos, D. and J.-L. Vila (2021). A preferred-habitat model of the term structure of interest rates. *Econometrica 89*(1), 77–112.

Weill, P.-O. (2020). The search theory of over-the-counter markets. *Annual Review of Economics 12*, 747–773.

Wittwer, M. and J. Allen (2023). Market power and capital constraints. Working paper.

ONLINE APPENDIX

Market and Product Specialization in Financial Markets

by Milena Wittwer, and Andreas Uthemann

Section A provides details regarding data cleaning.
Section B provides mathematical details for our specialization indices.

A Data cleaning

Data Restrictions. Our bond data includes all bond trades that are reported by CIRO-dealers to MTRS.20, excluding foreign sovereign bonds. We exclude primary market trades. In rare cases, trades are reported on a weekend. We treat those cases as trades that occur on the Monday following the weekend.

We include all stock-market trades, including those executed during the opening and closing auctions. In rare cases, trades are associated with negative trade amounts. We exclude those trades.

We keep regular derivative trades, and excludes rare cases of trades involving 'test futures'. In rare cases, trades are reported on a weekend. We treat those cases as trades that occur on the Monday following the weekend. We exclude a handful of dates where only Buy-ins are trading.

Quality Check. We compare the average monthly trade volume on the stock markets with the publically available information that is provided in CIRO's website to confirm that we observe close to 100% of the trades we should observe.

We also compare the derivative trade volume with information provided on MX's website. After restricting the raw data, as explained above, we observe roughly 90% of trade volume on average.

Our bond data is provided directly by the regulator and serves as the source for publicly available information on bond market trading volumes. We therefore do not cross-validate it against reported figures.

Type Classification for Dealers. The Bank of Canada classifies traders based on their LEIs into types following their in-house methodology. We replicate their approach to classify dealers on the stock exchanges and the derivative exchange, and to classify the parent-holding company of each LEI (see Appendix Table A4). Here we briefly describe their approach.

We use two types of information to classify entities – “Direct” and “Indirect” sources of information. “Direct” information refers to any information provided by the entity itself - either through its official website, internal documents, spokespersons, a regulatory organization to which it reports, etc. “Indirect” information refers to any information which is not direct information. The latter is further broken down into two sub-types: “Reliable" or “Weakly reliable". For the dealers in this project, all information comes from reliable sources, such as Bloomberg, Yahoo Finance, CapEdge, etc.

B Theil Index

To define the dealer-specific measure, consider a fixed dealer j. Let there be M markets, indexed by m, and P_m products within each market m, and $P = \sum_m P_m$ products overall, indexed by p. Denote dealer j's share of total volume traded by dealers in product-market segment mp by $s_{jmp} \in [0,1]$. The cross-product average for a dealer within a market is given by $\bar{s}_{jm} = \frac{1}{P}\sum_p s_{jmp}$, and $\bar{s}_j = \frac{1}{M\times P}\sum_m \sum_p s_{jmp}$ the overall average.

The standard Theil T index in this setting is defined as:

$$T_j = \frac{1}{M \times P} \sum_m \sum_p \left(\frac{s_{jmp}}{\bar{s}_j}\right) \ln\left(\frac{s_{jmp}}{\bar{s}_j}\right). \tag{6}$$

This index captures the distribution of dealer j's market share across product-market segments. If the dealer trades the same fraction of total volume in each segment, $T_j = 0$. A positive T_j indicates specialization, with higher values reflecting greater variation in the dealer's activity across segments.

For our purposes, the original index is not suitable because it fails to account for the fact that non-participation by dealers in a market or market-segment increases specialization. The original formulation only sums the trade volume of dealers who are active in a market, ignoring those who are inactive. To address this, we introduce a non-participation cost, ξ, which applies when a dealer does not trade in market mm (i.e., when $s_{jm} = 0$):

$$T_j = \underbrace{\frac{1}{M \times P} \sum_m \sum_p \mathbb{I}(s_{jmp} > 0) \left(\frac{s_{jmp}}{\bar{s}_j}\right) \ln\left(\frac{s_{jmp}}{\bar{s}_j}\right)}_{\text{Standard Theil index conditional on participation}} + \underbrace{\frac{1}{M \times P} \sum_m \sum_p \mathbb{I}(s_{jmp} = 0)\xi}_{\text{Non-participation}}.$$

The magnitude of the index depends on the size of the penalty, ξ, which can be chosen arbitrarily. As a result, the index by itself is not informative in absolute terms. However, it is valuable for comparing specialization within a market across products to specialization across

different markets, which is our primary objective.

To see this, note that the index decomposes into two components: one measuring within-market specialization, T_j^w, and another measuring across-market specialization, T_j^a:

$$T_j = T_j^a + T_j^w, \text{ with } T_j^w = \frac{1}{M}\sum_m T_{jm}^w, \text{ where}$$

$$T_{jm}^w = \frac{1}{P}\sum_p \mathbb{I}(s_{jmp} > 0 \cup \bar{s}_{jm} > 0)\left(\frac{s_{jmp}}{\bar{s}_j}\right)\ln\left(\frac{s_{jmp}}{\bar{s}_{jm}}\right) + \xi\mathbb{I}(s_{jmp} = 0 \cup \bar{s}_{jm} > 0)$$

measures the how dealer j's market share in a market-product segment is distributed across products in market m, with more uneven distributions meaning higher specialization; and

$$T_j^a = \frac{1}{M}\sum_m \mathbb{I}(\bar{s}_{jm} > 0)\left(\frac{\bar{s}_{jm}}{\bar{s}_j}\right)\ln\left(\frac{\bar{s}_{jm}}{\bar{s}_j}\right) + \xi\mathbb{I}(\bar{s}_{jm} = 0)$$

measures how the average of this market share across products is distributed across markets.

As for the standard Theil index, the minimum value for both measures is 0, which is the case when a dealer distributes their trading activity evenly across products in a market for T_{jm}^w, or on average across markets for T_j^a. The maximum value is given by M, P_m and ξ, namely, $\overline{T}_{jm}^w = \frac{1}{P_m}[MP_m\ln(P_m) + (P_m - 1)\xi]$, and $\overline{T}_j^a = \frac{1}{M}[M\ln(M) + (M-1)\xi]$.

C Additional evidence: product specialization and prices

To detect cross-product price effects, we add product indicators to regression (4), and estimate the following regression for each market separately:

$$\text{margin}_\tau = \alpha + \sum_p\sum_j \beta_{jp}\mathbb{I}(\text{dealer} = j \text{ and product} = p) + \gamma\cdot\text{control}_\tau + \zeta_t + \zeta_{ws} + \epsilon_\tau. \quad (7)$$

If the same dealer obtains similar margins across products within a market compared to the baseline, all β_{jp} coefficients would be similar in size. If the dealer is more successful when trading some products relative to others, these coefficients would differ. As before, we include day and security-week fixed effects to avoid potential biases that arise from time-variation in the traded securities.[25] For bonds, the baseline is a large primary dealer trading government bonds, for stocks it is that bank trading large stocks, and for derivative it is that bank trading

[25] As robustness, we also estimate a specification with only include day-fixed effect to exploit variation of margins across securities within the same product category. While the size of the coefficients differs, the main take away (Fact **??** is robust).

Treasury futures.[26]

We visualize the estimation outcome through heatmaps, one for each market, in Appendix Figure A22. Since estimating regression (7) is computationally intensive, especially for the stock market, we estimate it for each of the years in our sample separately, and report results for 2022. A row in the heatmap correspond to a dealer j. A column corresponds to a product p. When dealer j obtains systematically worse margins for product p, we color the corresponding jp cell red, meaning that the β_{jp} coefficient is negative and statistically different from zero. The cell is black if the dealer outperforms the other dealers, and empty if they either do not trade product p or the coefficient is not statistically significantly different from zero.

If the one dealer were to outperform (underperformed) the baseline across products, we would observe a black (red) line for that dealer. This is not the case for any dealer in any market—a take away that is robust across years, while the β estimates vary. Crucially, since the margin measure reflects price volatility over a day, this analysis does not imply that some products are inherently more profitable—we do not account for underlying trading costs. Rather, the key takeaway is that no dealer consistently outperforms all others across products, pointing towards product specialization in all markets.

[26]We clustered at the daily-level, and compute standard errors via wild-bootstrapping. This is useful not only because it circumvents issues that arise from uneven cluster sizes, but also because many indicator variables in regression (7) are zero, since dealers tend to specialize in specific products. This implies that the standard cluster-robust covariance matrix is close to singular (non-invertible) due to high correlation within some clusters with many zeros. Since wild-bootstrapping resamples residuals with cluster-dependent perturbations, and does not directly rely on inverting the covariance matrix, bootstrapping circumventing the issue.

Appendix Table A1: Fixed-income products

Product	Description
Government Bond	Government of Canada Bond, Government of Canada Real Return Bond, Government of Canada T-bill
Corporate Bond	Corporate Bond
Provie, Munie	Provincial Bill, Provincial Bond, Provincial Commercial Paper, Municipal Bond
Bank, Agency Paper	Bank Commercial Paper and Bank Security - Note/Bond/Debenture. Agency Bond and Agency Commercial Paper
Bankers' Acceptance	Bankers' Acceptance
ABS, MBS, CMB	Mortgage-Backed Security, Asset-Backed Security, Canada Mortgage Bond.
Strip	Agency Strip Bond, Bank Strip Bond, Corporate Strip Bond, Finance company Strip Bond, Government of Canada Strip Bond, Municipal Strip Bond, Provincial Strip Bond

Appendix Table A2: Equity products

Product	Description
Large Stock	Symbols without suffices (i.e., common shares) that are listed with missing sp-type with more than 2 billion of quoted market value
Small Stock	Symbols without suffices (i.e., common shares) that are listed with missing sp-type with less than 2 billion of quoted market value
Exchange Traded Funds	Symbols that are listed with sp-type being Exchange Traded Funds
Uncommon Shares	Symbols which suffices that aren't listed as Exchange Traded Funds, which include the the following types: preferred stocks, class A-C, notes, debentures, equity dividends, when-issued capital pool companies, warrants, redeemable common stocks, U.S. funds, units, subscr. receipts, and stocks that trade on the NEX market
Others or Missing	Symbols without suffices that have a non-missing sp-type, which include the following sp-types: Income Trust, Fund of Equities, Commodity Funds, Exchange Traded Receipt, Split Shares, Fund of Mortgages/MBS, Fund of Debt

Appendix Table A3: Derivative products

Product	Description/Symbols if available
Treasury futures	Government bond futures and future options; CGZ, CGF, CGB, LGB, OGZ, OGF, OGB
Short-term derivatives	BAX futures and future options, CORRA futures; BAX, OBW, OBX, OBY, or OBZ, CRA
Equity options and share futures	Equity option, weekly option, option on ETFs, share futures
Currency options	Options on USD; USX
Index options and futures	Index futures and options; SXF, SCF, SXB, SXY, SXK, SXJ, SEG, SXM, SXA, SXH, SXO, SXU, SXV, SCG, SDV
Bundles and spreads	User-defined strategy, inter-group strategies, spreads

Notes: Appendix Tables A1–A3 describe our product classification for the bond, stock, and derivative market, respectively.

Appendix Table A4: Dealer types classification

Broker	Financial entity whose purpose is to offer brokerage services
Investment Bank	Investment bank
Bank	Bank, retail bank or credit union, and any entity that is deposit taking
Asset Manger	Financial entity whose purpose is to manage assets (or investments) and/or offer investment advising services. Entities that manage multiple types of funds such as HF, MF or ETFs are also classified as such
Mutual Fund	Financial entity that is a mutual fund or a mutual fund manager
High-Frequency Trader	Financial entity that is a hedge fund or a hedge fund manager; Private Equity, or Proprietary Trader
Pension Fund and Insurance	Financial entity whose purpose is to manage investments (and/or provide services) related to pension, retirement, insurance, re-insurance, benefits, and superannuation funds
Retail	Financial entity whose purpose is to offer financial services to retail (non-institutional) investors
Other	This category includes all other types which we observe in our traded data. From the Bank of Canada classification we pool the following types under this category “Real Estate (a financial or non-financial entity that is involved in the construction, financing, management, or sale of commercial, industrial, or residential real estate), “Other" (Financial entity that does not fall in any of the aforementioned classifiers (e.g., Financial Planner, Financial Research Services, Execution Platform)), Uncategorized (entity that can neither be classified as a financial nor a non-financial entity due to lack of information), “Non-financial entity". We also include Buy-Ins that execute some trades on the exchanges here.

Notes: Appendix Table A4 explains the classification of trader types we adopt following the methodology of Bank of Canada staff.

Appendix Table A5: Daily trade volume, number of active dealers, trade-sizes, and prices

Variable	Mean	Median	Min	Max	Std
Daily Trade Volume					
Stocks (in mil)	661.053	613.824	155.859	1746.480	209.427
Bonds (in bn C$)	71.723	68.410	0.196	1,504.660	54.771
Derivatives (in k)	340.366	323.155	1.487	1,017.150	126.632
Number of Active Dealers (LEIs)					
Stocks	64.157	64.0	61.0	69.0	1.630
Bonds	61.041	62.0	5.0	84.0	7.794
Derivatives	53.617	54.0	21.0	57.0	3.019
Number of Active Dealers (Parents)					
Stocks	60.705	61.0	58.0	64.0	1.180
Bonds	57.394	58.0	11.0	71.0	6.636
Derivatives	47.737	48.0	21.0	51.0	2.668
Trade size					
Stocks	11875.7	1300.0	0.1	7.361×10^7	74,694.1
Bonds (in mil C$)	10.934	1.250	$1.0/10^6$	7.189×10^5	436.177
Derivatives	78.143	10.0	1.0	170,478.0	667.365
Trade price					
Stocks (in C$)	27.280	12.240	0.005	2,392.4	86.936
Bonds (in C$)	102.344	100.24	1.0	980.0	12.398
Derivatives (in C$)	20.809	1.060	-142.050	21,800.0	123.298

Notes: Appendix Table A5 summarizes trade data for stocks (TSX, TSXV, Alpha), bonds (MTRS), and derivatives (MX) from 2019 to 2022. It provides the mean, median, minimum, maximum, and standard deviation of daily trade volume, the number of active dealers ("Number of Dealers (LEI)") and parent institutions ("Parents"), trade size, and trade price. For derivatives, trade size and volume reflect the number of contracts, not the underlying asset value. There are 1,004 active trading days for stocks, 994 for derivatives, and 1,035 for bonds. Some bond trades occur on Canadian holidays, when the Investment Industry Association of Canada (IIAC) recommends pausing trading. These days, typically involving minimal activity, are excluded from the table but included in the analysis with either lower-frequency aggregation or day-fixed effects to account for special cases. The stock market features 6,449 symbols traded by 72 dealers and 66 parent institutions. The derivatives market, where symbols often include contract details like expiration dates, has 503,056 symbols traded by 64 dealers and 56 parent institutions. In the bond market, 107,516 CUSIPs are traded by 163 dealers (CIRO dealer members in the raw data) and 131 parent institutions. Only a subset of these dealers is active daily.

Appendix Table A6: Member intersection: RHS: LEI-level of members; LHS—Parent level

Intersection	TSX	ALPH	TSXV	Intersection	TSX	ALPH	TSXV
All	99.97	100.0	99.68	All	99.97	100.0	99.68
TSX and ALPH	0.0	0.0	0.0	TSX and ALPH	0.0	0.0	0.0
TSX and TSXV	0.03	0.0	0.32	TSX and TSXV	0.03	0.0	0.32
ALPH and TSXV	0.0	0.0	0.0	ALPH and TSXV	0.0	0.0	0.0
TSX only	0.0	0.0	0.0	TSX only	0.0	0.0	0.0
ALPH only	0.0	0.0	0.0	ALPH only	0.0	0.0	0.0
TSXV only	0.0	0.0	0.0	TSXV only	0.0	0.0	0.0

Notes: Appendix Table A6 shows the percentage of total volume traded in each of the three stock exchanges (TSX, TSXV, and Alpha) that is traded by brokers who trade on all segments (All), on TSX and Alpha, etc. Each column sums to 100%. Total volume traded is computed by summing all quantities of all brokers including both sides of the trade.

Appendix Table A7: Avg. weekly share traded by dealers per type (parent-level)

Dealer type	MTRS	TSX	TSX-IN	MX	MX-IN
Asset Manager	0.27	1.70	0.98	0.68	0.00
Bank	74.13	56.05	35.77	27.12	19.05
Broker	12.32	13.47	16.46	9.12	2.02
High-Frequency Trader	0.86	10.70	24.24	34.94	72.07
Investment Bank	11.88	17.21	22.49	27.98	6.75
Other	0.00	0.00	0.00	0.11	1.11
Pension Fund and Insurance	0.50	0.23	0.03	0.00	0.00
Retail	0.02	0.60	0.00	0.07	0.00

Notes: Appendix Table A7 shows the fraction of trade volume by dealer type (at the parent-level) per market in columns MTRS, TSX, and MX, respectively. In columns TSX-IN, and MX-IN we show the analogue but excluding trades for client accounts for TSX and MX.

Appendix Table A8: Symbol Suffixes on TSX/TSXV/Alpha

Symbol suffix	Description
None	Common shares
A, B, C	Class A, B, C of shares is typically related to voting rights, access to dividend
DB	Debenture, stock type that makes fixed payments at scheduled intervals of time, operates similar to preferred stock
E	Equity dividend
F	
G	
H	NEX market provides a trading forum for listed companies that no longer meet the TSX Venture's ongoing listing standards; designed for companies that have low levels of business activity or have ceased to carry on active business. It benefits such companies by giving their stocks a degree of liquidity and providing visibility that may attract potential acquirers or investors.
K	NEX market
IR	Installment receipts, is an equity issuance in which the purchaser does not pay the full value of the issue up front. In the purchase of an installment receipt, an initial payment is made to the issuer at the time the issue closes; the remaining balance must be paid in installments, usually within a two-year period. Although the purchaser has not paid the full value of the issue, he or she is still entitled to full voting rights and dividends.
J	
L	Legended shares. A legend is a statement on a stock certificate noting restrictions on the transfer of the stock. A stock legend is typically put in place due to the requirements established by the Securities and Exchange Commission (SEC) for unregistered securities. A stock legend may or may not be legally required on the certificate itself, depending on state laws.
M	Booms
N	Subscription receipts (second issue trading)
O	Subscription receipts (third issue trading)
NO, NS, NT	Exchange traded note (ETN) are unsecured debt securities that tracks an underlying index of securities.
P	Capital pool company (CPC) is an alternative way for private companies in Canada to raise capital and go public. The capital pool company system was created and is currently regulated by the TMX Group, and the resulting companies trade on the TSX Venture Exchange in Toronto, Canada.
Q	
PR, PF, PS	Preferred shares; similar to common shares, no maturity date, ownership, fixed distribution rate, no voting rights
PR.CLASS, PS.CLASS, PF.CLASS	Preferred class
R	Subscription receipts are defined as those limited term securities issued via prospectus, which are convertible into another security class of the issuer (predominantly common shares) at a set conversion rate based on the successful completion of a planned reorganization or transaction. Where completion is not successful, security proceeds are either returned to the subscriber or a more generous conversion rate is made available to the subscriber.
RT	Rights are instruments issued by companies to provide current shareholders with the opportunity to preserve their fraction of corporate ownership. Rights are short-term instruments that expire quickly, usually within 30-60 days of issuance. The exercise price of rights is always set below the current market price, and no commission is charged for their redemption.

Appendix Table A9: Correlation between margins and specialization scores

	Stocks		Bonds		Derivatives	
market specialization	0.022***		0.007		0.440***	
	(0.002)		(0.005)		(0.039)	
	[0.000]		[0.001]		[0.008]	
product specialization		0.388***		0.062***		0.237***
		(0.004)		(0.009)		(0.043)
		[0.001]		[0.003]		[0.008]
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	111,051,211	111,051,211	6,757,118	6,757,118	4,529,585	4,529,585
R^2	0.008	0.009	0.017	0.017	0.036	0.036
Within-R^2	0.000	0.001	0.000	0.000	0.000	0.000

Appendix Table A9 shows the estimation results from regressing margins (3) on our specialization measures (1) and (2),respectively, for each market separately, using all trades. In all regressions we include the same control variables and fixed effects as in regression (6): trade size, account-types for the exchange, trade-type for the bond market, dealer-types, date fixed effects and security-year-week fixed effects. We cluster standard errors at the daily-level and report conventionally computed robust clustered standard errors in round brackets, and wild-bootstrapped standard errors in squared brackets. The stars reflect to the larger standard errors.

S	Special U.S. terms
T	Special US trading terms (second issue trading)
U	U.S. dollar
V	U.S. dollar (second issue trading)
UN	Units are a securities that is made up of one common share and half a warrant. Units are commonly offered by special-purpose acquisition companies, or SPACs that are seeking to raise money in a public stock offering and trade on a stock exchange with the primary goal of merging with a private business and taking it public.
W	When issued
WB	
WR	
I	When issued (second issue trading)
WT	Warrants give the holder the right to purchase a company's stock at a specific price and at a specific date.
X	
Y	Redeemable common. Redeemable shares are shares that a company has agreed it will, or may, redeem (in other words buy back) at some future date. The shareholder will still have the right to sell or transfer the shares subject to the articles of association or any shareholders' agreement.

Notes: Appendix Table A8 describes the meaning of all suffixes of symbols trades on TSX, TSXV, or Alpha. An empty cell means that we were not able to find the description of a symbol that we observe in the raw data.

Appendix Table A10: IV regressions of margins on specialization scores for derivatives

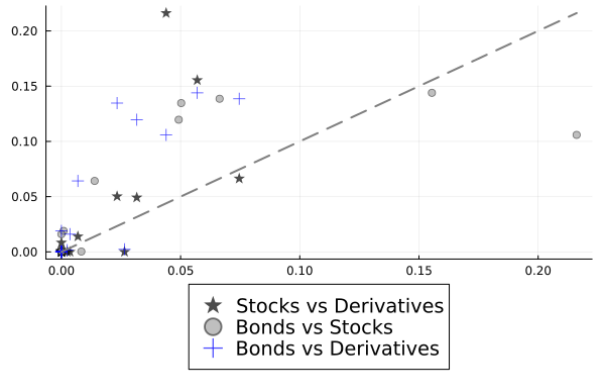
	(First Stage)	(OLS)	(IV)
s^c_{yjm}	-1.291***		
	(0.061)		
	[0.014]		
Market specialization		0.440***	1.691***
		(0.039)	(0.258)
		[0.008]	[0.007]
Controls	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes
N	2,911,210	4,529,585	2,911,210
R^2	0.495	0.036	0.125
Within-R^2	0.032	0.000	0.000

	(First Stage)	(OLS)	(IV)
s^c_{yjmp}	-1.353***		
	(0.039)		
	[0.007]		
Product specialization		0.237***	0.389*
		(0.043)	(0.155)
		[0.008]	[0.008]
Controls	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes
N	2,911,210	4,529,585	2,911,210
R^2	0.534	0.036	0.125
Within-R^2	0.164	0.000	0.000

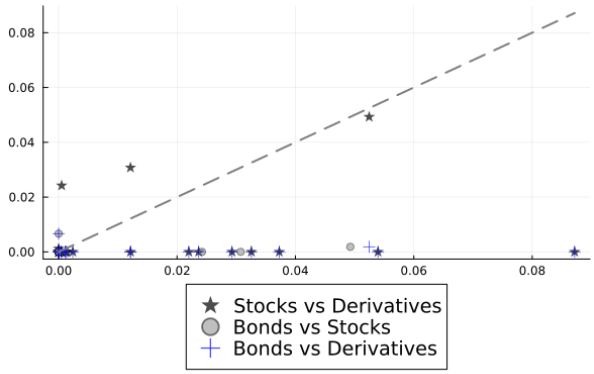
Appendix Table A10 show the IV estimation results for all trades on the derivatives market. Consider the LHS of 8. In column (First Stage), we show the first stage of the two stage least square estimator—regressing the market specialization score (1) on the fraction of all client-orders dealer j executes in market m in year y relative to other dealers, $s^c_{yjm} \in [0,1]$. In column (OLS) we present the OLS coefficient from regressing margins on market specialization, using trades for the dealer's own account for the stock market, and all trades for the derivatives market. In column (IV) we depict the corresponding IV estimate. The table on the RHS shows the analogous for product specialization, where the instrument is the fraction of all client-orders for product p dealer j executes in market m in year y relative to other dealers, $s^c_{yjmp} \in [0,1]$. In all regressions we include the same control variables and fixed effects as in regression (6): trade size, account-types for the exchange, dealer-types, date fixed effects and security-year-week fixed effects. We cluster standard errors at the daily-level and report conventionally computed robust clustered standard errors in round brackets, and wild-bootstrapped standard errors in squared brackets. The stars reflect to the larger standard errors.

Appendix Figure A1: Dealer market shares, $s_{yjm} \in [0,1]$, in an average year (parent-level)

(a) Banks



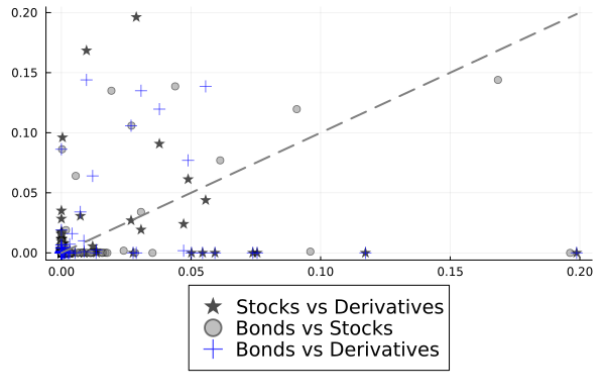
(b) High-frequency traders



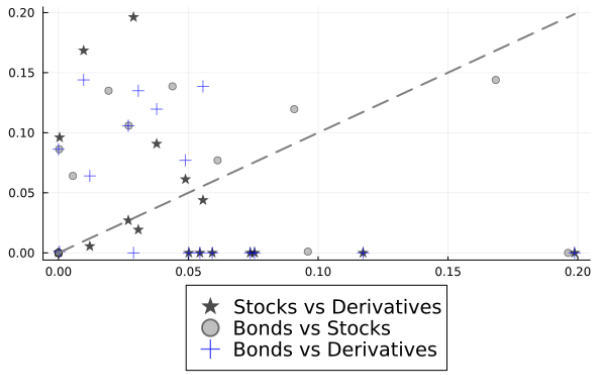
Notes: Appendix Figures A1a and A1b are analogous to Figure 2a but only includes banks, and high-frequency traders, respectively. It plots all bank/high-frequency dealer j's market shares for each market m, averaged across years. The stars show each dealer's stock market share on the y-axis and their derivatives market share on the x-axis; the circles show the stock versus bond market shares and the crosses the bond versus derivative market shares, on the y-axis and x-axis respectively.

Appendix Figure A2: Dealer market shares, excluding client-accounts, in an average year

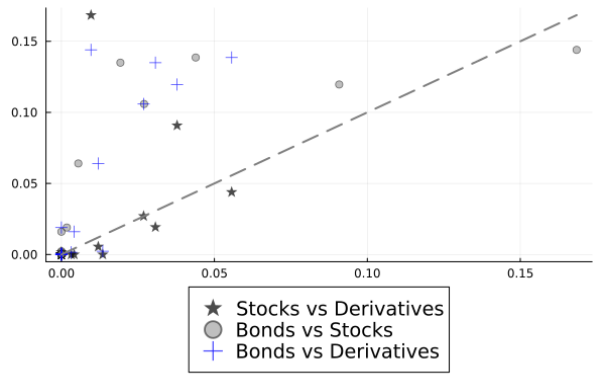
(a) All dealers



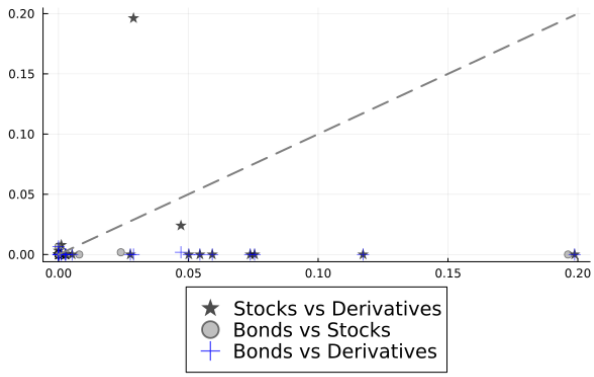
(b) Largest dealers



(c) Banks



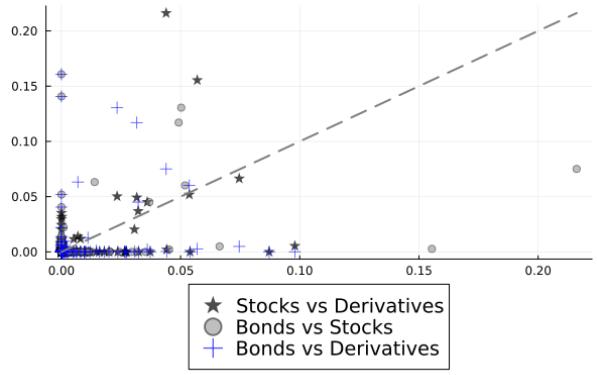
(d) High-frequency traders



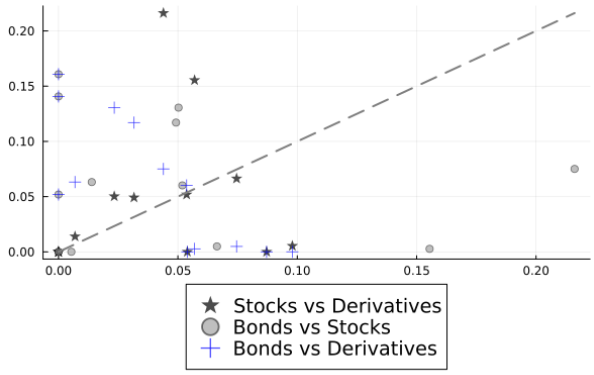
Notes: Appendix Figure A2a is analogous to Figure 2a but excludes trades for client accounts on the exchanges. It plots all dealer j's market shares (of trades for non-client accounts) for each market m, averaged across years. The stars show each dealer's stock market share on the y-axis and their derivatives market share on the x-axis; the circles show the stock versus bond market shares and the crosses the bond versus derivative market shares, on the y-axis and x-axis respectively. Figure zooms in on dealers who trade at least 5% of the non-client market share in one of the three markets. In Figures A2c and A2d we only consider banks, and high-frequency traders, respectively.

Appendix Figure A3: Dealer market shares, excluding client-accounts, in an average year (LEI-level)

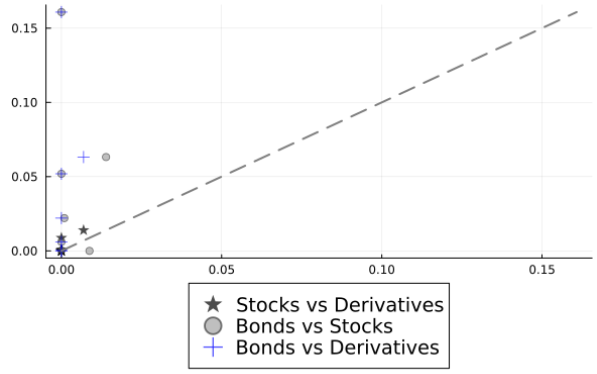
(a) All dealers



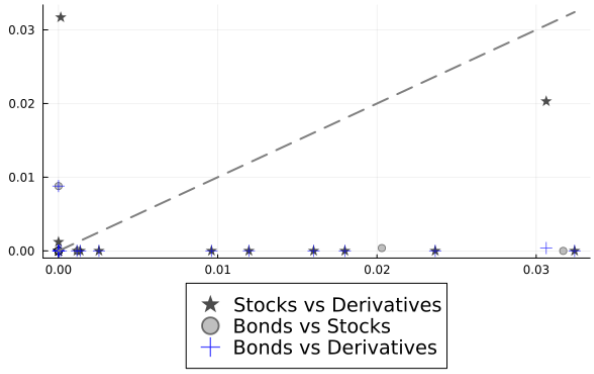
(b) Largest dealers



(c) Banks



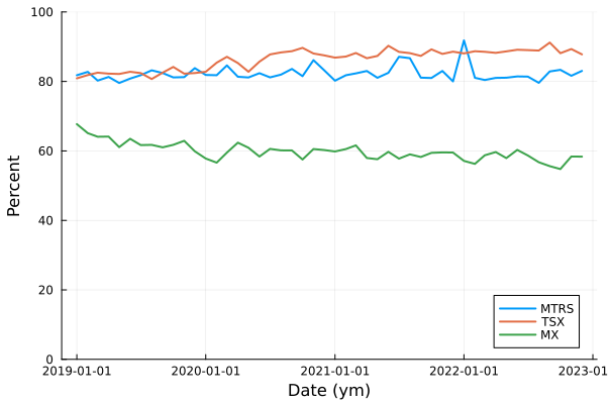
(d) High-frequency traders



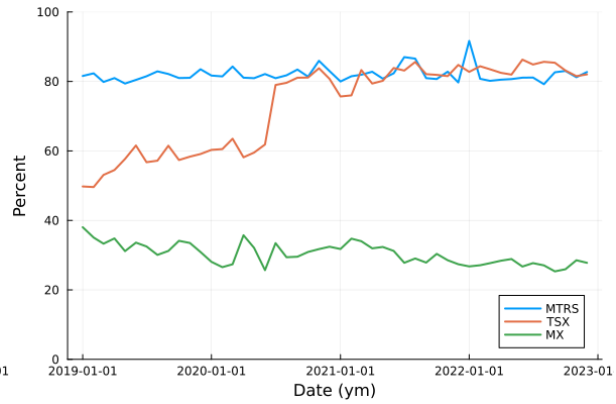
Notes: Appendix Figure A3 is analogous to Figure 2a but for dealers at the LEI-level rather than the parent-level. It plots all dealer j's market shares for each market m, averaged across years. The stars show each dealer's stock market share on the y-axis and their derivatives market share on the x-axis; the circles show the stock versus bond market shares and the crosses the bond versus derivative market shares, on the y-axis and x-axis respectively. Figure zooms in on dealers who trade at least 5% of the non-client market share in one of the three markets. In Figures A3c and A3d we only consider banks, and high-frequency traders, respectively.

Appendix Figure A4: Fraction of monthly trade volume by dealers active in all markets

(a) Including all trades



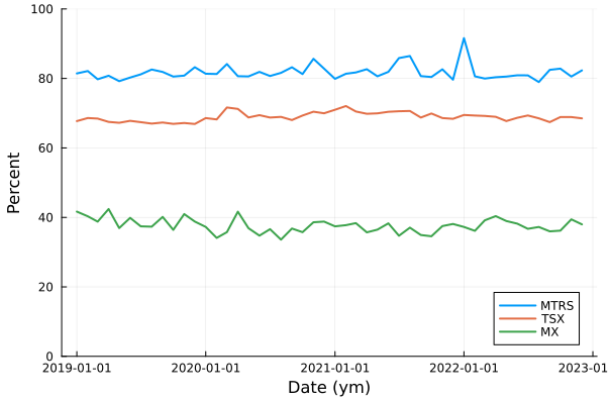
(b) Excluding client trades (for TSX/MX)



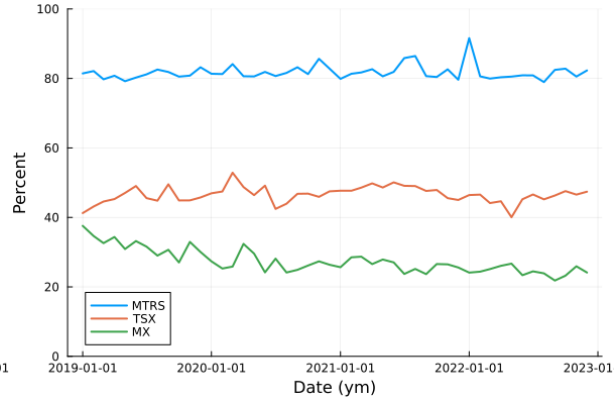
Notes: Appendix Figure A4a shows the fraction of monthly trade volume by those active in all markets, for each market (MTRS, TSX, MX). Appendix Figure A4b excludes trades for client accounts.

Appendix Figure A5: Fraction of monthly trade volume by primary dealers active in all markets

(a) Including all trades

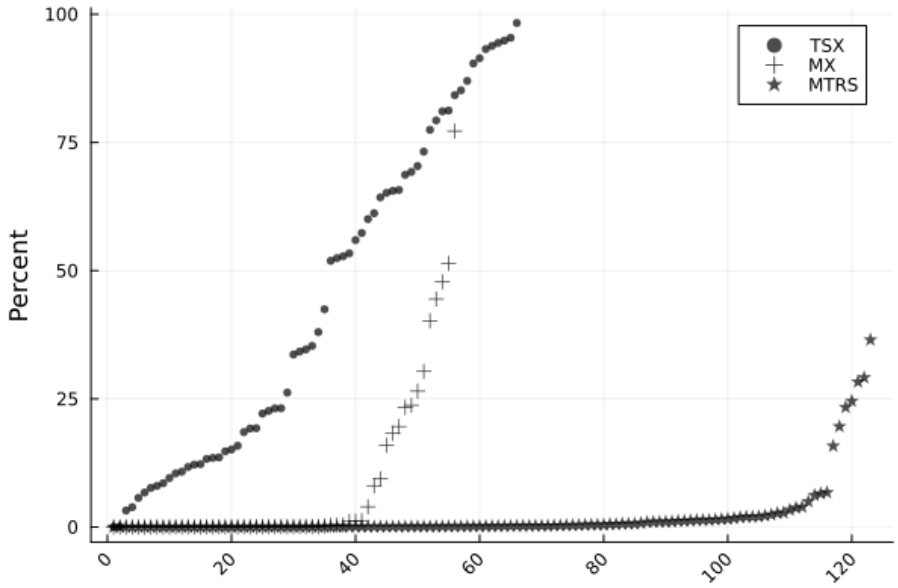


(b) Excluding client trades (for TSX/MX)



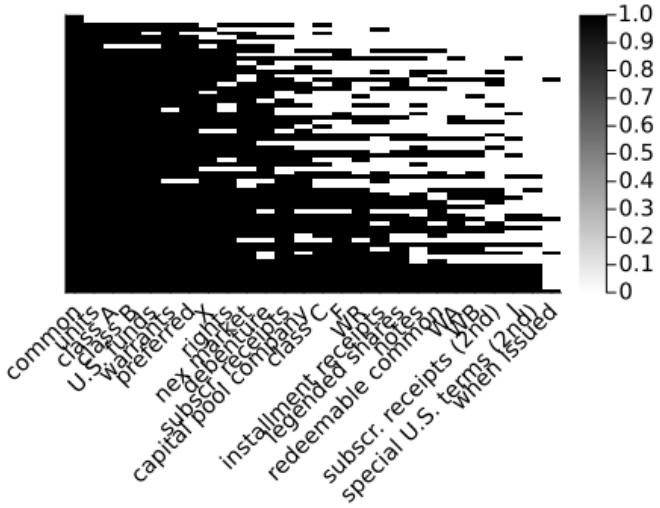
Notes: Appendix Figure A5a shows the fraction of monthly trade volume by primary dealers who are active in all markets. The graph implies that, in the bond market, essentially all dealers who are active in all markets are primary dealers. Roughly 68% of trade volume on TSX is executed by primary dealers who are active in all markets, which means that that 80%-68%=12% of trade volume is executed by dealers who are active across markets but are not primary dealers. Figure A5b excludes trades for client accounts.

Appendix Figure A6: Fraction of securities traded by each dealer out of all securities on TSX, MX, MTRS



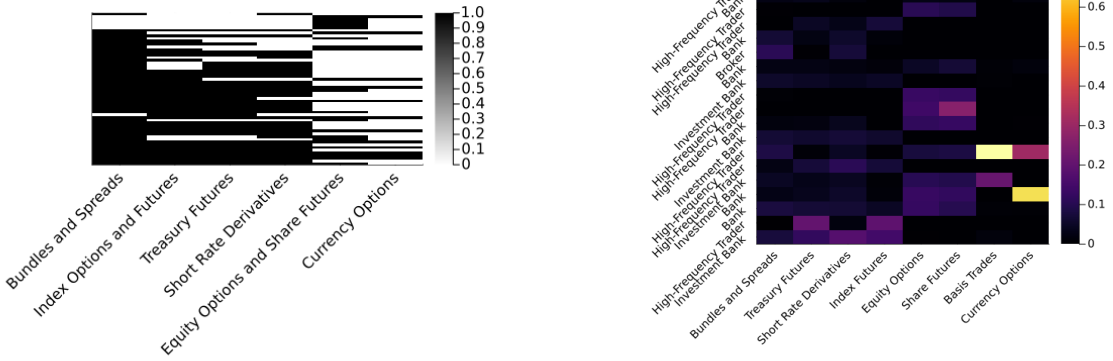
Notes: Appendix Figure A6 shows the fraction of symbols (x-axis) that each dealer trades (at the parent-level, on the y-axis) out of all traded symbols for each market. Since we sort by size within each market, the y-axis doesn't represent dealer-IDs that are common across markets. This is an alternative way of showing product differentiation: it is highest in the fixed-income market, and lowest on the stock exchange. The derivative exchange is in between.

Appendix Figure A7: Dealer presence across asset-types (defined by the symbol suffix)



Notes: Appendix Figure A7 shows whether each of the dealers is active (i.e., trades at least ones) in white, versus in-active in white for each asset-class, defined according to the symbol-suffix within the stock markets at the parent-level. Suffices are explained in Appendix Table A8.

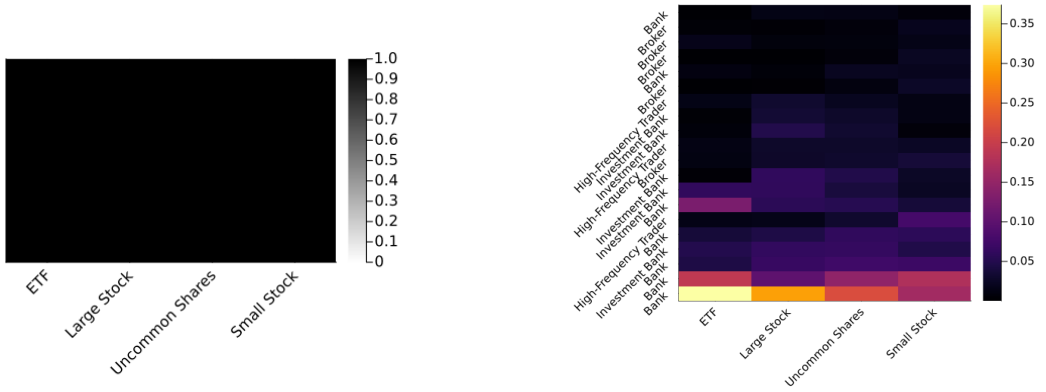
Appendix Figure A8: Dealer presence and market shares across all derivative products



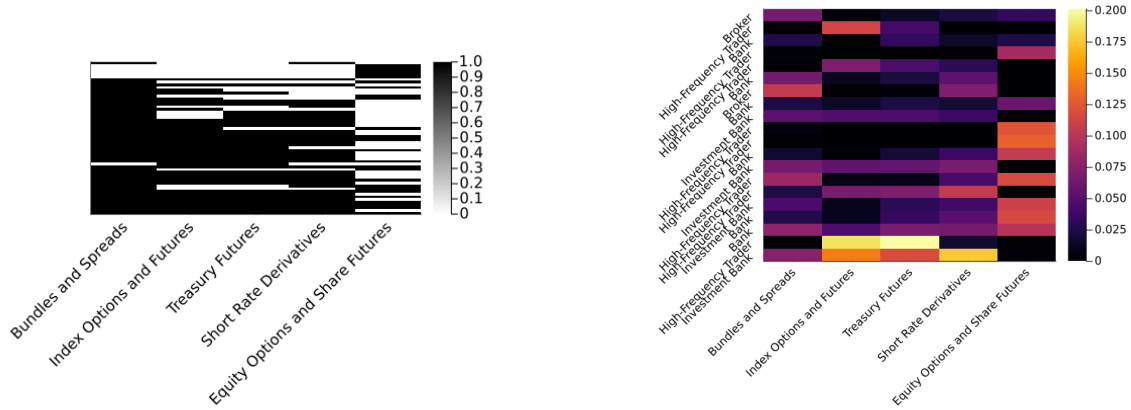
Notes: The RHS of Appendix Figure A8 is the analogue to Figure 4c but includes the small product "Currency Options". On the RHS we show whether dealers trade each product at least once in black. On the LHS we see the average annual product market shares of the largest dealers (on the LHS) for each market. In all figures, each row represents a dealer, sorted by total trade volume in the respective market. Dealers with the highest overall trade volume appear at the bottom, while those with the lowest appear at the top.

Appendix Figure A9: Dealer presence and market shares across products in the stock and derivative market (excluding market-maker accounts)

(a) Stock market



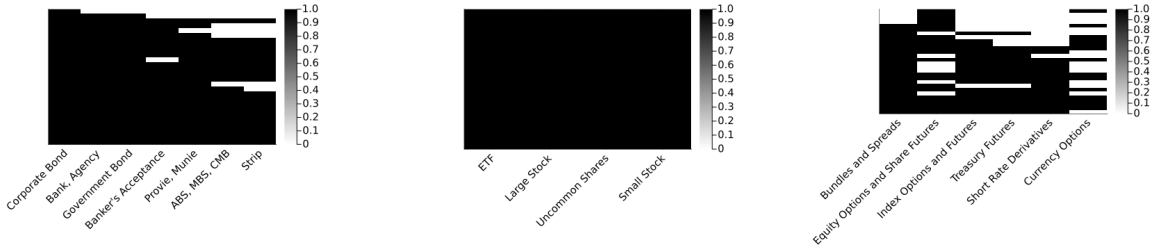
(b) Derivatives market



Notes: Appendix Figure A9 is similar to Figure 4 but excludes trades for market-marker accounts. On the RHS we show whether dealers trade each product at least once in black. On the LHS we see the average annual product market shares of the largest dealers (on the LHS) for each market. In all figures, each row represents a dealer, sorted by total trade volume in the respective market. Dealers with the highest overall trade volume appear at the bottom, while those with the lowest appear at the top.

Appendix Figure A10: Dealer presence across products of dealers present in all markets

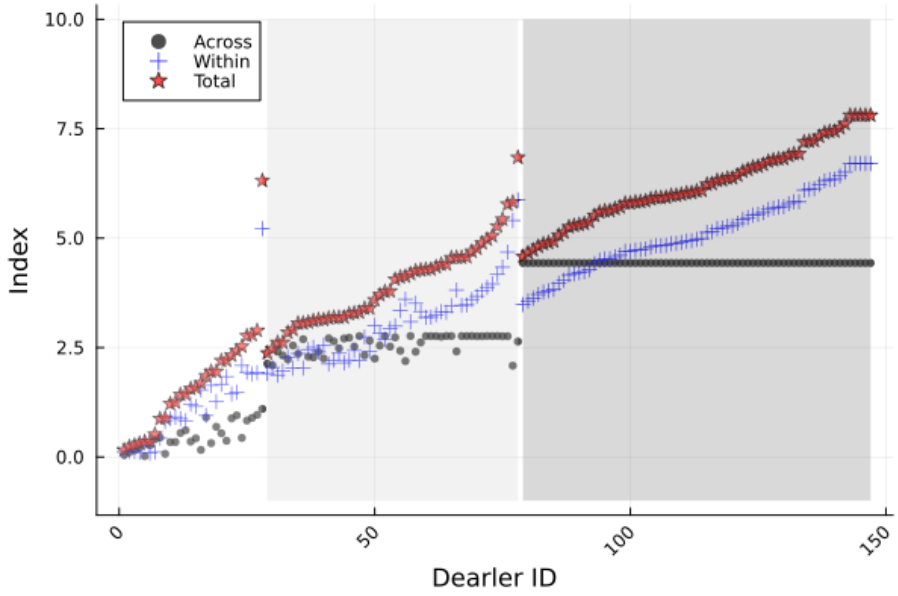
(a) MTRS (b) TSX (c) MX



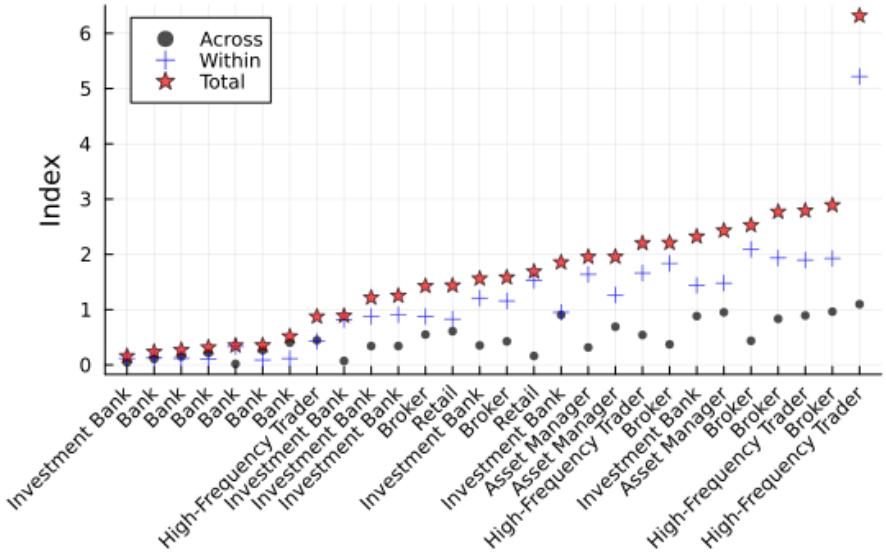
Notes: Appendix Figure A10 is the analogue to Figure 4, but includes the small product “Currency Options" for MX. It shows whether dealers who trade in all markets trade a product at least once in black, versus not in white within the bond (MTRS), stock (TSX), and derivatives market (MX).

Appendix Figure A11: Within-and cross-market segmentation indices

(a) Indices of all dealers

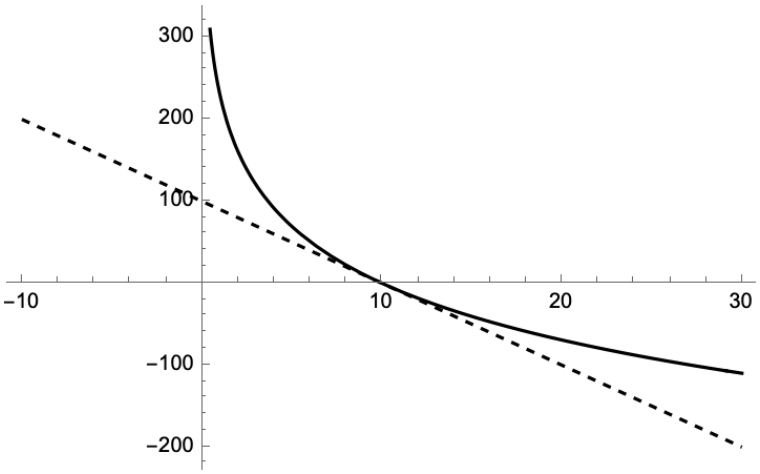


(b) Indices of largest dealers



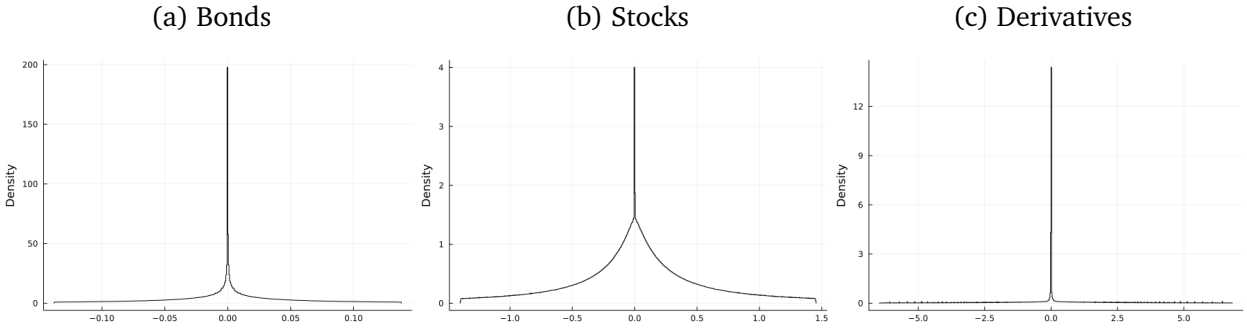
Notes: Appendix Figure A11a shows the within-market, across-market, and total adjusted Theil index for each dealer ID (at the parent-level). In the white area are dealer's who are active in all markets, in the light gray area are dealers who are active in only two markets, and in the darker gray shaded area are dealers who are active in only one market. The punishment term is 5; the maximal across market index is 4.431, and the maximal within market index is 8.82 for the stock market, 10.1234 for the bond market, and 10.6133 for the derivatives exchange. Figure A11b zooms in on the dealers who are active in all three markets.

Appendix Figure A12: Margins



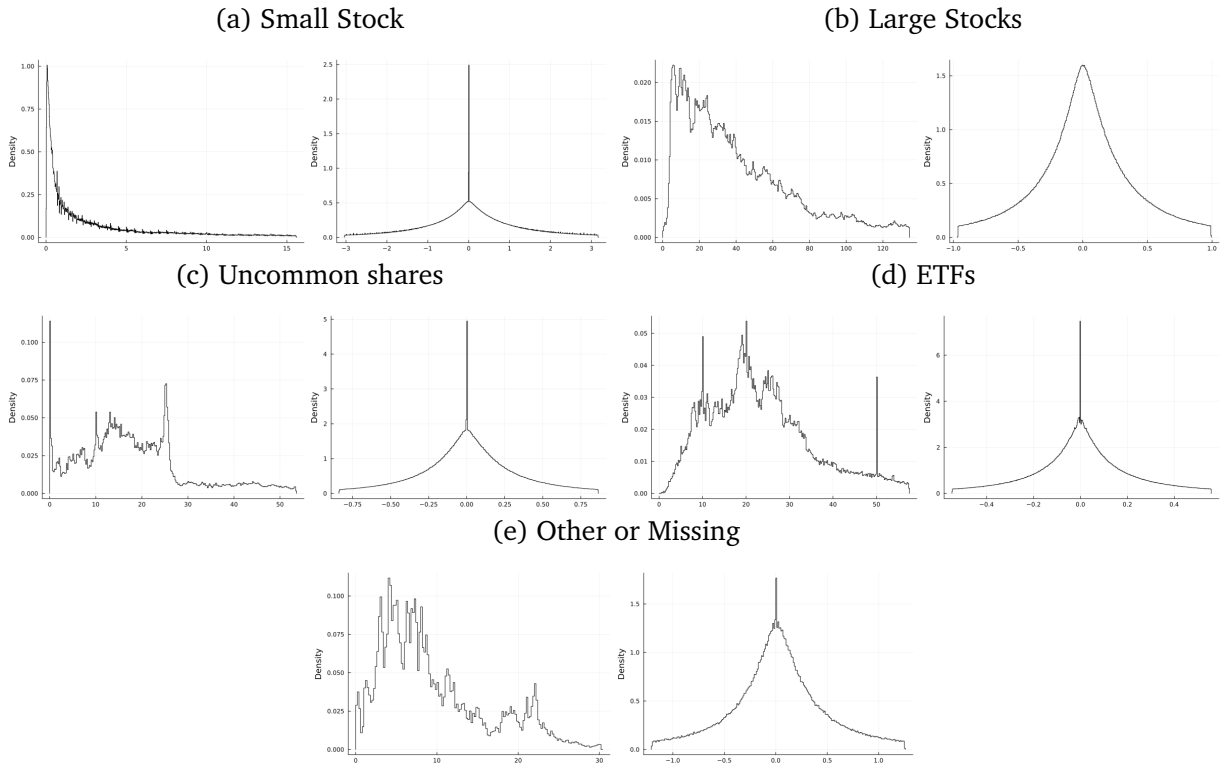
Appendix Figure A12 shows the margin (3) for an average price of 10 in black, and the linear approximation in dashed lines.

Appendix Figure A13: Margin distribution in each market



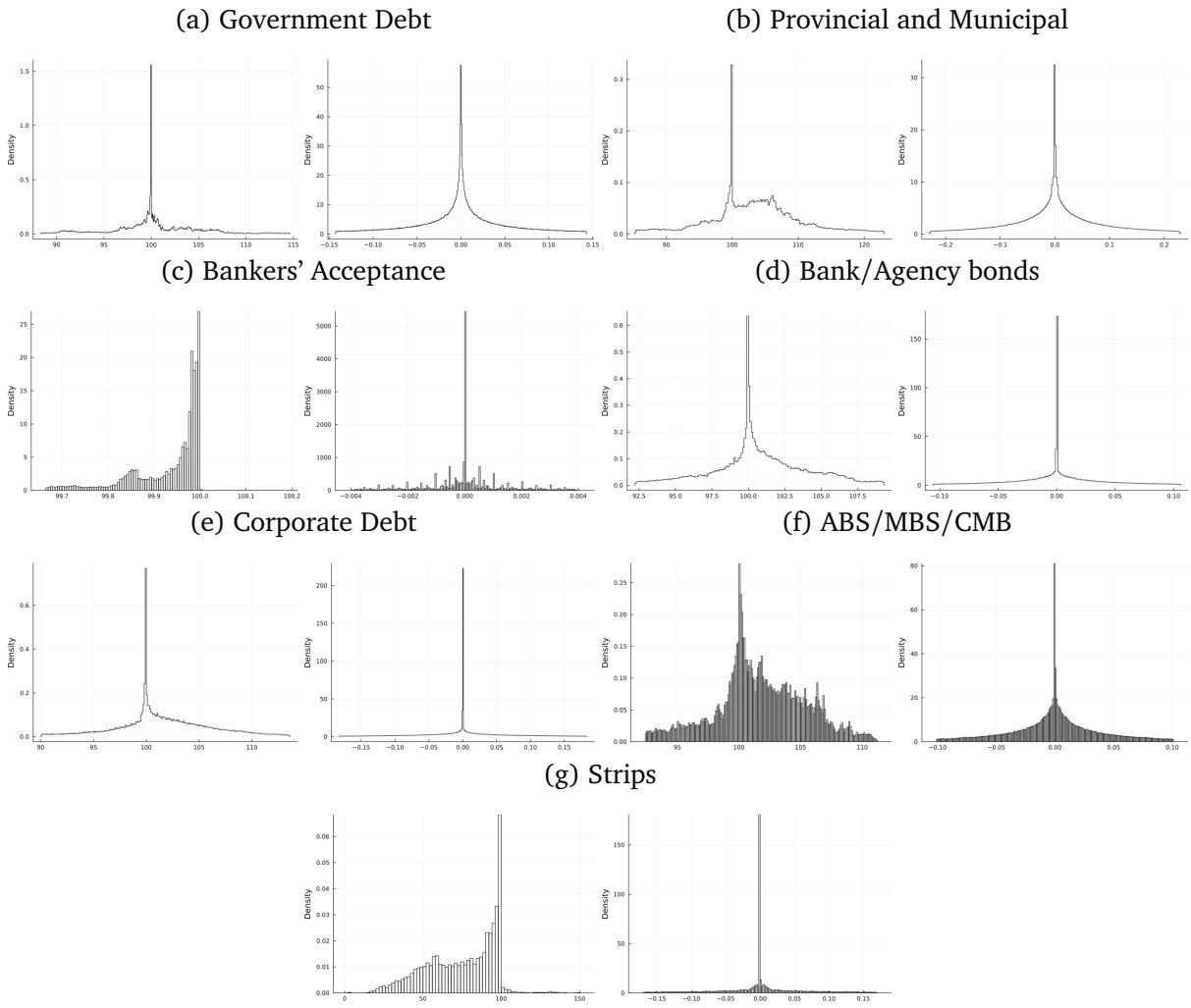
Appendix Figure A13 shows the distribution of our margin measure (3), which approximates how much less (more) a trader paid compared to the average price for a security in a day when buying (selling) for for each market. We exclude outliers, which are outside of the interquartile range. The median (average) margin is 0% for bonds, 0.005% for stocks, and 0% (0.18%) for derivatives. The standard deviation of margins is 1.59 for bonds, 2.03 for stocks, and 12.50 for derivatives. In comparison, the median trade price is C$100.29 for bonds, C$12.27 for stocks and C$1.1 for derivatives. The standard derivation in prices is 12.11 for bonds, 87.18 for stocks, and 157.35 for derivatives (where most of the variation is coming from the cross section of derivative contracts).

Appendix Figure A14: Price and margin distribution for equity products



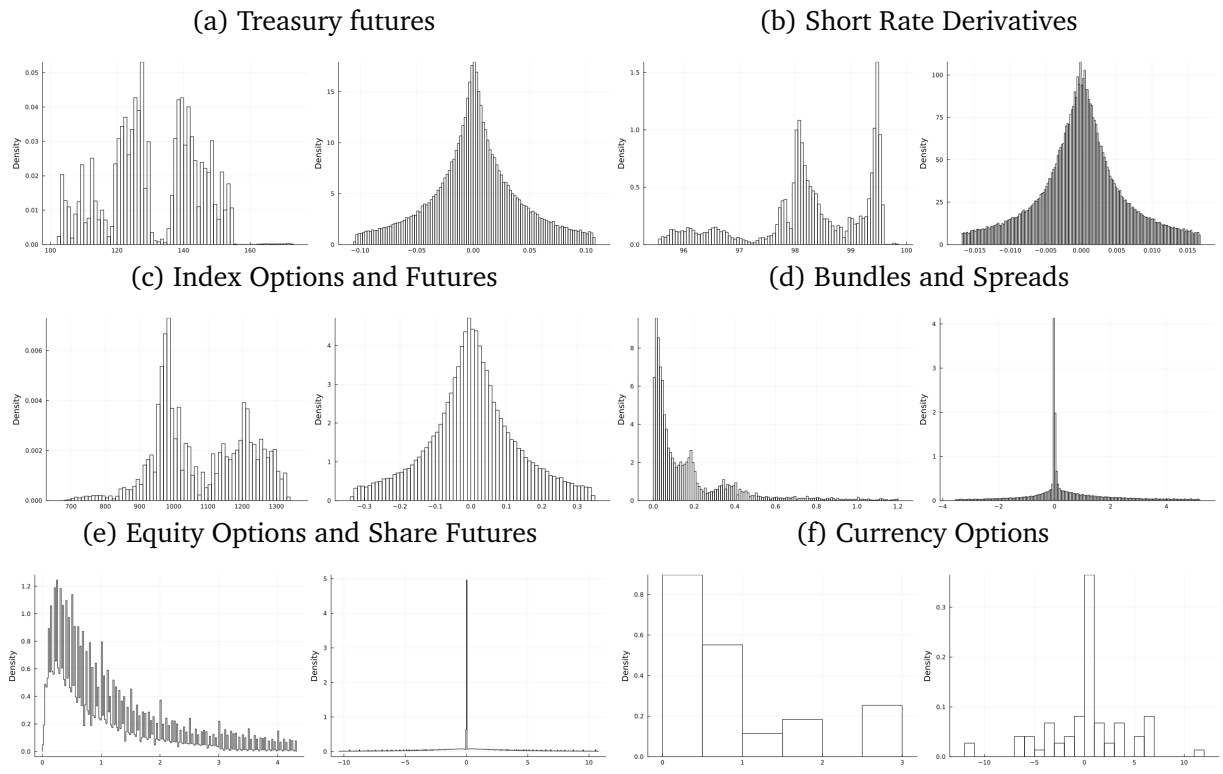
Notes: Appendix Figure A14 shows density histograms of prices for each product on TMX, excluding observations outside of the inter-quartile range.

Appendix Figure A15: Price and margin distribution for bonds



Notes: Appendix Figure A15 shows density histograms of prices for each product on the fixed-income market, excluding observations outside of the inter-quartile range.

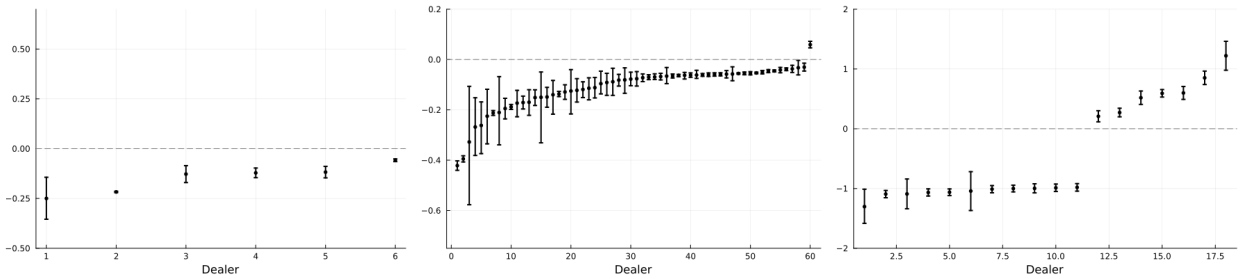
Appendix Figure A16: Price and margin distribution for derivatives



Notes: Appendix Figure A16 shows density histograms of prices for each product on MX, excluding observations outside of the inter-quartile range.

Appendix Figure A17: Dealer coefficients for dealers who exclusively trade in one market

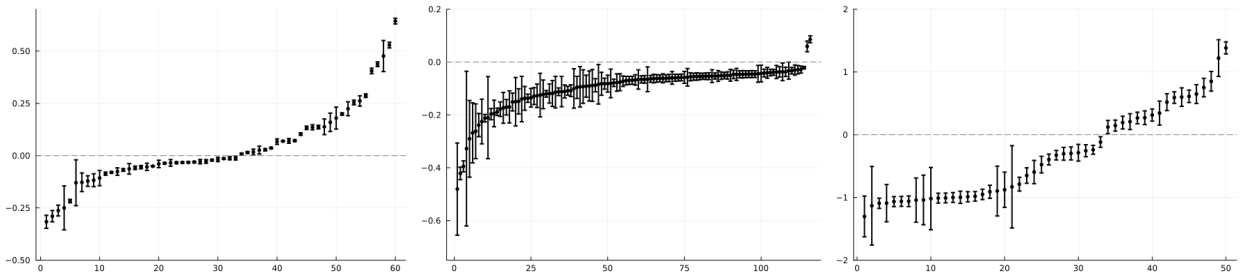
(a) Stock market (b) Bond market (c) Derivative market



Notes: Appendix Figures A17a shows the dealer coefficients and 95% confidence intervals, which are obtained via WCR bootstrapping, when regressing margins (3) of a trade on indicator variables for each dealer that is only active on stock exchanges, and not on the bond market, (at the LEI-level) in addition to control variables (trade-size, the account-type, security-week and day fixed effects). Figure A17b shows the analogue for the fixed-income market, where we replace the account-type with a variable that indicates the type of trade. Figure A17c shows the analogue for the derivatives market. In all graphs, we exclude dealer coefficients that aren't significantly different from zero at a significance level of 5% according to bootstrapped and conventional inference to be conservative. We sort coefficients from small to large. Therefore, the x-axis are not comparable across markets, since they don't reflect the dealer's IDs.

Appendix Figure A18: Robustness—Dealer coefficients that are statistically different from zero at 5% significance level (conventional inference)

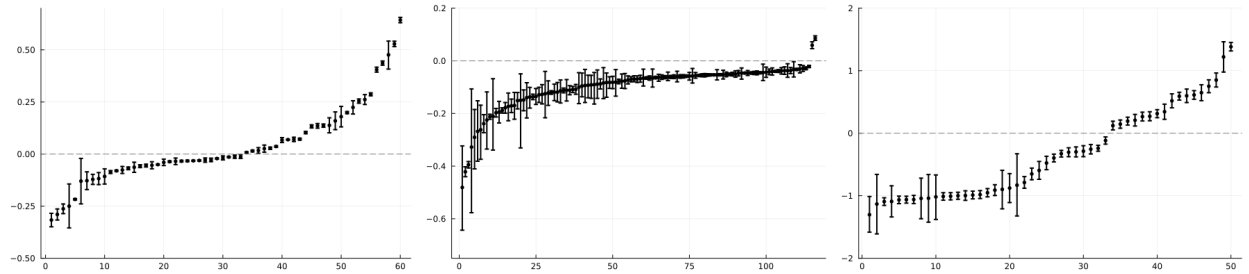
(a) Stocks (b) Bonds (c) Bonds



Notes: Appendix Figure A18 is the analogue to Figure 6 but with confidence intervals that are computed in the conventional way (without bootstrapping). For bonds, where clusters are most uneven in size, conventionally computed standard errors and confidence intervals differs slightly from bootstrapped confidence intervals—confirming expectations. Figure A18a shows the dealer coefficients when regressing margins (3) of a trade on indicator variables for each dealer active on the stock exchanges in addition to control variables (trade-size, the account-type, security-week and day fixed effects). Figure A18c shows the analogue for the derivatives exchange. Figure A18b shows the analogue for the fixed-income market, where we replace the account-type with a variable that indicates the type of trade (dealer-dealer, dealer-client, dealer-broker). In both graphs we exclude dealer coefficients that aren't significantly different from zero at a significance level of 5% according to bootstrapped and conventional inference. We sort coefficients from small to large. Therefore, the x-axis are not comparable across markets, since they don't reflect the dealer's IDs.

Appendix Figure A19: Robustness—Dealer coefficients that are statistically different from zero at 5% significance level (parent-level)

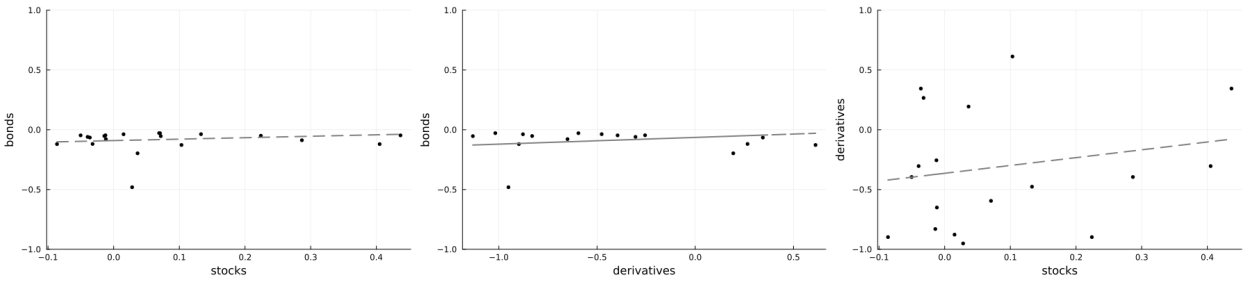
(a) Stocks (b) Bonds (c) Bonds



Notes: Appendix Figure A19 is the analogue to Figure 6 but aggregating dealer LEIs to the parent-level. Figure A19a shows the dealer coefficients when regressing margins (3) of a trade on indicator variables for each dealer active on the stock exchanges in addition to control variables (trade-size, the account-type, security-week and day fixed effects). Figure A19c shows the analogue for the derivatives exchange. Figure A19b shows the analogue for the fixed-income market, where we replace the account-type with a variable that indicates the type of trade (primary dealer/broker with non-primary dealer/non-broker, on-primary dealer/non-broker with non-primary dealer/non-broker, or primary dealer/broker with primary dealer/broker). In both graphs we exclude dealer coefficients that aren't significantly different from zero at a significance level of 5%. We sort coefficients from small to large. Therefore, the x-axis are not comparable across markets, since they don't reflect the dealer's IDs.

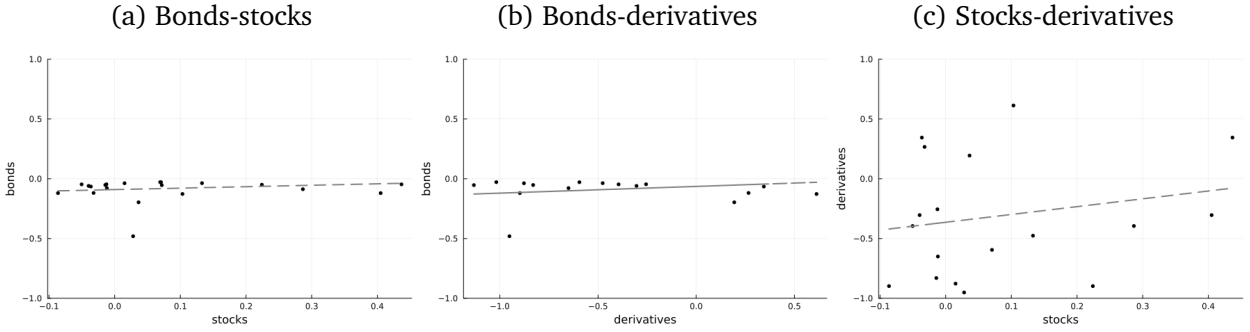
Appendix Figure A20: Robustness—Cross-market correlation between dealer coefficients of dealers active in all markets (conventional inference)

(a) Bonds-stocks (b) Bonds-derivatives (c) Stocks-derivatives



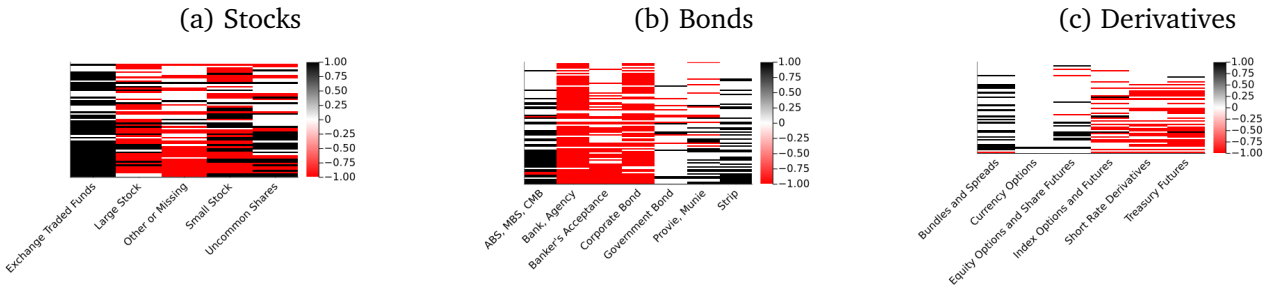
Notes: Appendix Figure A20a is analogous to Figure 7a but with confidence intervals computed in the conventional way without bootstrapping. It shows the within-dealer correlation of coefficients in the bond (y-axis) versus stock market (x-axis), (b) and (c) show the correlation for the other two market pairs. We exclude dealer coefficients that aren't significantly different

Appendix Figure A21: Robustness—Cross-market correlation between dealer coefficients of dealers active in all markets (parent-level)



Notes: Appendix Figure A21 is analogous to Figure 7a but uses dealer LEIs at the parent-level. It shows the within-dealer correlation of coefficients in the bond (y-axis) versus stock market (x-axis), (b) and (c) show the correlation for the other two market pairs. We exclude dealer coefficients that aren't significantly different from zero at a significance level of 5%.

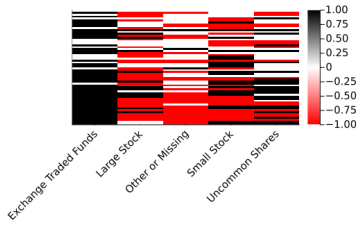
Appendix Figure A22: Dealer specialization across products within a market — 2022



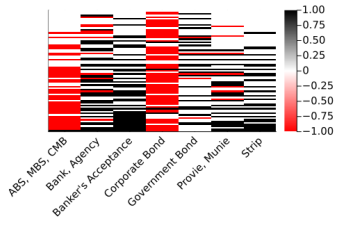
Notes: Appendix Figure A22a visualizes the dealer-product coefficients, β_{jp}, from regression (7)—which regresses trade margins (3) on indicator variables for each dealer-product combination plus control variables (trade-size, the account-type, security-year-week and day fixed effects)—using data from 2022. A row in each heatmap correspond to a dealer j, a column corresponds to a product p. When the corresponding β_{jp} is positive and statistically significant from zero at a 5% significance level, a p-j cell is black; if it is negative it is red; and empty if the coefficient is not statistically different from zero. Figure A22b and A22c show the analogue for bonds and derivatives. For bonds, the baseline β_{pj} coefficient is a large primary dealer trading government bonds; for stocks it's that bank trading large stocks, and for derivative it's that bank trading Treasury futures. In all graphs, dealers are sorted according to their trade-volume, with the dealer trading the most in the given market being at the bottom, and the dealer trading the least at the top. Standard errors are clustered at the daily-level.

Appendix Figure A23: Robustness: Dealer specialization across products within a market (parent-level) — 2022

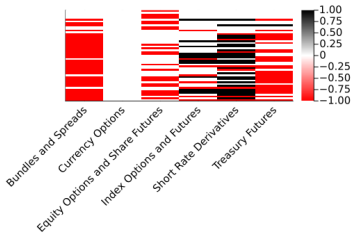
(a) Stocks



(b) Bonds

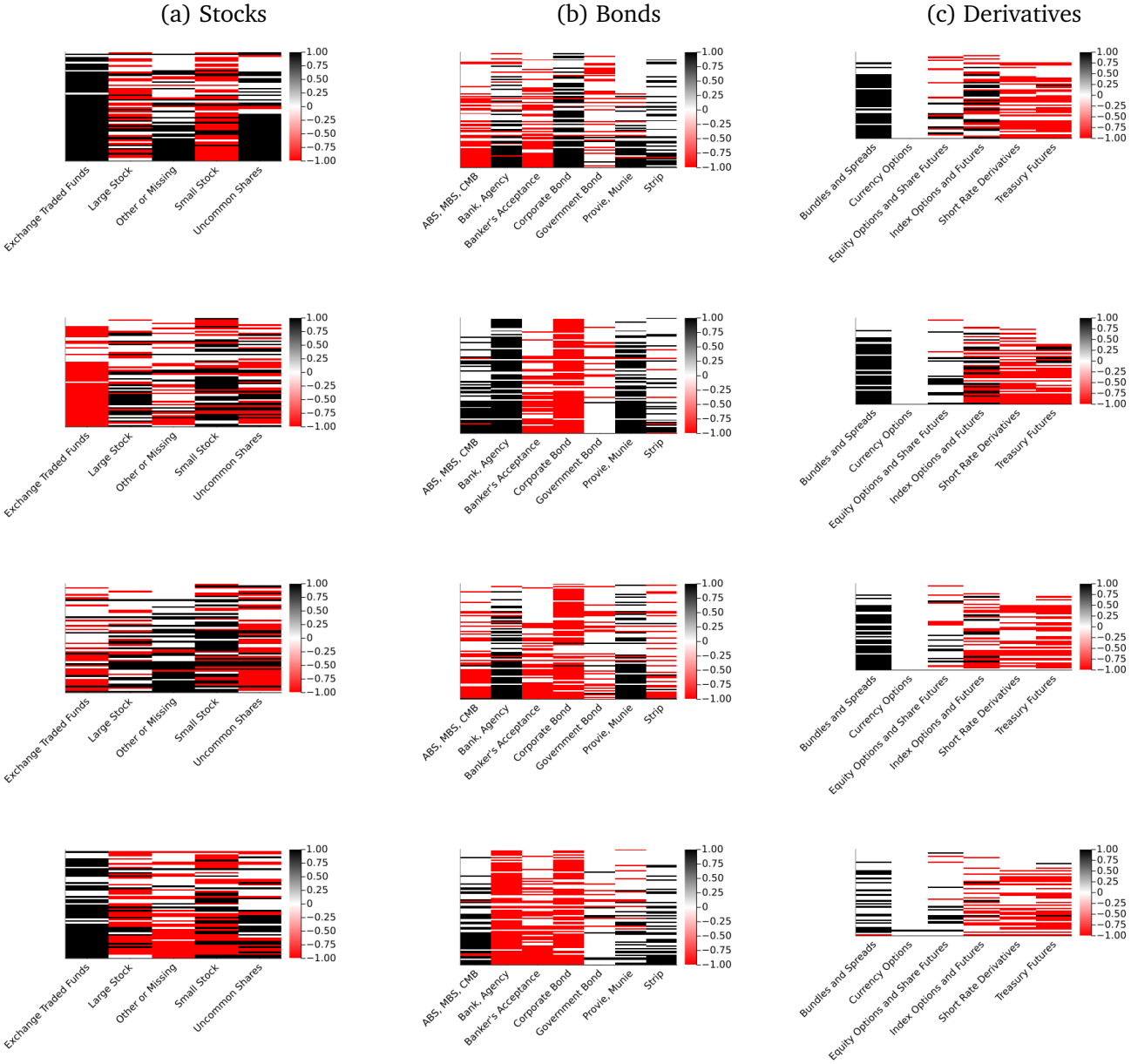


(c) Derivatives



Notes: Appendix Figure A23 is analogous to Appendix Figure A22 but when aggregating dealers to the parent-level. Appendix Figure A23a visualizes the dealer-product coefficients, β_{jp}, from regression (7)—which regresses trade margins (3) on indicator variables for each dealer-product combination plus control variables (trade-size, the account-type, security-week and day fixed effects)—using data from 2022. Each row in a heatmap correspond to a dealer j. A column corresponds to a product p. When the corresponding β_{jp} is positive and statistically significant from zero at a 5% significance level, a p-j cell is black; if it is negative it is red; and empty if the coefficient is not statistically different from zero. Appendix Figures A23b and A23c show the analogue for bonds and derivatives. For bonds, the baseline β_{pj} coefficient is a large primary dealer trading government bonds; for stocks it's that bank trading large stocks, and for derivative it's that bank trading Treasury futures. In all graphs, dealers are sorted according to their trade-volume, with the dealer trading the most in the given market being at the bottom, and the dealer trading the least at the top. Standard errors are clustered at the daily-level.

Appendix Figure A24: Dealer specialization across products within a market — 2019, 2020, 2021, 2022



Notes: Appendix Figure A24 is analogous to Figure A22, but for the other years in our sample. We note that dealer coefficients vary across years. However, the main takeaway that no dealer outperforms across products is robust for all years. Appendix Figures A24a visualizes the dealer-product coefficients, β_{jp}, from regression (7)—which regresses trade margins (3) on indicator variables for each dealer-product combination plus control variables (trade-size, the account-type, security-week and day fixed effects) using data from 2019, 2020, and 2021, respectively. Each row in a heatmap correspond to a dealer j. A column corresponds to a product p. When the corresponding β_{jp} is positive and statistically significant from zero, a p-j cell is black; if it is negative it is red; and empty if the coefficient is not statistically different from zero. Appendix Figures A24b and A24c show the analogue for bonds and derivatives. For bonds, the baseline β_{pj} coefficient is a large bank trading government bonds; for stocks it's that bank trading large stocks, and for derivative it's that bank trading Treasury futures. In all graphs, dealers are sorted according to their trade-volume, with the dealer trading the most in the given market being at the bottom, and the dealer trading the least at the top.

Trade-Off? What Trade-Off: Informative Prices without Illiquidity*

Thierry Foucault **Kostas Koufopoulos** **Roman Kozhan**

February 5, 2025

Abstract

Private information production in financial markets enhances asset price informativeness, aiding efficient decision-making. Investors pay for information to profit from trading, creating a trade-off between market informativeness and illiquidity costs. Using a mechanism design approach, we show price informativeness can be achieved without illiquidity, at a cost equal to producing information. This mechanism incentivizes efficient information production, avoiding the inefficiency of profits obtained at the expense of less informed investors.

Keywords: asymmetric information, optimal mechanism, information production, initial public offering.
JEL Classification: G20; G32; D82.

*Thierry Foucault is with HEC Paris, e-mail: foucault@hec.fr. Kostas Koufopulos is with the University of Sussex, e-mail: kkoufopoulos@gmail.com. Roman Kozhan is with Warwick Business School, University of Warwick, e-mail: roman.kozhan@wbs.ac.uk.

1 Introduction

An important role of financial markets is to produce information about asset payoffs. This information can then be used by decision makers (e.g., firms' managers) for making more efficient investment decisions, thereby increasing firm value. However, investors' incentives to produce information derives from the profits that they can make at the expense of less informed investors (see, e.g., Grossman and Stiglitz, 1980). This happens because of the prevalence of pooling equilbria that necessarily arise as a result of chosen mechanisms. Thus, asymmetric information makes financial markets less liquid, which lowers asset values. Consequently, there is a trade-off between the benefits of more informative prices and market liquidity.

Financial markets should therefore be designed to best solve this trade-off, that is, to maximize price informativeness while minimizing illiquidity due to adverse selection costs. In this paper, we use a mechanism design approach to study this question. For concreteness, we consider an entrepreneur (the "issuer") with one asset. The payoff of this asset can be high or low and the entrepreneur does not have the expertise to discover what is the exact realization of the payoff. To do so, he can sell a fraction of the asset to investors who have the ability to discover its payoff by collecting additional data. Doing so is costly and uncertain: with some probability, no information can be discovered about the payoff. The entrepreneurs' expected profit from the sale of the asset is equal to the proceeds from the sale plus a gain proportional to the reduction in the uncertainty on the asset payoff (e.g., this could be the gain derived from investing in other projects whose payoffs are correlated with the asset payoff). The entrepreneur chooses to sell a fraction of the asset if the maximal value of this expected profit exceeds the expected payoff of the asset (that is, the entrepreneur's outside option is to do nothing).

The entrepreneur's objective is to design the issue to maximize her expected profit. As all investors are rational and competitive, all costs borne by investors are ultimately passed back to the entrepreneur. Thus, the entrepreneur's expected proceeds from the issue cannot exceed the expected payoff of the asset net of information acquisition costs borne by investors. However, they can be less than this upper bound.

Indeed, to incentivize some investors to pay the information cost, the issuer must either pay them directly or let them earn profits at the expense of investors who do not buy information (e.g., as in Rock (1986) or Holmström and Tirole (1993)). In the first case, the issuer faces an agency problem (investors may misreport the information that they obtain or not pay the cost of information acquisition). To satisfy incentive constraints, the issuer might have to leave informational rents to investors. In the second case, uninformed investors will pass expected losses ("adverse selection costs") to the issuer by discounting the price at which they buy the asset. In sum, the entrepreneur faces both agency and adverse selection costs and seeks to design the issue to minimize these costs.

In this setting, we show that there is a mechanism that makes the entrepreneur's expected profit arbitrarily close to the maximum expected profit she can expect (the one obtained in the absence of agency and adverse selection costs). The mechanism has two stages. In the first stage, investors are sequentially offered the possibility to buy two derivatives securities, one that pays only if the asset payoff is high and one that pays only if it is low. If an investor refuses to participate, she retains the possibility to participate to stage 2. The entrepreneur optimally decides when to stop stage 1 and move to the second stage in which he sells the issue at a fixed price, after announcing publicly the outcome of the first stage (that is, the number of investors who participated to this stage, the number of derivatives sold and the type of derivatives traded). The entrepreneur chooses (i) derivatives' prices in stage 1, (ii) the payoff of each derivative, (iii) the number of investors participating to stage 1 (when to stop), (iv) the investors who can participate to stage 2 (he can exclude some of the investors who participated to stage 1) and (v) the price of the issue.

We show that the entrepreneur can design the derivatives (their payoff and price) in such a way that an investor who participates to stage 1 finds optimal to (i) produce information and (ii) select the derivative security that truthfully reveals the asset payoff if she learned this payoff. Moreover, if an investor does not discover information, she optimally abstains from buying or selling a derivative. Given these choices, the first investor who buys a security fully reveals the payoff of the asset. Thus, to minimize information acquisition costs, it is optimal for the issuer to stop

stage 1 as soon as one investor trades a derivative security and moves to stage 2. In this stage, the entrepreneur sells the asset at a price equal to its payoff.

As stage 1 takes place sequentially, the entrepreneur and investors become increasingly pessimistic about whether information about the asset payoff exists as the number of investors contacted to participate to stage 1 increases. Intuitively, this makes the cost of incentivizing information production higher over time because investors increasingly expect to pay the search cost without discovering information. Thus, to be incentivized to pay the information acquisition cost, investors must expect an increasingly higher payoff from the derivatives, which is costly to the issuer. As a result, unless information is available with probability 1, the entrepreneur optimally stops stage 1 at some point even if no investor bought a derivative. In this case, no information is produced in stage 1. Anticipating this outcome, some investors might refuse to participate to stage 1 and acquire information before participating to stage 2. However, we show that the entrepreneur can optimally avoid this outcome by pushing further the moment at which she stops stage 1.[1] In this way, the entrepreneur avoids underpricing the issue to induce uninformed investors to participate.

When information is produced in stage 1, the entrepreneur realizes the gains of obtaining information without paying illiquidity costs due to the risk of adverse selection for uninformed investors. When information is not produced, the entrepreneur does not obtain gains associated with information production but she can issue shares at the average payoff of the asset, that is, without paying illiquidity costs due to informed investors participating to the issue.[2] For these reasons, the expected profit of the entrepreneur with this mechanism is arbitrarily close to the one she can obtain when there are no agency and adverse selection frictions. The entrepreneur just needs to compensate investors who search for information.

Our paper relates to two strands of the literature. First, it relates to the literature on the informational benefits of financial markets for firms. These benefits can stem

[1] Intuitively, the entrepreneur delays the closure of stage 1 until the likelihood that information exists is so small that the expected profit from informed trading in stage 2 is less than the information cost.

[2] In this case, the entrepreneur is indifferent between issuing shares or not.

from the use of information in stock prices for contracting (see, for instance, Holmström and Tirole (1993)) or for making investment decisions (e.g., Edmans, Goldstein, and Jiang (2015); see also Bond, Edmans, and Goldstein (2012) and Goldstein (2022) for surveys)). In some papers in this literature, firms choose the fraction of shares to issue facing a trade-off between the benefits of informative prices (the gain associated with using the information in prices) and illiquidity costs (see, for instance, Holmström and Tirole (1993), Subrahmanyam and Titman (1999), Faure-Grimaud and Gromb (2004) or Foucault and Gehrig (2008). However, we are not aware of papers that seek to analyze how firms should optimally design the sale of shares to investors when they face this trade-off.

The literature on initial price offerings has analyzed the sale of shares to the public using a mechanism design approach (see, for instance, Beneviste and Wilhelm (1990) or Biais, Bossaerts, and Rochet (2002)). However, in most the literature on this topic, informed investors are supposed to be exogenously endowed with private information and firms do not derive gains from the information produced during their price offering. Exceptions are Sherman (2005) and Sherman and Titman (2002) and our framework is closely related to their modeling approach (in particular, the information structure is identical). However, they do not consider the possibility of using a sequential mechanism with two stages as we do. As discussed at the end of our paper, this possibility makes the issuer better off (that is, the mechanism considered in our paper dominates that considered in Sherman (2005) and Sherman and Titman (2002)).

2 The Problem: Illiquidity versus Informativeness under Asymmetric Information

In this section we first illustrate the tension between illiquidity and informativeness of the trading process using a standard modeling approach for the sale of a risky asset. Importantly, we assume that some investors have private information about the payoff of the asset. This information is exogenous. Hence, to obtain information, the issuer just needs to incentivize these investors to reveal their information, not to produce

it. We relax this assumption in the next section, which constitutes the core of our contribution.

The model is as follows. One agent owns $Q+N$ shares that are claims on the payoff of a risky asset of which it wishes to sell Q shares. The payoff of the asset (per share) is v_H with probability μ or v_L with probability $(1-\mu)$. There are $H+I$ potential buyers (henceforth investors), where I is the number of investors with information about the payoff of the asset. These investors perfectly know the realization of v while the remaining investors only know the distribution of v. Each investor can buy only up to one share and $Q < H$. Thus, the asset seller does not need participation of informed investors to execute her trade. The seller cannot observe who is informed and who is not (or cannot price discriminate based on investors' types).

There are several possible interpretations of this set-up. First, one can interpret the asset seller as a firm selling shares to the public in an initial price offering (IPO). This is our leading example and, for this reason, we refer to the seller as the issuer. Alternatively, one can see the seller as an entrepreneur selling a fraction of its stake to venture capitalists or business angels.

One can consider several ways to organize the sale of the asset. We first contrast two methods. The first is such that the issuance process fully reveals the payoff the asset but it results in underpricing due to adverse selection. The second is such that there is no underpricing because it excludes participation from informed investors. However, as a result, the issuance process provides no information about the asset payoff. These are just manifestations of the trade-off between illiquidity, due to adverse selection, and illiquidity.

In the first method, the issuer sets a price p_{issue} and investors decide whether they want to participate or not at this price. If there is excess demand, the issuer allocates shares pro-rata to each investor willing to buy one share at p_{issue}. This is a fixed price offering, as in the Rock (1986)'s model. For the issue to succeed, the issuer must guarantee the participation of uninformed investors. Suppose that $v_L < p_{issue} < v_H$ and consider a situation in which it is optimal for each uninformed

investor to buy one share at this price. At this price, each informed investor finds it optimal to buy one share if $v = v_H$ and to abstain otherwise. Thus, when $v = v_H$, each uninformed investor only receives $q_u(v_H) = \frac{Q}{H+I}$ shares (pro-rata rationing), while when $v = v_L$ each uninformed investor receives $q_u(v_L) = \frac{Q}{H}$. Thus, the expected profit of uninformed investors is:

$$E(q_u(v)(v - p_{issue})) = \mu q_u(v_H)(v_H - p_{issue}) + (1 - \mu)q_u(v_L)(v_L - p_{issue}). \quad (1)$$

To guarantee the participation of uninformed investors (which is necessary for the issue to succeed) and maximize the proceeds of the issue, the issuer must choose the largest price such that $E(q_u(v)(v - p_{issue})) \geq 0$, which is the price solving $E(q_u(v)(v - p_{issue})) = 0$. Thus, the issuing price is:

$$p^*_{issue} = \beta v_H + (1 - \beta)v_L,$$

with $\beta = \frac{\mu H}{H+(1-\mu)I}$. As $I > 0$, $\beta < \mu$ and therefore $p_{issue} < E(v)$.

Thus, the issue must be underpriced for it to succeed. Note that in this case, the issuing price does not reveal information about v since it is identical whether informed investors participate or not in the issue. However, total demand in the issue fully reveals the asset payoff. Thus, the trading process fully reveals investors' private information about the payoff of the asset. However, this information is obtained by the issuer at the cost of underpricing (illiquidity).

Now consider a more complex method for issuance. With this method, the issuer is allowed to make the issuance price contingent on demand. Specifically, let D be the total demand in the issue and consider the following price schedule posted by the issuer:

$$p_{issue} = \begin{cases} v_H + \epsilon, & \text{if} \quad D > H \quad \text{and} \quad \epsilon > 0, \\ E(v), & \text{if} \quad D \leq H. \end{cases} \quad (2)$$

In this case, the following decisions for investors form a Nash equilibrium: (i) informed investors do not participate, (ii) uninformed investors offer to buy 1 share. To see that this is an equilibrium, consider informed investors first. As the issuing price is

always strictly larger than v_L, it is never optimal for an informed investor to buy when $v = v_L$. When $v = v_H$, if an informed investor buys, she expects total demand to exceed H and therefore the price to be $v_H + \epsilon$. Thus, not participating is a best response to the issuer's price schedule and uninformed investors' strategy. Given that informed investors never participate, uninformed investors anticipate that they will receive $q_u = q_u(v_H) = q_u(v_L) = \frac{Q}{H}$ whether $v = v_H$ or $v = v_L$ and that total demand will always be $D = H$. Thus, their expected profit is:

$$E(q_u(v)(v - p_{issue})) = q_u \left(\mu(v_H - p_{issue}) + (1 - \mu)(v_L - p_{issue})\right) = 0.$$

Thus, uninformed investors are indifferent between participating or not, and participation is therefore a best response to the issuer's price schedule. The issuer cannot do better intuitively since any price larger than p_{issue} cannot satisfy uninformed investors' participation constraint. Thus, this equilibrium maximizes the expected proceeds for the issuer. However, ex-post, the issue price and total demand are completely uninformative since they are identical whether the payoff of the asset is high or low. This issuance method avoids underpricing (illiquidity) by removing adverse selection, at the cost of informativeness. This is again a manifestation of the standard trade-off.

We refer to this second mechanism as the "no-informed trading" mechanism. It is optimal for the issuer (it maximizes the expected proceeds from the sale of the asset) if the latter does not derive any benefit from the information produced during the issuance process. However, if it does (e.g., it could use the information for making new investments) and if this benefit is large enough, the first method can dominate the second. However, we show below that there is another mechanism that (i) avoids underpricing and (ii) is fully revealing. Thus, this mechanism eliminates the trade-off between illiquidity and informativeness and dominates the two previous methods. We refer to this mechanism as the "divide and conquer" mechanism.

In this mechanism, the issuance process is organized in two stages. In the first stage, investors are contacted sequentially and offered the possibility to buy 2 derivative contracts from the issuer whose payoffs are contingent on the realization of the fundamental value v, when this is finally observed. The first contract, labelled C_L

pays $F + \epsilon$ if $v = v_L$ and zero otherwise, where $F, \epsilon > 0$ are some predetermined positive values. The second contract, labelled C_H pays $F + \epsilon$ if $v = v_H$ and zero otherwise. All derivative contracts expire right after the end of the trading round after the fundamental value of the asset is observed. The price of each contract is F. The first stage stops when one investor has decided to buy one of the contract or when all investors have been contacted.

The issuer reveals the outcome of this stage to all investors and then move to the second stage. In the second stage the underwriter allocates the Q shares among the remaining $H + I - 1$ investors at $p_{issue} = v_L$ if the investor participated in the first stage has chosen C_L and $p_{issue} = v_H$ if the investor participated in the first stage has chosen C_H. If no investor participates to the first stage then the underwriter cancels the issue and no allocations is done (this never happens in equilibrium).

We say that this mechanism induces full revelation if (i) only informed investors buy in stage 1 and (ii) an informed investor selects contract C_w when she observes that $v = v_w$ for $w \in \{L, H\}$.

Proposition 1. *If the issuer chooses* $F > \max\{\frac{(1-\mu)}{\mu}, \frac{(\mu)}{(1-\mu)}\}\epsilon$ *and* $\epsilon > 0$*, the mechanism induces full revelation and the expected proceeds per share from the asset sale are* $E(v) - \epsilon$*. In this case, the Nash equilibrium of the issuance process is that (i) uninformed investor do not trade in stage* 1*, (ii) the first informed investor contacted by the issuer in stage 1 chooses contract* C_w *when she observes that* $v = v_w$ *for* $w \in \{L, H\}$ *and (iii) the issuing price in the second stage is* $p_{issue} = v$ *so that investors participating in the second stage are indifferent between buying the asset or not.*

With this divide and conquer mechanism, the expected proceeds from the sale of the asset, $E(v) - \epsilon$, are arbitrarily close to the maximum expected proceeds, $E(v)$ because ϵ (the net payoff of the derivative contracts) can be arbitrarily small (it just needs to be strictly positive). Thus, the mechanism optimally solves the trade-off between illiquidity and informativeness in the framework considered so far. Intuitively, the mechanism separates the problem of incentivizing informed investors to reveal

their private information from the problem of incentivizing uninformed investors to participate to the issue. In the two previous methods, these problems are bundled. The divide and conquer mechanism separates them and creates competition between informed investors to minimize the cost of information revelation for the issuer. Intuitively, this cost cannot be less than the cost of information production, as otherwise informed investors would not participate. However, so far, we assume that investors bear no information production cost (they are exogenously endowed with information). Thus, intuitively, by inducing competition among informed investors, the issuer can drive the cost of information revelation, ϵ, to almost zero. The condition $\epsilon > 0$ is just to make an informed investor strictly better off participating to the first stage.

In the next section, we show that this insight still obtains when information production is endogenous. In this case, the divide and conquer mechanism must not only induce investors with information to truthfully reveal their information via their choice in stage 1 but also induce them to produce information. We discuss the robustness of the mechanism to more general environments in Section 6. We think that the divide and conquer mechanism offers an interesting benchmark for assessing frictions in real-world financial markets. This mechanism solves the trade-off between informativeness and illiquidity. Hence, if it is not used, it must be that other frictions make it impractical or dominated by other mechanisms. Identifying reasons why divide and conquer mechanisms are not used more is then the question.

3 Costly Information Production

In this section, we now consider the case in which information production is endogenous. This case is more complex because the mechanism that is used by the issuer must incentivize investors both to reveal their information if they have some and to produce information. To consider this issue, we modify the previous framework as follows. As before, there are $H + I$ investors and each investor can buy only up to one share and H and I are large relative to Q. Only I investors have the ability to produce information about the asset. However, at the beginning of the asset sale, these investors have not yet information and must pay a cost to produce

it. We denote by $\mathcal{I}$ (resp., $\mathcal{H}$) the set of investors who (resp., don't) have the ability to produce information.

Information production is as follows. There is a probability $\pi \in (0,1)$ a probability that there information about the fundamental value of the firm. To produce information about v, an investor must pay a cost c without knowing whether information is available or not. After paying the cost c, if information exists, the investor is succesful, i.e., learns v perfectly with probability $\phi \in (0,1)$.Otherwise, that is, if the investor is unsuccessful or if information does not exist, the investor remains uninformed. Thus, the likelihood that an investor fails in producing information is $(1-\phi)\pi + (1-\pi) = (1-\phi\pi)$. Importantly failure to produce information does not imply that information does not exist since $\phi < 1$. To ensure that the information cost is not prohibitively high, we assume that $\frac{c}{\pi\phi} < Qv_L$ (that is, the expected cost of information acquisition is smaller than the value of the firm in bad state).

If instead the investor does not search for information, she remains uninformed and expects the value of the asset to be $v_U = E(v) = \mu v_H + (1-\mu)v_L$ (that is, she has access to the same information as the issuer and other uninformed investors). We assume that the issuer cannot acquire information.[3] This is a natural assumption since we want to analyze the trade-off between informativeness and illiquidity from the asset seller's viewpoint. If the asset seller could pay the cost of information, she would not need to incentivize information production in the first place.

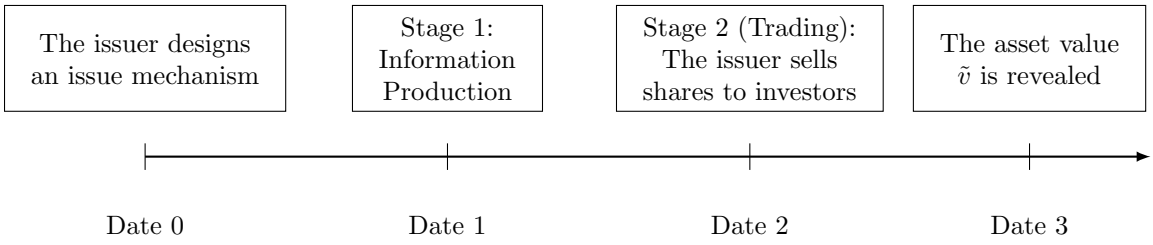


Figure 1. Timing of the model

Figure 1 presents the timing of the model. At date 0, the seller of the asset

[3]This does not mean that the issuer has no information. Indeed, one can assume that the issuer first collects information and arrives to an estimate of $E(v)$ for the firm. It just means that the cost of collecting incremental information is too high for the issuer.

designs and announces the mechanism that it will use to sell shares to investors. As explained below, this choice is made to maximize the proceeds from the sale and the "informativeness" of the sale. The mechanism is similar to the divide and conquer mechanism presented in the previous section. In the first stage ("information production"), the seller contacts investors sequentially and ask them to report their information about the asset. In the second stage ("trading"), the seller proceeds to the sale of the asset. In contrast to the divide and conquer mechanism presented before, in stage 1, investors directly report their information (or absence thereof) and receive a transfer from the issuer rather than pick a derivative. This difference is not important: The use of derivatives is just a way to implement the direct mechanism considered here. As explained in subsequent sections, the more substantial difference is that the mechanism must make sure that investor who reports information have indeed paid the cost of information production since their effort is not observable. Last, we assume that at some point in the future the fundamental value of the asset is realized, whether information production took place or not. This assumption plays a role in the design of the incentive mechanism considered in Section 5.[4]

We denote the price at which the asset is sold in stage 2 by p_{issue} and we denote by p_2 the price of the asset at date 2, just after stage 2. This price will depend on the information publicly available after stage 2 and therefore be different from the price at which the asset is sold at stage 2. For instance, if information is produced and fully revealed to market participants, p_2 will be v_H or v_L. However, this might not be the case for the price at which the asset is sold in stage 2, giving the rise to the possibility of underpricing or overpricing.

The seller's utility depends on her proceeds from the sale of the asset and the informativeness of the sale. Her proceeds are equal to $Qp_{issue} - C_{issue}$, where C_{issue} are total monetary transfers to investors participating to stage 1 (they can be zero or even negative; see below). The informativeness of the sale is measured by the residual uncertainty about the payoff of the asset after observing the outcome of stages 1 and 2. We denote the seller's information set at the end of stage 2 by Ω_2. It contains, for

[4]This is also the case in the divide and conquer mechanism considered in Proposition 1 since the payoff of the derivatives depends on the realization of the fundamental value.

instance, the reports in stage 1 and the price of the asset after stage 2, p_2. Residual uncertainty for a given realization of Ω_2 is measured by $Var(v \mid \Omega_2)$. The realized utility of the issuer is after the sale of the asset is therefore:

$$\Pi(p_{issue}, C_{issue}, \Omega_2)) = Qp_{issue} - C_{issue} - \gamma Var(v \mid \Omega_2), \tag{3}$$

where γ measures the utility gain for the seller from a marginal decrease in uncertainty about v after the sale of the asset. Parameter γ measures the importance of the informativeness of the mechanism for the seller. If $\gamma = 0$, the seller does not care about informativeness and, as we shall, see in this case she will organize the issue so that no information is produced. In this case, the illiquidity-informativeness trade-offis moot since information has no value. Thus, the more interesting case is $\gamma > 0$. As explained below, the seller designs the mechanism for selling the asset at date 0 to maximize $E(\Pi(p_{issue}, C_{issue}, \Omega_2)))$, the expected value of her realized utility after the issue.

3.1 Benchmark: Information Production is Observable

As a benchmark, we first consider the case in which the issuer can observe whether a given investor has the ability to produce information or not and that investors always truthfully report the outcome of their search for information. Moreover, we assume that the issuer can exclude informed investors from stage 2 (e.g., by using the no-informed mechanism described in Section 2). Thus, in this benchmark, we consider the case in which there is no moral hazard in stage 1 and no adverse selection in stage 2. In this case, the issuer's problem is to obtain information at the lowest possible expected cost.

In this case, the issuer faces no incentives compatibility constraints (investors don't need to be incentivized to report truthfully what they know). It must still design the issuing mechanism to guarantee participation by investors to each stage. This means in particular that, in stage 1, the issuer must compensate investors for their information production cost (as otherwise they would not produce information). Moreover, in stage 2, the issuer cannot sell the asset at a price larger than its expected

payoff conditional on the information produced during stage 1, as otherwise investors would not buy shares in Stage 2. Given this, the largest expected proceeds that the issuer can achieve are equal to $QE(v)$ minus the expected information acquisition costs for investors in stage 1. We show below that this is indeed the case. Moreover, the maximum expected utility achieved by the issuer in this case is an upper bound for its expected utility in the case in which the issuer does not observe whether investors acquire information because in this case the issuer face additional incentives compatibility constraints (see Section 5).

In stage 1, the issuer contacts investors with the ability to produce information sequentially, that is, investors in $\mathcal{I}$.[5] Each contacted investor optimally chooses to produce information or not and reports the outcome of her search to the issuer. If she chooses to produce information, the investor pays the information acquisition cost, observes the outcome of her search for information and finally reports a message $s \in \{H, L, U\}$ to the issuer, where $s = H$ means that the investor has discovered $v = v_H$, $s = L$ means that the investor has discovered $v = v_L$ and $s = U$ means that the investor has found nothing. To compensate the i^{th} investor, the issuer pays a fee f_{i,s_i} which can depend on the investor's report (s_i) and his position (i) in the queue of contacted investors.[6] If the investor chooses not to produce information, she receives no reward.

Importantly, this process brings information about whether information about v is available or not. Indeed, since the information is not present with certainty ($\pi < 1$), investors (as well as the issuer) update their beliefs about availability of information after every unsuccessful round of information acquisition. The probability that there is information available about the payoff of the asset conditional on observing $i-1$ uninformative signals in a row is:

$$\pi_i = \frac{(1-\phi)^{i-1}\pi}{(1-\phi)^{i-1}\pi + (1-\pi)}. \tag{4}$$

[5] Contacting investors in $\mathcal{H}$ is useless for the issuer since they cannot help the issuer to obtain information.

[6] We assume that investors know their position in the queue.

Observe that $\pi_1 = \pi$ and that π_i decreases with i. Thus, investors participating to stage 1 and the issuer becomes increasingly more pessimistic about the possibility of finding information as the length of stage 1 increases.

The i^{th} investor produces information if her expected reward exceeds the cost of information production, that is, if:

$$\pi_i \left[\phi\mu f_{i,H} + \phi(1-\mu) f_{i,L} + (1-\phi) f_{i,U}\right] + (1-\pi_i) f_{i,U} \geq c, \quad i \in \{1, ..., \tau\}, \quad (5)$$

This equation is the participation constraint of the i^{th} contacted investor in Stage 1. The L.H.S is the expected fee received by the investor producing information and the R.H.S is the cost of producing information. Thus, eq.(5) is the participation constraint of the i^{th} investor.

As π_i decreases over time when $\pi < 1$, there is a information production round K^* after which contacting subsequent investors to obtain information is not optimal. Thus, when the K^*th investor fails to find information, the issuer's expected utility is

$$QE(p_{issue}) - c - \gamma\mu(1-\mu)(v_H - v_L)^2.$$

If instead, the issuer contacts one extra investor and then moves to stage 2, his expected utility is:

$$QE(p_{issue}) - \gamma(1 - \pi_{K^*+1}\phi))\mu(1-\mu)(v_H - v_L)^2,$$

because $\Pr(\Omega_1 = U) = (1 - \pi_{K^*+1}\phi)$ in this case. By definition of K^*, this course of action must be dominated by moving to stage 2 not optimal if and only if $\gamma\pi_{K^*+1}\phi\mu(1-\mu)(v_H - v_L)^2 > c$. This implies that $K^* = K_{max}$.

Once an investor has found information, there is no incentive for the issuer to keep contacting investors in $\mathcal{I}$ since uncertainty about v has been fully resolved and the outcome of stage 1 is publicly announced. Thus, stage 2 should optimally stop when one investor reports $s = H$ or $s = L$. The issuer could also optimally stop when $s = u$ after many trials because inducing investors to produce information becomes

increasingly costly as i increases when $\pi < 1$ (see the participation constraint eq.(5)). Thus, we let K be the total number of contacted investors in stage 1 be another choice variable for the issuer. This number can be smaller or larger than I because one informed investor can be asked repeatedly to produce information. We denote by τ_{stop} the number of rounds in stage 1. This number is the minimum of K and the first time at which an investor finds information. For a given realization of τ_{stop}, the total cost of stage 1 for the issuer is therefore:

$$C_{issue} = \sum_{i=0}^{i=\tau_{stop}} f_{i,s_i} = (\tau_{stop} - 1) f_{i,U} + f_{\tau_{stop},s_{\tau_{stop}}}. \tag{6}$$

Observe that C_{issue} is random because the stopping time for stage 1 is random since whether investors discover or not information in stage 1 is random.

After stage 1 is completed, the issuer announces the outcome of this round and sets a price p_{issue} for the issue. The outcome, Ω_1 is H if one investor has reported $s = H$, L if one investor has reported $s = L$ and U otherwise. As we assume that an investor with information cannot participate to stage 2, we must have $p_{issue} \leq E(v \mid \Omega_1)$ to guarantee participation of uninformed investors to stage 2. Last, as the trading process in stage 2 is uninformative (since no informed investors participate to this stage), $\Omega_2 = \Omega_1$. Thus,

$$\mathrm{E}(Var(v \mid \Omega_2)) = \Pr(\Omega_1 = U)\mu(1-\mu)(v_H - v_L)^2, \tag{7}$$

where $\Pr(\Omega_1 = U)$ is the probability that no information is produced during stage 1.

Thus, for a given design of stages 1 and 2, we deduce from eq.(3) and eq.(7) that the expected utility of the issuer is:

$$\Pi(p_{issue}, \{f_{i,s_i}\}, K\}) = QE(p_{issue}) - E(C_{issue}) - \gamma \Pr(\Omega_1 = U)\mu(1-\mu)(v_H - v_L)^2. \tag{8}$$

At date 0, the issuer chooses $\{p_{issue}, \{f_{i,s_i}\}, K\}$ to maximize her expected utility, under the constraints that investors participate to stages 1 and 2. Thus, she solves

the following problem:

$$\Pi_{bench} = \max_{\{p_{issue}, \{f_{i,s_i}\}, K\}} \Pi(p_{issue}, \{f_i\}), \tag{9}$$

subject to the participation constraints:

$$\pi_i \left[\phi\mu f_{i,H} + \phi(1-\mu) f_{i,L} + (1-\phi) f_{i,U}\right] + (1-\pi_i) f_{i,U} \geq c, \quad i \in \{1, ..., \tau\}, \tag{10}$$

$$p_{issue}(s) \leq \mathrm{E}(v \mid \Omega_1) \tag{11}$$

for every $s \in \{H, L, U\}$. Observe that K affects the expected utility of the issuer because it determines the distribution of the stopping time. One can solve the problem in two steps. First, for a given $(p_{\text{issue}}, \{f_{i,s_i}\})$, one can solve for the optimal $K^*(p_{\text{issue}}, \{f_{i,s_i}\})$. Then, in a second step, one can solve for the $\{p_{\text{issue}}, \{f_{i,s_i}\}\}$ that maximizes: $\Pi_{\text{bench}}(p_{\text{issue}}, \{f_{i,s_i}\}, K^*(p_{\text{issue}}, \{f_{i,s_i}\}))$.

Define $K_{\max}$ to be the maximal i satisfying

$$\frac{c}{\pi_i \phi} < \gamma\mu(1-\mu)(v_H - v_L)^2. \tag{12}$$

We assume that this condition holds for $i = 1$, i.e., for $\pi_i = \pi$ (the case in which it does not is discussed below) so that $K_{max} > 1$. The solution to the issuer's problem in this case is as follows.

Proposition 2. *In the benchmark case, the issuer's optimal issuance strategy is as follows.*

1. *Stopping time: the issuer stops contacting investor as soon as it obtains a positive ($s = H$) or a negative ($s = L$) report or the number of rounds exceeds $K_{\max}$;*
2. *Fees: conditional upon observing $s_{i-1} = U$, the issuer sets $f_{i,L} = f_{i,H} = \frac{c}{\phi\pi_i}$ and $f_{i,U} = 0$.*
3. *Price: the issuer sells shares to Q investors (chosen randomly) at price $p_{issue} = E(v \mid \Omega_1$.*

4. Value of objective function:

$$\begin{aligned}\Pi^*_{bench} &= M + (Q+N)E(v) - cK_{\max}(1-\pi) - \frac{c\pi(1-(1-\phi)^{K_{\max}})}{\phi} \\ &- \gamma\mu(1-\mu)\left(1-\pi+\pi(1-\phi)^{K_{\max}}\right)(v_H - v_L)^2. \end{aligned} \tag{13}$$

Proof. See Appendix. □

Given our assumptions, it is straightforward that the issuer should sell shares in stage 2 at $p_{issue} = E(v \mid \Omega_1)$. A lower price would leave rents to investors while at a larger price investors would not buy shares. As a result, the issuer expects to sell shares at $E(v)$.

The information produced in stage 1 is useless to increase the proceeds from the issue because there is no adverse selection in stage 1. However, it is useful to reduce uncertainty about the payoff of the asset. In designing stage 2, the issuer trades-off the benefit of reducing uncertainty with the cost of producing information.

The issuer always sets its fees for information production so that the participation constraint of each investor contacted to produce information is binding. Thus, when an issuer contacts an investor, he expects to pay c to the investor. However, the issuer's optimal fee structure is to reward the investor only if the search for information is successful. Thus, it pays the investor more than c (in fact c/ϕ) when the investor is successful in finding information and nothing otherwise.

We have assumed that the issuer contacts investors sequentially one by one in stage 2. An alternative is to contact investors by batches of M_i investors in each round i. We call this the "batched process". the next proposition that the optimal size of a batch is $M_i = 1$ for each $i \in \{1, 2, ..., K^*\}$. Thus, the process we have considered so far is the optimal way to organize information production in stage 2.

Proposition 3. *In the batched procedure, the optimal size of a batch is $M_i = 1$ in any round. Thus, the sequential procedure where the issuer contacts exactly one investor per round is optimal for the issuer.*

Proof: see Appendix

The intuition is as follows. Suppose that the issuer deviates from the previous policy by contacting $M_1 > 1$ in the first batch. In this case, the issuer must pay $M_1 c$ for sure to all investors contacted in the first batch (as each must expect a payment of c to produce information) and the likelihood that none of these investors find information is $(1 - \pi) + \pi(1 - \phi)^{M_1}$. The likelihood of this event is identical to that if investors are contacted sequentially. However, in the latter, the expected payment to investors is strictly smaller than $M_1 c$ because there is the possibility that one investor finds information before all investors are contacted, in which case the issuer optimally stops the search for information. Last, conditional on none of the M_1 investors finding information, the continuation value for the issuer is exactly the same if he contacts the M_1 investors sequentially or not. Thus, the issuer is strictly better off not contacting the M_1 first investors in a batch. The same argument can show that this is also the case at any round.

In sum, Π^*_{bench} is the largest possible expected utility for the issuer. It serves as benchmark to measure the efficiency of the various mechanisms that the issuer can use in the more complex case in which (i) the issuer does not observe investors' information acquisition decision and the signals received by informed investors and (ii) the issuer cannot prevent investors from choosing to secretly produce information when contacted to participate to stage 2. In this case, the issuer faces a moral hazard problem in stage 1 and there is adverse selection in stage 2, which may force the issuer to sell the asset at a discount, as explained in Section XXX. One may think that these frictions will reduce the expected utility that the issuer can achieve due to the trade-off between liquidity and informativeness. However, in Section, we show how the issuer can design a mechanism that makes the issuer's expected profit arbitrarily close to Π^*_{bench}. This implies that this mechanism dominates any other mechanism that the issuer could use in the context of our model (in particular that proposed by Sherman and Titman (2002) in the same environment or modification of Rock (1986) model to costly information acquisition).

4 Mechanisms

In this section we describe several competing mechanisms that can be considered as candidates to implement the first-best allocation described in the benchmark model.

4.1 Fixed price mechanism (FP)

The pooling mechanism (**FP**) is an extension of the fixed price offering of Rock (1986) where the costly information has to be endogenously acquired. In this mechanism the issuer sets a pooling price p_{issue} and let investors decide whether they want to participate or not in the issue at this price. If there is excess demand, the issuer allocates shares pro-rata to each investor willing to buy one share at p_{issue}. Those investors who have the ability to search and acquire the information endogenously decide where or not to do so.

The optimal strategies of the issuer and the investors are as follows.

Proposition 4. *Under* **FP** *mechanism:*

- *the issuer offer the issue price with underpricing* $p_{issue} < E(v)$*;*
- *there is a number* $0 \leq K_{FP} \leq I$ *of investors who participate in the information production;*
- *the issuer objective function under this strategy is:*

$$\begin{aligned} \Pi_{FP} &= M + (Q+N)E(v) - cK_{FP} \\ &- \frac{\gamma\mu(1-\mu)\left(1-\pi+\pi(1-\phi)^{K_{FP}}\right)(v_H - v_L)^2}{\mu\left(1-\pi+\pi(1-\phi)^{K_{FP}}\right)+(1-\mu)}. \end{aligned} \tag{14}$$

Proof: See Appendix

4.2 No-information production mechanism (NI)

Now consider another mechanism in which the issuer sets a price of $v_H + \epsilon$ if the total demand in the IPO is strictly larger than H and a price equal to $p_{issue} = E(v)$

otherwise. Using the same argument as before with this mechanism, no investor searches for information. Thus, all investors are indifferent between participating or not and the case in which just H investors participate is an equilibrium. There also equilibria in which less than H but more than Q investors participate. In this "mechanism", the issuer gets an expected utility of:

$$\Pi_{NIM} = M + (N + Q)E(v) - \gamma\mu(1 - \mu)(v_H - v_L)^2. \tag{15}$$

5 Optimal mechanism

In this section we describe the sequential mechanism (**SEQ**) that implements an allocation provided in the benchmark model.

5.1 Description of the mechanism

As described in the benchmark case, the issuance is happened in two stages. At the beginning of date 1 investors to apply for Stage 1 trading. The issuer randomly chooses one of the applicants to trade in Stage 1. Before trading, the chosen investor can (but might choose not to) attempt to acquire information. During Stage 1 the investor can either buy one of the offered by the issuer derivative contracts whose payoffs are contingent on the realization of the firm value v or decide (upon observing the information acquisition process) not to trade derivatives. If the chosen investor decides not to trade in Stage 1, that investor is excluded from the allocation and the issuer may either re-open Stage 1 for the remaining investors or proceed to Stage 2.

In Stage 2, the issuer decides on the price of shares and equally allocates them among the rest of the investors who is willing to accept the offered price. The investors before deciding whether or not to accept the offered price in Stage 2 may privately attempt to acquire information. After observing investors actions during Stage 1 and the reported signal (if any), the issuer decides on price at which to allocate shares during Stage 2.

In Stage 1, the issuer issues derivative contracts whose payoffs are contingent on

the realization of the fundamental value v. Given that the firm value can take only two distinct values, it is sufficient to offer only two different contracts corresponding to each of the realization of the firm's value. The idea behind this is that an informed trader by choosing the specific derivative contract will reveal the information they possess about the realization. At the beginning of Stage 1, the issuer invites investors to apply for the allocation of the derivative contract on the "first come first served" basis (or alternatively, the issuer could randomly select an investor among those who applied). Application for Stage 1 trading is optional and each investor might choose not to apply and wait for Stage 2 allocation instead.

If an investor is selected to trade in round i of Stage 1, he has to pay the issuer a fixed fee F and then he gets the right to buy one contract of either C_H or C_L. The contract corresponding to the bad state C_L pays $F + f_{i,L}$ if $v = v_L$ at date 3 and zero otherwise; the contract corresponding to the good state C_H pays $F + f_{i,H}$ if $v = v_H$ and zero otherwise. The fee f_i is determined by the issuer before the issuance process, and depends in general on the information cost c, probability of successful information acquisition and the number of rounds in information acquisition.[7] It is designed to incentivize investors to acquire information (which is optional for them and they can choose not to pay information costs and not to acquire the information).

If the investor who has applied for Stage 1 and has been chosen to trade decides to report the neutral signal (either because he was unsuccessful in acquiring information or and he decided strategically to misreport and hide the information) does not trade any derivative contract and is excluded from the allocation. Exclusion of investors who refuse to trade in Stage 1 is needed to ensure the efficiency of the allocation and minimization of the cost of issuance. Suppose the investor receives a positive signal and strategically misreports the neutral signal betting on the issuer failing to acquire informative signal and offering the shares in Stage 2 at some average prices. In order to ensure truth telling and avoid this scenario the issuer has to offer higher compensation to those who disclose informative signal (by trading the derivative

[7]Each of these contracts can be replicated by issuing "butterfly spread" – a portfolio of call options written on the underlying asset, for example, C_L contract payoff is equivalent to the payoff of a long position in call option with strike price $f_{i,L} - F$, a short position in two call options with strike price v_L and a long position in a call option with strike price $f_{i,L} + F$ for given round i.

contract in Stage 1) which, in turn, increases the cost of issue. By excluding investors who declare neutral signal during Stage 1 eliminates this possibility as these investors will have no chance to exploit acquired information in the subsequent stage. On the other hand, investors who did not apply for Stage 1 trading can still participate in the following rounds of Stage 1 (should it have been announced) or being considered for allocation in Stage 2.

If during Stage 1 an investor bought one of the derivative contracts, the Stage 1 trading is ended and the issuer opens Stage 2 allocation. If the initial attempt to sell the derivative in Stage 1 fails, i.e., the investor who applied to trade derivatives decided not to close the trade (e.g., the signal received by the investor appeared to be neutral), then the issuer may call for the second round of application to trade in Stage 1. Any of the remaining investors (except those who have been excluded from the issue due to declaring neutral signal in one of the previous rounds) are allowed to participate. If Stage 1 results in unsuccessful trade after round K, the issuer terminates Stage 1 and proceed to Stage 2 allocation.

In Stage 2 the issuer allocates the Q shares among the remaining investors at price $p_{issue} = v_H$ if Stage 1 ends with the purchase of C_H contract, at price $p_{issue} = v_L$ if Stage 1 ends with the purchase of C_L contract, and at price $p_{issue} = \mu v_H + (1-\mu)v_L$ if Stage 1 ends with no transaction after $K_{\max}$ rounds. Each investor receives at most one share.

5.2 Issuer's objective and constraints

Similarly to our benchmark model, we define the issuer's problem as minimization of a separable function of accuracy of the price at Date 2, the expected amount of underpricing and expected cost of derivative trading (8). The main difference is that the issuer faces additional constraints relative to the benchmark model.

There are three main types of constraints that the optimal mechanism has to satisfy. The issuer needs to give investors the incentive both to buy the information and to report it accurately. As part of mechanism design problem, the issuer must design an allocation and pricing schedule that elicits accurate information from investors.

Since the issuer uses the reported information to price the issue, the pricing and allocation strategy must counteract investor incentives to withhold favorable information that will lead to a higher issue price. We will be considering Nash equilibria where, conditioned on the issuer's strategy, investors have an incentive to truthfully reveal their information, given their expectation that other investors will also report information accurately.

Let $R(s_i, \sigma)$ be the expected profit to an investor i who has been chosen to participate in the information acquisition process, receives signal s_i but decides to report the state σ instead (by means of choosing to trade the derivative C_σ, for $\sigma \in \{H, L\}$ or not to trade if $\sigma = U$). The assumption that the investors are excluded from the allocation when declaring neutral signal implies that $R(s_i, U) = 0$ for any $s_i \in \{H, L, U\}$. In equilibrium, investors are induced to report their information truthfully, which implies that the following truth-telling constraints must be satisfied:

$$R(s_i, s_i) \geq R(s_i, \sigma) \text{ for all } s_i, \sigma \in \{H, L, U\}. \tag{16}$$

It should be noted that the cost of acquiring information does not affect the information reporting conditions, since it is a sunk cost by the time the investor decides what signal to report. On the other hand, whether or not the investor plans to accurately report the signal certainly affects the incentive to buy a signal. After all, if the investor planned to report U (or H or L) regardless of the actual signal, then there would be no reason to buy a signal.

In addition to the truth-telling conditions, a constraint is needed to guarantee that investors choose to acquire information. The first set of conditions is that buying and reporting a signal offers at least as high an expected profit as not purchasing a signal and falsely reporting either H or L during Stage 1 trading:

$$\begin{aligned} \pi_i\phi(\mu R(H, H) + (1 - \mu)R(L, L)) + (1 - \pi_i\phi)R(U, U) \\ \geq R(\emptyset, \sigma) + c, \quad \sigma \in \{H, L, U\}, \quad i \leq \tau, \end{aligned} \tag{17}$$

where $R(\emptyset, \sigma)$ is the expected profit to an investor who reports σ without observing a

signal. The profit of truthful reporting is equal to the profit from the corresponding derivative contract $R(s_i, s_i) = f_i(s_i)$ for $s_i \in \{H, L\}$ and the profit from reporting the neutral signal $R(U, U) = f_i(U) = 0$. The expected profit to an investor who reports σ without observing a signal is

$$R(\emptyset, \sigma) = \begin{cases} \mu(-F) + (1-p)f_{i,L}, & \sigma = L, \\ \mu f_{i,H} + (1-\mu)(-F), & \sigma = H. \end{cases} . \tag{18}$$

As a result, the condition (17) is equivalent to the following two conditions:

$$f_{i,H}(\pi_i \phi \mu) - f_{i,L}(1 - \pi_i \phi)(1-\mu) \geq c - \mu F, \tag{19}$$

$$f_{i,L}(\pi_i \phi (1-\mu)) - f_{i,H}(1 - \pi_i \phi)\mu \geq c - (1-\mu)F. \tag{20}$$

The second condition reflects the incentives of the investors to purchase a signal relative to patiently waiting for State 2 offering and not to apply for State 1 information acquisition:

$$\begin{aligned} \pi_i \phi(\mu R(H,H) + (1-\mu)R(L,L)) + (1 - \pi_i \phi)R(U,U) - c \\ \geq P(s = H)(v_H - p_{issue}(H)) + P(s = L)(v_L - p_{issue}(L)) \\ + P(s = U)R^{Stage2}(U), \quad i \leq \tau, \end{aligned} \tag{21}$$

where $R^{Stage2}(U)$ is the return that the investor receives from observing state U announced by the issuer in Stage 2. This return is either equal to $(\mu v_H + (1-\mu)v_L - p_{issue}(U))$ when the investor is allocated a share and accepts price $p_{issue}(U)$ or 0 if the investor is not allocated the share or refuses to participate.

5.3 Equilibrium

The optimal strategy of the investors (in terms of whether to acquire the information and whether to tell the truth) depends on their beliefs about issuer's commitment to start the Stage 2 even if it is unsuccessful in revealing the true fundamental value or not. Given that the information acquisition process can be lengthy and requires several rounds to obtain the information, we assume that there is no time discount-

ing for the issuer. We discuss the implications of time discounting in the following sections.

The following analysis characterizes a Nash equilibrium in which each investor applies to Stage 1 trading, optimally pays for acquiring information and truthfully reveals the information to the issuer via purchasing the corresponding derivative contract.

Let us consider the state of the market when the investors failed to acquire information after i rounds ($s_i = U$ for all i). If the i traders who received those uninformative signals in the preceding i rounds genuinely tried to acquired information and did not hide informative signals, then the probability that there is additional information to be gained in the market conditional on observing i uninformative signals in a row is π_i, as defined in Equation (4).

Due to the fact that $\pi < 1$, investors might have incentives not to participate information acquisition and wait for Stage 2 betting that the information is not revealed during K rounds and try and acquire information privately. This happens, for example, when the number of rounds $K_{\max}$ is small (e.g., due to very low value of γ) and there is sufficiently high conditional probability of information acquisition π_i. In order to eliminate this possibility, the issuer has to invoke some additional mechanism that prevents information production. One example of such a mechanism can be **NI** described in the previous section.

The following proposition shows that the combination of **SEQ** and **NI** mechanisms implements the first-best allocation.

Proposition 5. *Consider the following mechanism:*

- *The issuer contacts each investor sequentially and asks them to produce information;*
- *The issuer stops contacting investor as soon as it obtains a positive ($s_i = H$) or a negative ($s_i = L$) report during round i or the number of rounds with unsuccessful reports exceeds $K_{\max}$;*

- *Conditional upon observing* $s_i = U$*, it sets* $f_{i,L} = f_{i,H} = \varepsilon + \frac{c}{\pi_i \phi}$ *with arbitrarily small* $\varepsilon > 0$ *and* $F > \left(\varepsilon + \frac{c}{\phi \pi_{K_{\max}}}\right)$*;*
- *If* $s_i = H$ *or* $s_i = L$ *for some* $i \leq K_{\max}$ *then the issuer sells* Q *shares to investors (chosen randomly) at price* $p_{issue} = p_1(s_i)$*.*
- *If* $s = U$ *for* $i = K_{\max}$ *then the issuer invokes* **NI** *mechanism and sells* Q *shares to investors at price* $p_{issue} = E(v)$*.*

Then each of I *investors applies for Stage 1 trading and the expected utility of the issuer is given by*

$$\begin{aligned} \Pi^*_{\mathbf{SEQ}} &= M + (Q+H)E(v) - cK_{\max}(1-\pi) - \frac{c\pi(1-(1-\phi)^{K_{\max}})}{\phi} \\ &- \gamma\mu(1-\mu)(1-\pi+\pi(1-\phi)^{K_{\max}})(v_H - v_L)^2. \end{aligned}$$

Proof. See Appendix. □

Corollary 6. FP *mechanism is never optimal.*

Proof. See Appendix. □

6 Discussion of results and limitations

The optimality of our mechanism is dependent on several key assumptions. In this section we discuss their relevance and limitations of the mechanism with respect to these assumptions.

One aspect where we are substantially different from Sherman and Titman (2002) as well as other papers in the literature is the sequential nature of our mechanism. In order to minimize the information acquisition costs, the issuer pays only one investor at each point in time rather than a number of investors at once. Optimality of our mechanism does not rely on the necessity of several investors having the same

information to ensure truth-telling (as in Sherman and Titman, 2002). However, it might take several rounds for the investors to obtain information. In the model, the costs associated with a delay of the issue do not enter the objective function of the issuer and we assume that it is patient enough to wait as long as needed in order to produce the information (e.g., $E(\tau^*)$ increases as γ increases). If time were to enter the preferences of the issuer, the equilibrium solution would have to exhibit a trade-off between time preferences (speed of information acquisition) and its precision.

It should also be noted our model is that there is no secondary market for the derivatives. This means that the investor has to hold the derivative until maturity in order to cash out the reward for information production. Introducing a secondary market for derivatives is not straightforward as this might alter incentives of the investors for truth-telling in anticipation of potential derivative resale price.

Another important feature of our mechanism is the absence of an active market for shares before the derivative contracts trade (stage 1). This makes the IPO an ideal application of our mechanism. In the presence of an active parallel market (for example, SEO) the mechanism would still lead to the production of information and its full revelation in equilibrium. However, the availability of a market where the investor could trade after having acquired information would improve his outside option, maker his truth-telling constraint more binding and so increase the cost of information production for the issuer.

7 Conclusions

In this paper, we use a mechanism design approach to show that price informativeness can be achieved without illiquidity, at a cost equal to the information production cost. We build a model of stock issuance where the issuer incentivizes investors to search for costly information and truthfully disclose it. This is achieved by organizing the issue process in two stages where in the first stage, investors are sequentially offered the possibility to buy two derivatives securities, one that pays only if the asset payoff is high and one that pays only if it is low. We show that the entrepreneur can

design the derivatives in such a way that an investor who participates to stage 1 finds it optimal to produce information and select the derivative security that truthfully reveals the asset payoff if she learned this payoff. Moreover, if an investor does not discover information, she optimally abstains from buying or selling a derivative. As a result, in the second stage, the issuer sells the asset at a price equal to its expected payoff.

The proposed two-stage mechanism allows the issuer to pay the information costs directly to the investor while efficiently relaxing incentive constraints (such as misreporting the information that investors obtain or not pay the cost of information acquisition). This, in turn, allows the issuer to avoid adverse selection costs.

References

Beneviste, L., and W. Wilhelm (1990): "A comparative analysis of IPO proceeds under alternative regulatory environments," *Journal of Financial Economics*, 28(1-2), 173–207.

Biais, B., P. Bossaerts, and J.-C. Rochet (2002): "An Optimal IPO Mechanism," *The Review of Economic Studies*, 69, 117–146.

Bond, P., A. Edmans, and I. Goldstein (2012): "The Real Effects of Financial Markets," *Annual Review of Financial Economics*, 4, 339–360.

Edmans, A., I. Goldstein, and W. Jiang (2015): "Feedback Effects, Asymmetric Trading, and the Limits to Arbitrage," *American Economic Review*, 105(12), 3766–3797.

Faure-Grimaud, A., and D. Gromb (2004): "Public Trading and Private Incentives," *The Review of Financial Studies*, 17(4), 985–1014.

Foucault, T., and T. Gehrig (2008): "Stock price informativeness, cross-listings, and investment decisions," *Journal of Financial Economics*, 88(1), 146–168.

Goldstein, I. (2022): "Information in Financial Markets and Its Real Effects," *Review of Finance*, 27(1), 1–32.

Grossman, S., and J. Stiglitz (1980): "On the Impossibility of Informationally Efficient Markets," *American Economic Review*, 70(3), 393–408.

Holmström, B., and J. Tirole (1993): "Market Liquidity and Performance Monitoring," *Journal of Political Economy*, 101(4), 678–709.

Rock, K. (1986): "Why new issues are underpriced," *Journal of Financial Economics*, 15(1-2), 187–212.

Sherman, A. (2005): "Global trends in IPO methods: Book building versus auctions with endogenous entry," *Journal of Financial Economics*, 78, 615–649.

Sherman, A., and S. Titman (2002): "Building the IPO order book: underpricing and participation limits with costly information," *Journal of Financial Economics*, 65(1), 3–29.

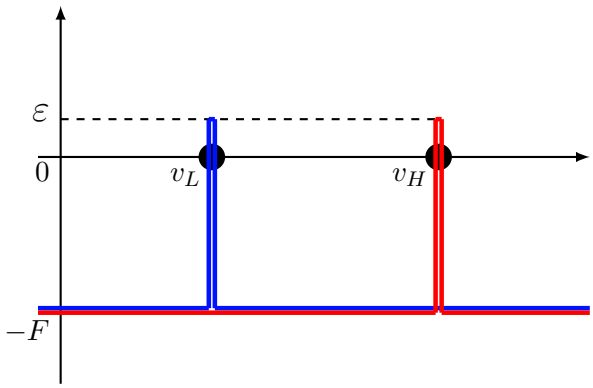
Subrahmanyam, A., and S. Titman (1999): "The Going-Public Decision and the Development of Financial Markets," *The Journal of Finance*, 54(3), 1045–1082.

Appendix

Proof of Proposition 1. The mechanism must be incentive compatible both for informed and uninformed investors. Suppose that an informed investor applies to participate in Stage 1, observes the realization of the fundamental value v and chooses the derivative contract C_w, $w \in \{L, H\}$. The profit of the investor is:

$$\text{Profit}^I_{\text{Stage 1}}(w|v) = \begin{cases} \epsilon, & v = v_w, \\ -F, & v \neq v_w. \end{cases} \tag{22}$$

Figure below plots the profits for each of the contracts (the black line for C_L and the red line for C_H) as function of v.



The informed investor, upon participation in Stage 1, has incentive to disclose the information to the market maker via choosing the set of contracts corresponding to the true fundamental value. The profit of the informed in this case is $\text{Profit}^I_{\text{Stage 1}} = \epsilon > 0$.

Given a strictly positive profit in Stage 1, informed investors have incentives to participate in Stage 1 rather than Stage 2:

$$E\left[\text{Profit}^I_{\text{Stage 2}}\right] = v - p_{issue} = 0 < E\left[\text{Profit}^I_{\text{Stage 1}}\right].$$

The profit of an uninformed investor who decides to participate in stage 1 and chooses

the contract C_w is

$$E\left[\text{Profit}^U_{\text{Stage 1}}|w = v_L\right] = (1-\mu)\epsilon - \mu F < 0 = E\left[\text{Profit}^U_{\text{Stage 2}}\right].$$

Hence, only informed investors choose to participate in the first stage, which leads to full information revelation. Finally, given that the value of ϵ is arbitrarily chosen, the loss of the issuer in this can be arbitrarily small. In a limit case when $\epsilon \to 0$, the loss of an issuer approaches to zero. □

Proof of Proposition 2. We first show that it is optimal to stop whenever $s_i = H$ or $s_i = L$. In this case, $s = s_i$ and since the information is revealed truthfully by assumption of the benchmark model, we have that $\text{E}(Var(v \mid p_2)) = 0$. Furthermore,

$$\begin{aligned}\Pi_i(p_{issue}, \{f_j\}) &= M + (N+Q)p_2(s_i) - \sum_{j=0}^{i} f_j(s_j) \\ &\geq M + (N+Q)p_2(s_i) - \sum_{j=0}^{k} f_j(s_j) = \Pi_k(p_{issue}, \{f_j\}).\end{aligned}$$

for any $k > i$.

Next, we prove that $\tau^* \leq K_{\max}$. To do so, we first need to calculate the continuation value of the objective function at any round i for the next k number of rounds given that it is optimal to stop after the informative signal. First, note that

$$\text{E}(Var(v \mid p_2)) = (1 - \pi_{i+1} + \pi_{i+1}(1-\phi)^k)\mu(1-\mu)(v_H - v_L)^2. \tag{23}$$

Indeed, if the process does not continue after k rounds, this means that the information revealed and $Var(v \mid p_2) = 0$. If this is not the case, then $s_{i+j} = U$ for any $j \leq k$ and the information is not revealed. Hence $Var(v \mid p_2) = \mu(1-\mu)(v_H - v_L)^2$. The latter case happens if there is either no information in the market (with probability $1 - \pi_{i+1}$) or the investors were unlucky to find one (with probability $\pi_{i+1}(1-\phi)^k$).

Next we calculate the expected fee needed to be paid for the next k rounds.

Lets denote by I_h the event that the issuer gets an informative signal exactly after contacting the hth investor (i.e., either $s_{i+h} = H$ or $s_{i+h} = L$ and all other previously contacted investors produces uncertain signal U). We also denote by U_h the event that the issuer gets uninformative signals from each investors $i+1, ..., i+h$ (i.e., $s_{i+j} = U$ for all $j \leq h$). In order to calculate the future expected fee (ignoring already paid sunk costs to the previous i investors) note that for any $h = 2, ..., k-i$:

$$\begin{aligned}
E\left(\sum_{j=1}^{k} f_{i+j,s_{i+j}} | I_h\right) &= \sum_{j=1}^{h-1} f_{i+j,U} + \mu f_{i+h,H} + (1-\mu) f_{i+h,L} \equiv \sum_{j=1}^{h-1} f_{i+j,U} + \bar{f}_{i+h}, \\
Pr(I_h) &= \pi_{i+1}(1-\phi)^{h-1}\phi, \\
E\left(\sum_{j=1}^{k} f_{i+j,s_{i+j}} | U_k\right) &= \sum_{h=1}^{k} f_{i+h,U}, \\
Pr(U_k) &= 1 - \pi_{i+1} + \pi_{i+1}(1-\phi)^{i+k}.
\end{aligned}$$

By the low of total expectations,

$$\begin{aligned}
& E\left(\sum_{j=1}^{k} f_{i+j,s_{i+j}}\right) = \sum_{h=1}^{k} E\left(\sum_{j=1}^{k} f_{i+j,s_{i+j}} | I_h\right) Pr(I_h) + E\left(\sum_{j=1}^{k} f_{i+j,s_{i+j}} | U_k\right) Pr(U_k) \\
= & \sum_{h=1}^{k} \left(\bar{f}_{i+h} + \sum_{j=1}^{h-1} f_{i+j,U}\right) \pi_{i+1}(1-\phi)^{h-1}\phi + \left(\sum_{h=1}^{k} f_{i+h,U}\right) \left(1 - \pi_{i+1} + \pi_{i+1}(1-\phi)^{i+k}\right) \\
= & \sum_{h=1}^{k} \bar{f}_{i+h}\pi_{i+1}(1-\phi)^{h-1}\phi + \sum_{h=1}^{k} \pi_{i+1}(1-\phi)^{h-1}\phi \sum_{j=1}^{h-1} f_{i+j,U} \\
+ & \sum_{h=1}^{k} f_{i+h,U} \left(1 - \pi_{i+1} + \pi_{i+1}(1-\phi)^{i+k}\right) \\
\overset{Ineq.(10)}{\geq} & \sum_{h=1}^{k} \frac{c\pi_{i+1}(1-\phi)^{h-1}}{\pi_{i+h}} - \sum_{h=1}^{k} f_{i+h,U} \left(\frac{1-\phi\pi_{i+h}}{\pi_{i+h}}\right) \pi_{i+1}(1-\phi)^{h-1} \\
+ & \sum_{h=1}^{k} \pi_{i+1}(1-\phi)^{h-1}\phi \sum_{j=1}^{h-1} f_{i+j,U} + \sum_{h=1}^{k} f_{i+h,U} \left(1 - \pi_{i+1} + \pi_{i+1}(1-\phi)^{i+k}\right)
\end{aligned}$$

with the equality whenever the constraint (10) is binding.

Given that

$$\frac{\pi_{i+1}(1-\phi)^{h-1}}{\pi_{i+h}} = 1-\pi_{i+1}+\pi_{i+1}(1-\phi)^{h-1}, \quad (24)$$

$$\left(\frac{1-\phi\pi_{i+h}}{\pi_{i+h}}\right)\pi_{i+1}(1-\phi)^{h-1} = 1-\pi_{i+1}+\pi_{i+1}(1-\phi)^{h}, \quad (25)$$

$$\sum_{h=1}^{k}(1-\phi)^{h-1}\sum_{j=1}^{h-1}f_{i+j,U} = \sum_{h=0}^{k-1}f_{i+h,U}\frac{[(1-\phi)^h-(1-\phi)^k]}{\phi} \quad (26)$$

we have the following inequality:

$$\begin{aligned}
E\left(\sum_{j=1}^{k}f_{i+j,s_{i+j}}\right) &\geq cK(1-\pi_{i+1})+\frac{c\pi_{i+1}(1-(1-\phi)^k)}{\phi}-\sum_{h=1}^{k}f_{i+h,U}\left(1-\pi_{i+1}+\pi_{i+1}(1-\phi)^h\right)\\
&+ \pi_{i+1}\sum_{h=0}^{k-1}f_{i+h,U}[(1-\phi)^h-(1-\phi)^k]+\sum_{h=1}^{k}f_{i+h,U}\left(1-\pi_{i+1}+\pi_{i+1}(1-\phi)^k\right)\\
&= ck(1-\pi_{i+1})+\frac{c\pi_{i+1}(1-(1-\phi)^k)}{\phi}+\pi_{i+1}\sum_{h=0}^{k-1}f_{i+h,U}[(1-\phi)^h-(1-\phi)^k]\\
&- \pi_{i+1}\sum_{h=1}^{k}f_{i+h,U}[(1-\phi)^h-(1-\phi)^k] = ck(1-\pi_{i+1})+\frac{c\pi_{i+1}(1-(1-\phi)^k)}{\phi}.
\end{aligned}$$

Hence, the continuation value for up to k rounds is

$$\begin{aligned}
E\left(\Pi_k(p_{issue},\{f_i\})|\tau=i\right) &\leq M+(N+Q)\mathrm{E}(v)-ck(1-\pi_{i+1})-\frac{c\pi_{i+1}(1-(1-\phi)^k)}{\phi}\\
&- \gamma(1-\pi_{i+1}+\pi_{i+1}(1-\phi)^k)\mu(1-\mu)(v_H-v_L)^2. \quad (27)
\end{aligned}$$

We are ready to prove that it is sub-optimal to continue with the information search process if $\tau \geq K_{\max}$. We prove this by showing that the continuation value for any number of rounds $k>1$ is smaller than the expected value of the objective function $E\left(\Pi_0(p_{issue},\{f_i\})|\tau=i\right)$ when the process is stopped at τ.

Indeed, suppose that $\tau \geq K_{\max}$. Then

$$E\left(\Pi_0(p_{issue},\{f_i\})|\tau=i\right) = M+(N+Q)\mathrm{E}(v)-\gamma\mu(1-\mu)(v_H-v_L)^2.$$

Hence,

$$
\begin{aligned}
& E\left(\Pi_k(p_{issue},\{f_i\})|\tau=i\right) - E\left(\Pi_0(p_{issue},\{f_i\})|\tau=i\right) \\
\leq\ & \gamma\mu(1-\mu)(v_H-v_L)^2 \\
-\ & ck(1-\pi_{i+1}) - \frac{c\pi_{i+1}(1-(1-\phi)^k)}{\phi} - \gamma(1-\pi_{i+1}+\pi_{i+1}(1-\phi))\mu(1-\mu)(v_H-v_L)^2 \\
=\ & -ck(1-\pi_{i+1}) - \frac{c\pi_{i+1}(1-(1-\phi)^k)}{\phi} + \gamma\pi_{i+1}\left[1-(1-\phi)^k\right]\mu(1-\mu)(v_H-v_L)^2 \\
=\ & -ck(1-\pi_{i+1}) + \pi_{i+1}\left[1-(1-\phi)^k\right]\left[-\frac{c}{\phi}+\gamma\mu(1-\mu)(v_H-v_L)^2\right] \\
\leq\ & -ck(1-\pi_{i+1}) + \pi_{i+1}\left[1-(1-\phi)^k\right]\left[-\frac{c}{\phi}+\frac{c}{\phi\pi_{i+1}}\right] \\
=\ & -ck(1-\pi_{i+1}) + \frac{c(1-\pi_{i+1})\left[1-(1-\phi)^k\right]}{\phi} = c(1-\pi_{i+1})\left[-k+\frac{1-(1-\phi)^k}{\phi}\right] < 0.
\end{aligned}
$$

In order to finalize the proof we need to show that it is optimal to continue as long as $\tau^* \leq K_{\max}$. Suppose that the issuer managed to run $\tau = i$ information search rounds (with $0 < i < K_{\max}$) and all of them result in an uninformative signal U and

$$\frac{c}{\pi_{i+1}\phi} < \mu(1-\mu)(v_H-v_L)^2. \tag{28}$$

Then the expected cost of running at least one round of information search is less than or equals to

$$
\begin{aligned}
E\left(\Pi_1(p_{issue},\{f_i\})|\tau=i\right) &= M+(N+Q)\mathrm{E}(v) - \pi_{i+1}\phi\left[\mu f_{i+1,H}+(1-\mu f_{i+1,L})\right] \\
&- \gamma(1-\pi_{i+1}+\pi_{i+1}(1-\phi))\mu(1-\mu)(v_H-v_L)^2.
\end{aligned}
$$

According to the constraint (10), $\pi_{i+1}\phi\left[\mu f_{i,H}+(1-\mu f_{i,L})\right] \geq c$ but the issuer can achieve equality if it sets $\mu f_{i+1,H}+(1-\mu f_{i+1,L}) = \frac{c}{\pi_{i+1}\phi}$. Hence, the total expected cost in the case of one round of information search is

$$E\left(\Pi_1(p_{issue},\{f_i\})|\tau=i\right) = M+(N+Q)\mathrm{E}(v)-c-\gamma(1-\pi_{i+1}+\pi_{i+1}(1-\phi)\mu(1-\mu)(v_H-v_L)^2.$$

The difference in the expected objective functions is

$$\begin{aligned} & E\left(\Pi_1(p_{issue},\{f_i\})|\tau=i\right)-E\left(\Pi_0(p_{issue},\{f_i\})|\tau=i\right) \\ = \;& \gamma\mu(1-\mu)(v_H-v_L)^2-c-\gamma(1-\pi_{i+1}+\pi_{i+1}(1-\phi))\mu(1-\mu)(v_H-v_L)^2 \\ = \;& -c+\gamma\pi_{i+1}\phi\mu(1-\mu)(v_H-v_L)^2>0 \end{aligned}$$

(the last inequality follows from inequality (28).

The choice $p_{issue}=p_1(s)$ is attainable and maximizes the objective function given the constraint (11). Furthermore, the ex-ante expected costs of the issuer is minimized when $\bar{f}_i\equiv\mu f_{i,H}+(1-\mu)f_{i,L}=\frac{c}{\phi\pi_i}$ and $f_{i,U}=0$ and the expected objective function is equal to (13). □

Proof of Proposition 3. Let us suppose that the issuer decided to implement a hybrid procedure where it would call for M_i investors every round i who would search for the information simultaneously. The issuer promises to compensate them with fees $f_i(s_{i,m})$ depending on the signal they report, where $s_{i,m}$ is the signal reported by the investor m in round i. This compensation should satisfy for each m

$$\pi_i\left[\phi\mu f_{i,H}+\phi(1-\mu)f_{i,L}+(1-\phi)f_{i,U}\right]+(1-\pi_i)f_{i,U}\geq c. \tag{29}$$

So, as a result, the issuer's total expected fee in round i is

$$\sum_{m=1}^{M_i}\left\{\pi_i[\phi\mu f_{i,H}+\phi(1-\mu)f_{i,L}+(1-\phi)f_{i,U}]+(1-\pi_i)f_{i,U}\right\}\geq M_i c. \tag{30}$$

Suppose that the issuer selects the fee structure so that the equality holds in (30). The issuer's objective function for running k rounds of information search is

$$\begin{aligned} \Pi_k^{hybrid}\left(p_{issue},\{f_i\}\right) &= M+(N+Q)\mathrm{E}(v)-E(C_{issue})-\gamma E(Var(v\mid p_2)) \\ &= M+(N+Q)\mathrm{E}(v)-cM_1-Pr(i>1)E(C_{future}\mid i>1) \\ &- \gamma(1-\pi+\pi(1-\phi)^{M_1})E(Var(v\mid p_2)\mid i>1), \end{aligned}$$

where $E(C_{future} \mid i > 1)$ is the expected future costs that the issue expected to incur conditional one more than one round going forward.

Consider now an alternative procedure, where instead of calling M_1 investors during the round 1 simultaneously, the issuer calls M_1 one by one to search for the information. If all of them fail to produce an informative signal, then the remaining procedure is identical to the initial hybrid one. Then the issuer's objective function for running those $M_1 + k - 1$ rounds (insuring that the same number of potential investors participates) of information search is this case is

$$\tilde{\Pi}^{hybrid}_{M_1+k-1}\left(p_{issue}, \{f_i\}\right) = M + (N+Q)\mathrm{E}(v) - cM_1(1-\pi) - \frac{c\pi}{\phi}\left(1-(1-\phi)^{M_1}\right)$$
$$- \quad Pr(i > M_1)E(C_{future} \mid i > M_1) - \gamma(1-\pi+\pi(1-\phi)^{M_1})E(Var(v \mid p_2) \mid i > M_1).$$

Since, $M_1 > \frac{1-(1-\phi)^{M_1}}{\phi}$ we have that $\tilde{\Pi}^{hybrid}_{k}\left(p_{issue}, \{f_i\}\right) > \Pi^{hybrid}_{k}\left(p_{issue}, \{f_i\}\right)$.

This means that no matter what M_1 the issuer chooses for the hybrid procedure, it is always better off in running M_1 sequential rounds first with the ex-ante predetermined number of traders M_1 rather than calling them simultaneously. Given that the issue has flexibility of adjusting this M_1 ex-post (if the informative signal realizes sooner than M_1 rounds), this increases the expected objective function even further.

Finally, repeating this step for each round i with $M_i > 1$ shows that pure sequential procedure dominates the hybrid (or simultaneous) one. □

Proof of Proposition 4. For the issue to succeed, the issuer must guarantee the participation of uninformed investors. Suppose that $v_L < p_{issue} < v_H$ and consider a situation in which it is optimal for each uninformed investor to buy one share at this price. At this price, each informed investor finds it optimal to buy one share if $v = v_H$ and to abstain otherwise. Thus, when $v = v_H$, each uninformed investor only receives $q_u(v_H) = \frac{Q}{H}$ shares (pro-rata rationing), while when $v = v_L$ each uninformed investor receives $q_u(v_L) = \frac{Q}{H(1-\lambda)}$. Thus, the expected profit of uninformed investors

is:

$$E(q_u(v)(v - p_{issue})) = \mu q_u(v_H)(v_H - p_{issue}) + (1 - \mu)q_u(v_L)(v_L - p_{issue}).$$

To guarantee the participation of uninformed investors (which is necessary for the issue to succeed) and maximize the proceeds of the issue, the issuer must choose the largest price such that $E(q_u(v)(v - p_{issue})) \geq 0$, which is the price solving $E(q_u(v)(v - p_{issue})) = 0$. Thus, the issuing price is:

$$p^*_{issue} = \beta v_H + (1 - \beta)v_L,$$

with $\beta = \frac{\mu(1-\lambda)}{1-\mu\lambda}$. As $\lambda > 0$, we have: $p_{issue} < E(v)$. Thus, the issue must be underpriced for it to succeed. Note that in this case, the issuing price does not reveal information about v since it is identical whether informed investors participate or not in the issue. However, total demand in the issue fully reveals the asset payoff. Thus, if total demand is revealed ex-post, one obtains accuracy but at the cost of underpricing. This is a manifestation of the trade-off between illiquidity (here measured by underpricing) and informativeness.

To simplify, suppose that investors with the ability to produce the signal but who do not participate in the IPO (they will be indifferent in equilibrium). Likewise, suppose that informed investors who search for information but don't find it don't participate to the IPO. Now suppose that the price of the issue is such that $v_L < p_{issue} < v_H$. Thus, it is optimal for informed investors with information to demand one share when $v = v_H$ and to demand no shares when $v = v_L$. Moreover suppose that p_{issue} is such that it is optimal to buy one share for uninformed investors. Let $q_u(v)$ be the allocation to uninformed investors when the payoff of the asset is v and let $q_i(v)$ be the allocation to informed investors when the payoff of the asset is v. If $0 \leq k \leq K$ investors find information, we have:

1. $q_u(v_H) = q_i(v_H) = \frac{Q}{k+H}$
2. $q_u(v_L) = \frac{Q}{H}$, $q_i(v_L) = 0$

Note that k the number of informed participants in the IPO is random and that $Prob(k = j) = \pi\binom{K}{k}(1-\phi)^{K-k}(\phi)^k$ for $0 < k < K$ and $Prob(k = 0) = \pi(1-\phi)^K + (1-\pi)$.

The expected profit of an uninformed investor when the price of the issue is p_{issue}:

$$\Pi_u(p_{issue}) = E(q_u(v)(v - p_{issue})). \tag{31}$$

The largest price of the issue that guarantees the participation of uninformed (which is necessary for the success of the issue when $v = v_L$) is therefore such that $\Pi_u(p_{issue}) = 0$, that is:

$$p_{issue} = \frac{E(q_u(v)v)}{E(q_u)}. \tag{32}$$

One can compute the price of the issue differently. Observe that the clearing condition in the IPO implies:

$$Hq_u(v) + kq_i(v) = Q \quad \text{for} \quad \forall k \text{ and } \forall v. \tag{33}$$

Thus,

$$H\Pi_u(p_{issue}) = E(Hq_u(v)(v - p_{issue})) = Q(E(v) - p_{issue}) - E(kq_i(v)(v - p_{issue})) = 0,$$

implying

$$Q(E(v) - p_{issue}) = E(kq_i(v)(v - p_{issue}). \tag{34}$$

This means that the total amount left on the table by the issuer is equal to informed investors' total expected profit. Moreover:

$$p_{issue} = \frac{E(v)Q}{Q - E(kq_i(v))} - \frac{E(kq_i(v)v)}{Q - E(kq_i(v))}. \tag{35}$$

Now, let $\tau(k) \equiv \frac{k}{k+H}$. $\tau(k)$ is the fraction of the issue allocated to informed investors when $v = v_H$. When $v = v_L$, informed investors do not trade. Thus, we have:

$$E(kq_i(v)v) = E(\tau(k))Q\mu v_H, \text{ and}$$

$$E(kq_i(v)) = E(\tau(k))Q\mu.$$

We deduce that

$$p_{issue} = \beta v_H + (1-\beta)v_L \tag{36}$$

with $\beta = \frac{\mu(1-E(\tau(k))}{1-E(\tau(k))\mu}$. Observe that $\beta < \mu$ if $E(\tau(k)) > 0$. Thus, informed trading in the IPO generates underpricing.

Given our assumptions, one can compute $E(\tau(k))$:

$$E(\tau(k)) = \pi \sum_{k=1}^{K} \binom{K}{k} (1-\phi)^{K-k} \phi^k \left(\frac{k}{k+H}\right). \tag{37}$$

Let $p^*_{issue}(K)$ be the equilibrium issue price when K investors search for information. In equilibrium, the aggregate expected profits of these investors is (from eq.(34)):

$$E(kq_i(v)(v - p^*_{issue}(K))) = Q(E(v) - p^*_{issue}(K)). \tag{38}$$

Thus, the aggregate expected profit of informed investors searching for information is equal to the expected loss of the issuer in the IPO (relative to an issue at the unconditional expected value of the asset).

Now consider the determination of K. Each informed investor who searches for information expects a profit of $\Pi_i(K) = \frac{E(kq_i(v)(v-p^*_{issue}(K)))}{K} = \frac{Q(E(v)-p^*_{issue}(K))}{K}$. As K increases, $\Pi_i(K)$ decreases (to be checked). And thus, K_{FP} is the largest value of K such that:

$$\Pi_i(K) \geq c. \tag{39}$$

Let K_{FP} be this value. We have:

$$K_{FP}\Pi_i(K_{FP}) \approx K^{FP}c. \tag{40}$$

In this approach, the aggregate demand in the IPO provides a more complex signal about the payoff of the asset. Let $D(v) = Q(q_i(v) + q_u(v)$ be this demand. It is either equal to H if $v = v_L$ or $v = v_H$ and $k = 0$ or strictly larger than H if $v = v_H$ and $k > 0$. Thus, when $D > H$, the IPO outcome reveals that $v = v_H$. If $D = H$, however, the

IPO demand is not fully revealing. Let $\mu(D = H, K) = Pr(v = v_H \mid D = H)$ when K investors search for information. We have:

$$\mu(D = H, K) = \frac{\mu(1 - \pi + \pi(1 - \phi)^K)}{\mu(1 - \pi + \pi(1 - \phi)^K) + (1 - \mu)} \tag{41}$$

Observe that $\mu(D = H, K) < \mu$. Observing that $D = H$ is bad news as it indicates the possibility that $v = v_L$. Note also that $\mu(D > H, K) = Pr(v = v_H \mid D > H) = 1$. It follows that:

$$\begin{aligned} E(Var(v \mid D)) &= \mu(D = H, K)(1 - \mu(D = H, K))(v_H - v_L)^2 \\ &= \frac{\mu(1 - \mu)(1 - \pi + \pi(1 - \phi)^K)}{\mu(1 - \pi + \pi(1 - \phi)^K) + (1 - \mu)}(v_H - v_L)^2. \end{aligned} \tag{42}$$

Thus, in equilibrium, the expected objective function of the issuer is:

$$\begin{aligned} &\Pi_{FP} = M + NE(v) + QE(p^*_{issue}) - E(C_{issue}) - \gamma E(Var(v \mid p_2(s))) \\ &\approx M + (Q + N)E(v) - cK_{FP} - \frac{\gamma\mu(1 - \mu)\left(1 - \pi + \pi(1 - \phi)^{K_{FP}}\right)(v_H - v_L)^2}{\mu\left(1 - \pi + \pi(1 - \phi)^{K_{FP}}\right) + (1 - \mu)}. \end{aligned} \tag{43}$$

□

Proof of Proposition 5. Let us start with verifying the truth-telling condition (16). Suppose that an investor applies to participate in Stage 1 and is chosen to acquire information. The investor observes the informative signal $s_i \in \{H, L\}$ and hence learns the realization of the true fundamental value v with probability ϕ. Conditional on observing the informative signal the investor purchases the corresponding derivative contract C_H if $s_i = H$ or C_L if $s_i = L$. The investor's profit is $R(H, H)$ or $R(L, L)$ respectively, and given that $f_{i,H} = f_{i,H} \equiv f_i$ is equal to:

$$R(H, H) = R(L, L) = F + \varepsilon + f_i - F = \varepsilon + \frac{c}{\phi\pi_i}.$$

Conditional on observing the neutral signal the trader is better off not participating

in the trade as long as

$$R(U,H) = \mu\left(\varepsilon + \frac{c}{\phi\pi_i}\right) - (1-\mu)F < 0, \tag{44}$$

$$R(U,L) = (1-\mu)\left(\varepsilon + \frac{c}{\phi\pi_i}\right) - \mu F < 0. \tag{45}$$

Both inequalities (44) and (45) hold is we choose large enough F, that is, if

$$F > \max\left\{\frac{\mu}{1-\mu}, \frac{1-\mu}{\mu}\right\}\left(\varepsilon + \frac{c}{\phi\pi_{K_{\max}}}\right) \geq \max\left\{\frac{\mu}{1-\mu}, \frac{1-\mu}{\mu}\right\}\left(\varepsilon + \frac{c}{\phi\pi_i}\right).$$

Moreover, $R(H,L) = R(L,H) = -F < 0$. This verifies truth-telling constraint (16). Finally, since an investor reporting U signal is excluded from Stage 2 allocation, $R(H,U) = R(L,U) = -c$, and hence the investor has no incentive to sabotage and not to disclose an informative signal.

Next, we verify the set of conditions (19) and (20) that buying and reporting a signal offers at least as high expected profit as not purchasing a signal and falsely reporting either H or L during Stage 1 trading. Since

$$\left(\varepsilon + \frac{c}{\phi\pi_i}\right) \leq \left(\varepsilon + \frac{c}{\phi\pi_{K_{\max}}}\right) < \frac{F}{\max\left\{\frac{\mu}{1-\mu}, \frac{1-\mu}{\mu}\right\}},$$

the following relationship holds:

$$\begin{aligned}
&f_{i,H}(\phi\pi_i\mu) - f_{i,H}(1-\phi\pi_i)(1-\mu) = \left(\varepsilon + \frac{c}{\phi\pi_i}\right)(\pi_i\phi\mu) - \left(\varepsilon + \frac{c}{\phi\pi_i}\right)(1-\pi_i\phi)(1-\mu) \\
&= \varepsilon\pi_i\phi + c - (1-\mu)\varepsilon - \frac{(1-\mu)c}{\phi\pi_i} > c - \left(\varepsilon + \frac{c}{\phi\pi_i}\right)(1-\mu) \\
&> c - \frac{(1-\mu)F}{\max\left\{\frac{\mu}{1-\mu}, \frac{1-\mu}{\mu}\right\}} \geq c - \mu F,
\end{aligned}$$

which proves condition (19). Similarly,

$$f_{i,L}(\pi_i\phi(1-\mu)) - f_{i,H}(1-\pi_i\phi)\mu = \left(\varepsilon + \frac{c}{\phi\pi_i}\right)(\pi_i\phi(1-\mu)) - \left(\varepsilon + \frac{c}{\phi\pi_i}\right)(1-\pi_i\phi)\mu$$
$$= \varepsilon\pi_i\phi(1-\mu) + c - \mu\varepsilon - \frac{\mu c}{\phi\pi_i} > c - \left(\varepsilon + \frac{c}{\phi\pi_i}\right)\mu$$
$$> c - \frac{\mu F}{\max\left\{\frac{\mu}{1-\mu}, \frac{1-\mu}{\mu}\right\}} \geq c - (1-\mu)F,$$

which proves that the condition (20) holds.

To verify condition (21) that the investors have incentives to purchase a signal relative to patiently waiting for State 2 offering, we note that $R^{Stage2}(U) = 0$. Indeed, after the issuer invokes the **NIP** mechanism, the investor has no incentives to acquire information privately is better off accepting price $p_{issue}(U) = \mu v_H + (1-\mu)v_L$, in which case $R^{Stage2}(U) = 0$. Hence, condition (21) follows from this argument.

As a result, all additional constraints are satisfied, and since the optimal stopping rule, and choice of the functions $\{f_i\}$ and p_issue are identical to the benchmark model, the allocation achieved in this mechanism coincides with the benchmark allocation. Hence, the expected value of the objective function is equal to

$$\begin{aligned}\Pi^*_{\mathbf{SEQ}} &= M + (Q+N)E(v) - cK_{\max}(1-\pi) + \frac{c\pi(1-(1-\phi)^K_{\max})}{\phi} \\ &- \gamma\mu(1-\mu)\left(1-\pi+\pi(1-\phi)^{K_{\max}}\right)(v_H - v_L)^2 \end{aligned} \tag{46}$$

□

Proof of Corollary 6. Consider the following two cases: a) $K_{FP} \leq K_{\max}$ and b) $K_{FP} > K_{\max}$.

a). Let us modify the sequential mechanism so that the we bound the stopping time from above by $\tau \leq K_{FP}$. Let us denote the expected objective function of the

issuer in this case by

$$\Pi_{\mathbf{SEQ}}(\tau \leq F_{FP}) = \max_{\{p_{issue},\{f_i\}\}} \Pi(p_{issue}, \{f_i\}, \tau \leq F_{FP}),$$

where $\Pi(p_{issue}, \{f_i\}, \tau \leq F_{FP})$ is the value of the optimal stopping problem

$$\Pi(p_{issue}, \{f_i\}, \tau \leq F_{FP}) = \sup_{0\leq\tau\leq F_{FP}} \Pi_\tau(p_{issue}, \{f_i\}),$$

Following the proof of Propositions 2 and 5 we can deduce that

$$\begin{aligned}\Pi^*_{\mathbf{SEQ}}(\tau \leq F_{FP}) &= M + (Q+N)E(v) - cK_{FP}(1-\pi) + \frac{c\pi(1-(1-\phi)^{K_{FP}})}{\phi} \\ &- \gamma\mu(1-\mu)\left(1-\pi+\pi(1-\phi)^{K_{FP}}\right)(v_H - v_L)^2 < \Pi^*_{\mathbf{SEQ}}.\end{aligned}$$

Hence, we have

$$\begin{aligned}&\Pi_{FP} - \Pi_{SEQ} < \Pi_{FP} - \Pi_{SEQ}(\tau \leq F_{FP}) \\ =\ & cK_{FP}(1-\pi) + \frac{c\pi(1-(1-\phi)^{K_{FP}})}{\phi} + \gamma\mu(1-\mu)\left(1-\pi+\pi(1-\phi)^{K_{FP}}\right)(v_H - v_L)^2 \\ -\ & cK^{FP} - \frac{\gamma\mu(1-\mu)\left(1-\pi+\pi(1-\phi)^{K^{FP}}\right)(v_H - v_L)^2}{\mu\left(1-\pi+\pi(1-\phi)^{K^{FP}}\right)+(1-\mu)} \\ <\ & -cK_{FP}\pi + \frac{c\pi(1-(1-\phi)^{K_{FP}})}{\phi} \\ +\ & \gamma\mu(1-\mu)\left(1-\pi+\pi(1-\phi)^{K_{FP}}\right)(v_H - v_L)^2 - \gamma\mu(1-\mu)\left(1-\pi+\pi(1-\phi)^{K_{FP}}\right)(v_H - v_L)^2 \\ =\ & \frac{c\pi}{\phi}\left(1-(1-\phi)^{K_{FP}} - K_{FP}\phi\right) < 0. \end{aligned} \tag{47}$$

b). Consider a round i such that $i > K_{\max}$, so we have

$$c > \pi_{i+1}\phi\gamma\mu(1-\mu)(v_H - v_L)^2. \tag{48}$$

In this case it is optimal for **SEQ** mechanism to stop. Let us show that in this case **NIP** mechanism also dominates **FP** mechanism. Indeed, note that **NIP** mechanism

is preferred over **FP** mechanism by the issuer if and only if:

$$U \equiv cK_{FP} - \frac{\gamma\pi_{i+1}\mu(1-\mu)^2(v_H-v_L)^2(1-(1-\phi)^{K_{FP}})}{\mu(1-\pi_{i+1}+\pi_{i+1}(1-\phi)^{K_{FP}})+(1-\mu)} > 0. \tag{49}$$

Since (48) holds, we can write:

$$\begin{aligned}
U &> K_{FP}\pi_{i+1}\phi\gamma\mu(1-\mu)(v_H-v_L)^2 - \frac{\gamma\pi_{i+1}\mu(1-\mu)^2(v_H-v_L)^2(1-(1-\phi)^{K_{FP}})}{\mu(1-\pi_{i+1}+\pi_{i+1}(1-\phi)^{K_{FP}})+(1-\mu)} \\
&= \pi_{i+1}\gamma\mu(1-\mu)(v_H-v_L)^2\left(K_{FP}\phi - \frac{(1-\mu)(1-(1-\phi)^{K_{FP}})}{\mu(1-\pi_{i+1}+\pi_{i+1}(1-\phi)^{K_{FP}})+(1-\mu)}\right) \\
&> \pi_{i+1}\gamma\mu(1-\mu)(v_H-v_L)^2\left((1-(1-\phi)^{K_{FP}}) - \frac{(1-\mu)(1-(1-\phi)^{K_{FP}})}{\mu(1-\pi_{i+1}+\pi_{i+1}(1-\phi)^{K_{FP}})+(1-\mu)}\right) \\
&= \pi_{i+1}\gamma\mu(1-\mu)(v_H-v_L)^2(1-(1-\phi)^{K_{FP}})\left(1 - \frac{1-\mu}{\mu(1-\pi_{i+1}+\pi_{i+1}(1-\phi)^{K_{FP}})+(1-\mu)}\right) \\
&= \pi_{i+1}\gamma\mu(1-\mu)(v_H-v_L)^2(1-(1-\phi)^{K_{FP}})\left(1 - \frac{1-\mu}{\mu(1-\pi_{i+1}+\pi_{i+1}(1-\phi)^{K_{FP}})+(1-\mu)}\right) \\
&> \pi_{i+1}\gamma\mu(1-\mu)(v_H-v_L)^2(1-(1-\phi)^{K_{FP}})\left(1 - \frac{1-\mu}{\mu(1-\pi_{i+1})+(1-\mu)}\right) > 0
\end{aligned}$$

This shows that whenever it is optimal to stop within **SEQ** mechanism, it is not optimal to invoke **FP** as an alternative.

Suppose now that $i > K_{\max}$. We can show that **SEQ** mechanism dominates **FP** in the similar way as in a). To do so, we can modify **SEQ** mechanism by forcing the issuer to continue the information search until round $K_{FP} > K_{\max}$ if $s_i \neq U$ for $i = K_{\max+1}, ..., K_{FP}$. The difference in the expected objective functions are then

$$\Pi_{FP} - \Pi_{SEQ} < \Pi_{FP} - \Pi_{SEQ}(K_{\max} \leq \tau \leq F_{FP}) < 0.$$

following the same logic as in (47). □