FTG Working Paper Series

Siphoned Apart: A Portfolio Perspective on Order Flow Segmentation

by

Markus Baldauf
Joshua Mollner
Bart Yueshen Zhou

Working Paper No. 00094-00

Finance Theory Group

www.financetheory.com

*FTG working papers are circulated for the purpose of stimulating discussions and generating comments. They have not been peer reviewed by the Finance Theory Group, its members, or its board. Any comments about these papers should be sent directly to the author(s).

Siphoned apart:
A portfolio perspective on order flow segmentation *

Markus Baldauf[†] Joshua Mollner[‡] Bart Zhou Yueshen[§]

March 28, 2023

Click here for latest version

* This paper benefits tremendously from helpful comments by Robert Battalio (discussant), Eric Budish, Isaías Chaves, Darrell Duffie, Naveen Gondhi, Zhiguo He, Shiyang Huang, Ralph Koijen, John Kuong, Katya Malinova (discussant), Robert McDonald, Maureen O'Hara, Christine Parlour (discussant), Ioanid Roşu (discussant), Gideon Saar, Alvaro Sandroni, John Shim, Konstantin Sokolov (discussant), Vincent van Kervel, and the seminar and conference participants at INSEAD, KU Leuven, Maastricht University, Monash University, Nanyang Technological University, NFA, SAFE Market Microstructure Conference, Singapore Management University, Texas Finance Festival, and Tilburg University. We also appreciate greatly insightful conversations with industry participants. Baldauf would like to thank the University of Chicago, Booth School of Business, where he was visiting when large parts of this paper were written. There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.

[†] University of British Columbia, Sauder School of Business. 2053 Main Mall, Vancouver, BC, Canada V6T 1Z2. Email: baldauf@mail.ubc.ca.

[‡] Northwestern University, Kellogg School of Management, 2211 Campus Drive, Evanston, IL 60657. Email: joshua.mollner@kellogg.northwestern.edu

[§] INSEAD, 1 Ayer Rajah Avenue, Singapore 276337. E-mail: b@yueshen.me.

Siphoned apart: A portfolio perspective on order flow segmentation

Abstract

We study liquidity provision in fragmented markets. Market makers intermediate heterogeneous order flows, trading off expected spread revenue against inventory costs. Applying our model to payment for order flow (PFOF), we demonstrate that portfolio-based considerations of inventory management incentivize market makers to segment retail orders by siphoning them off-exchange. Banning order flow segmentation unambiguously hurts welfare, can make trading more costly for all investors, and can resolve a prisoner's dilemma affecting market makers. These results differentiate our inventory-based model from the existing information-based theories of PFOF.

Keywords: order flow segmentation, payment for order flow, inventory management, market maker, retail investor

1 Introduction

Trade in modern financial markets is spread across many venues. Creation of the National Market System in the U.S. in 1975 spawned a sustained regulatory effort to create a reliable, integrated exchange trading environment. Yet, an enormous amount of trading still happens off-exchange—in recent months, as much as half of all equity volume. In part, this reflects differences in intermediation costs: investors who tend to be less costly for market makers to intermediate can be given better prices if they are segmented into separate venues. But what does this order flow segmentation imply for welfare and liquidity? This paper argues that the implications may be nuanced and may depend on *why* certain investors are less costly to intermediate than others.

There are two reasons for why investors may differ in their intermediation costs, each relating to one of the two classic frictions in the market microstructure literature: asymmetric information (Glosten and Milgrom, 1985; Kyle, 1985) and inventory costs (Stoll, 1978; Amihud and Mendelson, 1980; Ho and Stoll, 1981, 1983). On the one hand, certain orders may be less costly to intermediate because they are less informed about fundamentals. And several existing studies have analyzed order flow segmentation through this lens. What has so far received less attention in the literature—despite its empirical importance—is the other possibility: that certain orders may be less costly to intermediate because they tend to be less correlated in direction with the other orders in a market maker's portfolio.[1] This paper aims to fill that gap. The analysis introduces a portfolio perspective on order flow management, which highlights the incentive to endogenously segment orders. We obtain additional predictions regarding liquidity and the welfare of various liquidity-demanding investors. Finally, our analysis has implications for the potential consequences of regulatory intervention.

We first study a baseline model in Section 2, where order flow segmentation is taken as given:

[1] Indeed, the staff report of SEC (2021) states that "[retail] orders are more likely to be small, uncorrelated with one another, and thus 'one and done' (i.e., not the first in a series of orders intended to transact a large amount of stock), which also allows for a tighter spread."

order flows from various sources (e.g., retail investors or institutions) are exogenously split across marketplaces. These various marketplaces can represent exchanges, dark pools, 'upstairs' block trading, or any other form of fragmentation. Yet another form of order flow segmentation is payment for order flow (PFOF), a common order-handling practice in which retail orders are routed directly to market makers, who typically execute them against their own balance sheets. In Sections 3–5, we specialize the model to the setting of PFOF, where we show how considerations of inventory cost can cause this practice to arise endogenously.

Baseline model. A single security is traded on a fixed number of marketplaces, in each of which liquidity-demanding orders arrive at a flow rate that is price-elastic (i.e., elastic with respect to the bid-ask spread). One key characteristic of a marketplace is its order flow directionality, which captures the probability with which each order arriving there is a buy or a sell. These directionalities are modeled as random variables with an arbitrary joint distribution, and they capture order correlation within and across marketplaces. Such correlation can arise, for example, from the splitting of large institutional parent orders into child orders or from retail investors' sentiment-driven trades. A continuum of market makers provide liquidity by choosing liquidity supply intensities—for each marketplace, a Poisson rate at which they are willing to accept randomly assigned orders—balancing expected revenues from bid-ask spreads against quadratic inventory costs. An equilibrium consists of liquidity-supply intensities and spreads for each marketplace such that (i) each market maker's liquidity-supply intensities are optimal given the spreads, and (ii) each spread clears liquidity demand and supply in its marketplace.

A key insight from the baseline model is that market makers must consider their liquidity supply decisions across marketplaces as a *portfolio*: the correlation structure of directionalities determines the extent to which order flows will offset, hence the expected inventory cost of each marginal order. In fact, the market maker's problem is tightly connected to the optimization problem in standard portfolio theory. Our analysis highlights that portfolio considerations matter even for inventory management of a *single* asset.

The importance of inventory considerations in general—and of a portfolio perspective in particular—is consistent with recent empirical evidence. Daures-Lescourret and Moinas (2022) show that after a shock to her inventory from an execution on one venue, a market maker's liquidity provision on the same (opposite) side becomes less (more) aggressive on *all* venues. Barardehi et al. (2022) argue that market makers treat retail orders differently, depending on institutional liquidity demand imbalances. Portfolio-based inventory considerations also explain two other facts: (i) Eaton et al. (2022) find that outage-induced reductions in Robinhood retail activity improve on-exchange liquidity provision (while outages of other brokers harm liquidity), and (ii) Schwarz et al. (2022) experimentally document that Robinhood clients receive less price improvement than clients of other retail brokers (E*Trade, Fidelity, TD Ameritrade). An explanation for both findings is that Robinhood order flow tends to exacerbate inventory imbalances: compared with order flows from other retail brokerages, Robinhood orders are more concentrated (Barber et al., 2022) and more correlated with past returns (Eaton et al., 2022).

Application to PFOF. Whereas order flow segmentation is exogenous in the baseline model, we next investigate how it might *endogenously* arise, for example, in the form of PFOF.[2] Section 3 considers a setting with two marketplaces, on-exchange and off-exchange, and two order sources, R and I, for *r*etail and *i*nstitutional investors. I-orders must clear on-exchange, but an endogenous fraction of R-orders may be siphoned off-exchange. This siphoning endogenously affects both the characteristics of the on-exchange order flow and the correlation between the on-exchange and off-exchange order directionalities—and in turn, the equilibrium volumes and bid-ask spreads as well.

Depending on parameters, the equilibrium entails either (i) both R- and I-orders clearing on-exchange (i.e., a "no-siphoning" equilibrium) or (ii) all R-orders siphoned off-exchange, leaving

[2] In practice, "PFOF" may refer to either (i) the practice whereby retail orders are routed directly to market makers and executed off-exchange; or (ii) the payment transferred from market makers to retail brokers for such a purpose. In this paper, "PFOF" refers only to (i). We do not model the payment (ii) both for parsimony and because it tends to be very small (Schwarz et al., 2022).

only *I*-orders on-exchange (i.e., a "with-siphoning" equilibrium). To see how siphoning arises, conjecture an equilibrium in which all orders clear on-exchange. This implies a specific portfolio of *R*- and *I*-orders for market makers. A market maker might, however, benefit if she could alter the composition of *R*- and *I*-orders in her portfolio—due to the orders' possibly different characteristics, like their arrival rates and their directionalities. In particular, if it is beneficial to obtain more *R*-orders, they can be siphoned off-exchange with the promise of a slightly smaller spread. Doing so destroys the conjectured no-siphoning equilibrium, yielding a with-siphoning equilibrium instead.

We characterize the exact condition that separates the two regions of parameters. In words, the with-siphoning region is precisely where *R*-orders contribute less to inventory costs (at the margin) than *I*-orders do. For example, this happens when the variance of *R*-orders' directionality is sufficiently small. In this case, market makers find *R*-orders more attractive and siphon them off-exchange as a result. We argue in Section 3.3 that the realistic parameter values lie in the with-siphoning region, which is consistent with the fact that "in the equity markets right now, if you place a [retail] market order,. . . , 90–95 percent do not go to the lit exchange, do not go to Nasdaq or New York Stock Exchange, they go to wholesalers" (Gensler, 2022). In Section 5, we consider an extension of the model in which market makers can choose their liquidity supplies dynamically. *R*-orders continue to be siphoned off-exchange in this dynamic extension—in fact, siphoning is further incentivized by new dynamic considerations.

Most existing models of PFOF analyze it through the lens of asymmetric information, as in Easley, Kiefer, and O'Hara (1996) and Battalio and Holden (2001). These models view retail orders as less informed than institutional orders, meaning that they create less adverse selection and can therefore be cleared at a smaller spread if siphoned off-exchange. Our model proposes an entirely different mechanism: retail orders may contribute less to—and may even reduce—market makers' inventory risk. As we discuss below, this differing mechanism implies different predictions and different policy implications.

One set of predictions concerns the consequences of banning off-exchange retail trading (here-after, a "PFOF ban") on spreads, market maker profits, and total welfare. Our analysis adds to the ongoing policy debate regarding PFOF. While the practice is widely popular in the U.S., it would be affected by new rules that the SEC has recently proposed (SEC, 2022). In Europe, the issue remains contentious (Reuters, 2023).

When a PFOF ban has an effect (i.e., in the with-siphoning region of the parameter space), the model predicts that it harms *R*-investors, in the sense that it leads them to pay a larger spread. One might think such a ban would entail a countervailing benefit for *I*-investors—however, this is not necessarily so. Rather, for certain parameters, *I*-investors are also harmed. When PFOF is banned and *R*-investors are charged a larger spread, fewer *R*-investors opt to trade. If *R*-orders are sufficiently effective for hedging *I*-orders, this leads market makers to anticipate ending with a larger net inventory imbalance, and they charge a larger on-exchange spread to compensate. Moreover, there is reason to think that this is not only a theoretical possibility, but in fact the empirically-relevant case: Evidence from Jones et al. (2022) suggests that retail order imbalances negatively correlate withinstitutional imbalances, hence are effective for hedging against them.

These predictions reveal an interesting comparison with the existing information-based models of PFOF. These models posit that *R*-investors pay a smaller spread when siphoned off-exchange because they are less informed than *I*-investors (i.e., create less adverse selection). By pooling both investor types, a PFOF ban would lead to an intermediate spread, harming *R*-investors while unambiguously benefitting *I*-investors. In contrast, our theory makes a more nuanced prediction for *I*-investors. And although our theory makes the same prediction for *R*-investors, it is for an entirely different reason: not because their orders are less informed but rather because they contribute less inventory risk.

The model's most interesting prediction about market makers' profitability is that PFOF can sometimes function as a prisoner's dilemma: although each market maker has a unilateral incentive to siphon *R*-orders off-exchange, their collective siphoning creates a pecuniary externality, which may lead them to be worse off in equilibrium than if the practice were banned. In this way, our theory rationalizes market makers' seeming ambivalence toward regulatory discussions of PFOF bans.[3]

A PFOF ban unambiguously reduces total welfare in our model. This is because, absent a PFOF ban, the equilibrium in fact leads to the welfare-maximizing quantities of R- and I-investor volume, essentially due to the First Welfare Theorem. By constraining market makers' liquidity supply decisions, a PFOF ban distorts outcomes and necessarily reduces total welfare in our model. This analysis identifies a novel channel through which regulations on PFOF can reduce welfare, via market makers' inventory considerations.

Related literature. Our paper contributes to two strands of literature. First, a theoretical literature has studied market fragmentation from various angles: investors' venue choices (Pagano, 1989, Chowdhry and Nanda, 1991, Babus and Parlatore, 2022); competition among venue operators (Pagnotta and Philippon, 2018, Chao, Yao, and Ye, 2019, Baldauf and Mollner, 2020, Cespa and Vives, 2022); information and price discovery (Ye, 2011, Zhu, 2014); speed and latency arbitrage (Foucault and Menkveld, 2008; van Kervel, 2015); and price impact (Chen and Duffie, 2021). We complement the existing literature by, instead, focusing on market makers' inventory cost concerns and the order flow segmentation that endogenously results. The closest work to ours is Daures-Lescourret and Moinas (2022), which, like our paper, speaks to liquidity provision across exogenously fragmented exchanges in a setting with inventory frictions. Different from their paper, our model highlights that such inventory concerns, in fact, endogenously incentivize market makers to siphon certain orders off-exchange.

Second, our application to PFOF contributes to the theoretical literature on the practice. Battalio and Holden (2001) argue that PFOF and internalization can arise when orders' verifiable characteristics are correlated with informativeness. This is consistent with the evidence from Easley, Kiefer, and O'Hara (1996), who estimate that orders on the main exchange (NYSE) are more likely to be

[3] For example, Ken Griffin (CEO of Citadel) has said that "Payment for order flow is a cost to me . . . So if you're going to tell me that by regulatory fiat one of my major items of expense disappears, I'm OK with that" (FT, 2021).

informed than those diverted to the regional exchange (Cincinnati). More recently, Yang and Zhu (2020) show that by acquiring information about retail flows, "back-runners" like high-frequency trading firms can learn about institutional flows and profit from such inferred information, and may therefore be willing to pay for such retail flows. The above explanations share a common feature: they all analyze PFOF through various information channels, like adverse selection and learning. Our model departs from this focus on information and turns instead to inventory risk as the key friction.

Various other dimensions of PFOF have also been explored theoretically. Chordia and Subrahmanyam (1995) find the migration of order flows from exchange (NYSE) to off-exchange (non-NYSE) through PFOF arises when discrete ticks constrain market makers' price competition. Hagerty and McDonald (1996) analyze a broker's optimal portfolio of informed and uninformed clientele. Kandel and Marx (1999) explicitly study brokers' order-handling decisions: sending to the exchange (via Nasdaq's Small Order Execution System), selling to market makers via PFOF, or internalizing (vertical integration). Parlour and Rajan (2003) argue that PFOF can serve as an anticompetitive device, which raises market makers' profits. Glode and Opp (2016) argue that pre-trading order flow agreements, like PFOF, can sustain intermediation chains that reduce information asymmetry and improve efficiency.

On the empirical side, our model connects to the recent, growing literature studying new developments in retail trading. Jain et al. (2021) document how the rise of zero-commission retail brokers changed various volume shares, for example, across brokers or between exchanges and wholesale market makers. Adams and Kasten (2021) study the execution quality of small orders in the zero-commission regime using Rule 605 filings. Ernst and Spatt (2022) find that retail brokers receive larger PFOF payments in options than equities, meaning that they have an incentive to sway retail investors to the options market.

Our model also relates to the economics literature on price discrimination in competitive environments, surveyed by Stole (2007). One interesting feature of our model is that it is possible for price discrimination to lower spreads for all investors. This cannot happen in the classic model of third degree oligopolistic price discrimination of Holmes (1989), which features two firms and two consumer groups (the "weak" and "strong" markets). However, Corts (1998) demonstrates that if the competing firms differ in which markets they consider strong versus weak, then it is possible for price discrimination to lower prices in both markets, a phenomenon that he calls "all-out competition."

2 A model of liquidity supply

2.1 Setup

Overview. A single security is traded on multiple marketplaces. Its fair value is common knowledge. Liquidity-taking order flows arrive on these venues continuously and are met by liquidity-supplying market makers. We assume away information asymmetry so as to mute channels already analyzed by previous literature. That is, all liquidity-taking orders are submitted for non-fundamental, private-value reasons.

Marketplaces and order flows. Marketplaces are indexed $j \in \{1, \dots, J\}$. For each j, let $s_j \geq 0$ be its half bid-ask spread, which will be determined endogenously in equilibrium. Trading lasts for one unit of time. At each instant, a flow of $\lambda_j(s_j)$ liquidity-demanding orders arrives to marketplace j, with $\lambda_j(\cdot)$ decreasing and nonnegative.

Remark 1. For example, a model of PFOF might entail $J = 2$ marketplaces, with one as the exchange and the other as off-exchange execution on market makers' own balance sheets. In general, a marketplace can represent any of the possible trading venues—including also dark pools, crossing networks, systematic internalizers, single-dealer platforms, exchanges' retail liquidity programs, and the market for block trading.

Order directions. Each order arriving on marketplace j is independently either a one-unit buy or a one-unit sell, with respective probabilities $\frac{1}{2}(1 + D_j)$ and $\frac{1}{2}(1 - D_j)$, where $D_j \in [-1, 1]$ captures the average direction of the marketplace's order flow. The vector of directionalities $\boldsymbol{D} = (D_1, D_2, \ldots, D_J)^\top$ is random, with mean and variance denoted $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{D}]$ and $\Sigma = \mathrm{var}[\boldsymbol{D}]$, respectively, and realizes just before trading starts. As a convention, bold letters denote vectors (always columns) or matrices. We assume Σ is positive-definite.

Remark 2. Directionalities capture the possibility of correlation among orders. In reality, orders could be correlated due to several mechanisms. One is trading on private information about fundamentals. However, given our focus on market makers' inventory concerns, our model better fits correlation driven by forces other than private information. For example, correlation can arise from the splitting of institutional parent orders placed for non-informational reasons like portfolio rebalancing, portfolio transition, and fund flows. Several datapoints help quantify the importance of non-informational institutional trades. Index additions and deletions are a source of such trades, because they cause index funds and other investors to adjust their holdings for mechanical reasons, and they lead to a significant increase in trading volume (e.g., Harris and Gurel, 1986; Greenwood, 2005). In the data of Kyle and Obizhaeva (2016), an average portfolio transition accounts for 4.20% of corresponding stocks' daily trading volume, and this number increases to 16.23% for small stocks. Coval and Stafford (2007) find that flows out of (or into) mutual funds can result in severe uninformed fire selling (or purchasing), and such fund flows amount to as much as 13.9% (Israeli data, Ben-Rephael, Kandel, and Wohl, 2011) or 19.19% (U.S. data, Ben-Rephael, Kandel, and Wohl, 2012) of total trading volume in the market. Furthermore, non-informational correlation can also arise when retail investors coordinate on online platforms (e.g., "WallStreetBets" on Reddit) or herd on the same market sentiment.

Microfoundation. The downward-sloping liquidity demand $\lambda_j(\cdot)$ is as in Garman (1976) and Ho and Stoll (1981). Together with the directionality D_j, it can be microfounded as follows. Assume that order flow on marketplace j originates from a continuum of investors with measure κ_j, who arrive according to independent Poisson processes with intensity η_j. Upon arrival, an investor submits a one-unit immediate-or-cancel (IOC) order with a limit price reflecting her private value for the security—this order will be marketable only if the private value is extreme enough, exceeding the half spread s_j. Her private value is positive with probability $\frac{1}{2}(1 + D_j)$ (and negative otherwise) and its magnitude is drawn i.i.d. from c.d.f. $F_j(\cdot)$. We then obtain the decreasing aggregate liquidity demand $\lambda_j(s_j) = \kappa_j \eta_j \cdot \big(1 - F_j(s_j)\big)$.

Liquidity supply. A continuum of market makers, indexed by $m \in [0, M]$, compete to provide liquidity on all J marketplaces.[4] Before $\boldsymbol{D}$ realizes, every market maker m chooses her liquidity supply intensity x_{mj} for each marketplace j, taking as given the half spreads $(s_j)_{j=1}^J$, to maximize her expected payoff described below. The supply x_{mj} is the Poisson intensity at which she will accept (randomly assigned) orders on marketplace j. Therefore, at each instant, market makers in aggregate supply liquidity to a flow of $\int_0^M x_{mj} \mathrm{d}m$ orders on marketplace j.

Market makers' expected payoffs. A market maker's payoff has two components: spread revenue and inventory cost.[5] Let Q_{mj} and Z_{mj} denote, respectively, market maker m's realized volume and realized net inventory from supplying liquidity to marketplace j. For example, if she receives 2 buy and 3 sell orders on that venue, she gets a volume of $Q_{mj} = 2+3$ and a net inventory of $Z_{mj} = -2+3$. Each unit of volume earns her the half-spread s_j. We assume the inventory cost is quadratic, with $\gamma > 0$ as a common scaling parameter. Ex ante both Q_{mj} and Z_{mj} are random, with distributions depending on her liquidity supply x_{mj}. Therefore, her expected payoff

$$\mathbb{E}\left[\sum_{j=1}^{J} Q_{mj} s_j - \frac{\gamma}{2}\left(\sum_{j=1}^{J} Z_{mj}\right)^2\right] \tag{1}$$

is an endogenous function of the liquidity supply choices $(x_{mj})_{j=1}^J$.

Remark 3. Our model entails a very stylized description of liquidity provision: each market

[4] While in reality the market making sector is composed of a handful of large wholesale market makers (Citadel, Virtu, etc.), we model them as a continuum to simplify the analysis by ensuring price-taking behavior.

[5] Market makers' inventory costs can arise from risk aversion, price impact in portfolio rebalancing, capital pledged to clearing houses, or a moral hazard problem between market makers and their financiers (Bruche and Kuong, 2021).

maker m commits to a vector $(x_{mj})_{j=1}^{J}$, which determines the arrival process of orders that she will receive for the duration of the trading game. The interpretation of x_{mj} may depend on what marketplace j represents. For example, if j refers to an exchange, then x_{mj} can represent a market maker's posted limit orders, which constitute a short-term commitment of liquidity. Alternatively, the commitment x_{mj} can reflect the "groundwork" that a market maker needs to lay before trading starts, like the allocation of computational power, bandwidth, and staffing, data subscription fees, regulatory and compliance costs, etc. If j refers to market makers' siphoning of retail orders, then x_{mj} can refer to market maker m's negotiations with brokers over the terms at which retail orders will be processed.

Of course, in many settings, more elaborate—for example, dynamic—strategies would be feasible. Yet, a tradeoff between spread revenues and inventory costs should always remain, analogous to the tradeoff reflected in equation (1). Thus, the liquidity supply intensities can also serve as a crude, low-dimensional approximation of a market maker's full strategy space—an approximation that buys substantial tractability. Moreover, Section 5 considers dynamics more formally, showing that our main conclusions carry over to a two-period version of the model.

By letting market makers choose intensities x_{mj}, we effectively base our model of liquidity provision on quantity competition rather than price competition. Empirical evidence, e.g., from Brogaard and Garriott (2019), supports such a view of algorithmic market makers.

Equilibrium definition. An equilibrium consists of liquidity supply intensities $(x_{mj})_{j=1}^{J}$ for each market maker m and a half spread s_j for each marketplace j, such that (i) each market maker m's $(x_{mj})_{j=1}^{J}$ maximizes her expected payoff (1), and (ii) market clearing holds for each marketplace j:[6]

$$\int_0^M x_{mj} \mathrm{d}m = \lambda_j(s_j), \ \forall j \in \{1, \dots, J\}. \tag{2}$$

[6] In formulating this market-clearing condition, we follow convention in assuming an exact law of large numbers over a continuum of independent random variables. See Duffie, Qiao, and Sun (2020) for a rigorous formulation.

2.2 Equilibrium characterization

We first express a market maker's expected payoff (1) as a function of her supply intensities $(x_{mj})_{j=1}^{J}$. Suppose a market maker m has chosen x_{mj}. Her volume Q_{mj} is then Poisson distributed with mean $x_{mj} \cdot 1$, where 1 is the length of the trading period. She therefore expects spread revenue of $\mathbb{E}\left[Q_{mj} s_j\right] = x_{mj} s_j$ from marketplace j. Recalling that each market order is a buy with probability $\frac{1}{2}(1 + D_j)$, her net inventory from marketplace j can be written

$$Z_{mj} = \sum_{i=1}^{Q_{mj}} (-1)^{B_{imj}}, \tag{3}$$

where $\{B_{imj}\}$ are i.i.d. Bernoulli draws with success rate $\frac{1}{2}(1 + D_j)$. Total net inventory across all marketplaces is $\sum_{j=1}^{J} Z_{mj}$, where, to evaluate her quadratic inventory cost, we need the expectation of its square.

Lemma 1. Write a market maker m's liquidity supplies as a vector $\boldsymbol{x}_m := (x_{m1}, \ldots, x_{mJ})^{\top}$. Then $\mathbb{E}\left[\left(\sum_{j=1}^{J} Z_{mj}\right)^2\right] = \boldsymbol{x}_m^{\top}\mathbf{1} + \boldsymbol{x}_m^{\top}(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\top})\boldsymbol{x}_m$, where $\mathbf{1}$ is a length-J column vector of ones.

To understand Lemma 1, consider the special case with only one marketplace j and only sell orders (i.e., $\sigma_j = 0$ and $\mu_j = -1$). Since all orders are to sell, a market maker's realized inventory Z_{mj} equals her volume Q_{mj}, which is a Poisson random variable with mean x_{mj}. Hence, expected squared inventory is $\mathbb{E}\left[Z_{mj}^2\right] = \mathrm{var}\left[Z_{mj}\right] + \mathbb{E}\left[Z_{mj}\right]^2 = x_{mj} + x_{mj}^2$, consistent with Lemma 1. For general values of σ_j and μ_j, we derive in the proof that $\mathbb{E}\left[Z_{mj}^2\right] = x_{mj} + (\sigma_j^2 + \mu_j^2)x_{mj}^2$. Intuitively, potential for both buying and selling, on the one hand, allows certain offsetting in the net inventory Z_{mj}, thus lowering the x_{mj}^2 term to $x_{mj}^2\mu_j^2$; but on the other hand, randomness in the composition of buys versus sells creates additional inventory variation, thus adding the term $x_{mj}^2\sigma_j^2$. With multiple marketplaces, the expectation of the square of total inventory is as stated in Lemma 1.

Using Lemma 1, therefore, we can write the market maker's optimization problem as[7]

$$\max_{\boldsymbol{x}_m} \boldsymbol{x}_m^\top \Big(\boldsymbol{s} - \frac{\gamma}{2}\mathbf{1}\Big) - \frac{\gamma}{2}\boldsymbol{x}_m^\top \big(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\big)\boldsymbol{x}_m, \tag{4}$$

where $\boldsymbol{s} := (s_1, \ldots, s_J)^\top$ is the vector of half spreads. Next, we derive the market maker's optimal liquidity supply intensities. Because the optimization problem (4) is quadratic, first-order conditions suffice:

$$\boldsymbol{x}_m = \frac{1}{\gamma}\big(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\big)^{-1}\Big(\boldsymbol{s} - \frac{\gamma}{2}\mathbf{1}\Big). \tag{5}$$

The equilibrium half spreads $\boldsymbol{s}$ can then be pinned down via the market-clearing condition (2).

Proposition 1 (Equilibrium liquidity supply). There exists a unique equilibrium, where the half spreads $\boldsymbol{s}$ are the unique solutions of $\frac{M}{\gamma}\big(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\big)^{-1}\big(\boldsymbol{s} - \frac{\gamma}{2}\mathbf{1}\big) = \big(\lambda_1(s_1), \ldots, \lambda_J(s_J)\big)^\top$; and where each market maker m's liquidity supply $\boldsymbol{x}_m$ is given by (5), which moreover satisfies $\boldsymbol{x}_m \geq \mathbf{0}$.

2.3 Connections to portfolio theory

The objective (4) resembles the optimization problem in standard portfolio theory:

$$\max_{\boldsymbol{w}} \boldsymbol{w}^\top\big(\boldsymbol{r} - r_f\mathbf{1}\big) - \frac{a}{2}\boldsymbol{w}^\top\Sigma_r\boldsymbol{w},$$

where, fixing the risk-free rate r_f, an investor with risk-aversion coefficient a chooses a weight vector $\boldsymbol{w}$ over risky assets with expected returns $\boldsymbol{r}$ and variances Σ_r. Analogously, our market makers choose *liquidity supply portfolios* $\boldsymbol{x}_m$ for a single asset.

The solution to this standard portfolio problem is $\boldsymbol{w}^* = \frac{1}{a}\Sigma_r\big(\boldsymbol{r} - r_f\mathbf{1}\big)$. Naturally, our expression for the optimal supply (5) resembles it. Portfolio theory, therefore, also suggests an intuition for equation (5). In choosing her optimal portfolio, an investor trades off the benefit from the assets' expected returns $\boldsymbol{w}^\top\boldsymbol{r}$ against two sources of cost: (i) the opportunity cost of not investing in the

[7] We do not require the liquidity supplies $(x_{mj})_{j=1}^J$ to be nonnegative, although they always are in the unique equilibrium characterized by Proposition 1. For example, $x_{mj} < 0$ can be interpreted as the market maker demanding liquidity on marketplace j by crossing the spread.

risk-free asset $\boldsymbol{w}^\top r_f \mathbf{1}$ and (ii) the portfolio's return risk $\frac{a}{2}\boldsymbol{w}^\top \Sigma_r \boldsymbol{w}$. In optimizing her portfolio, the investor maximizes the Sharpe ratio $\frac{\boldsymbol{w}^\top(\boldsymbol{r}-r_f\mathbf{1})}{\sqrt{\boldsymbol{w}^\top \Sigma_r \boldsymbol{w}}}$ then scales according to her risk aversion a. In our setup, a market maker trades off the benefit from the spread revenues $\boldsymbol{x}_m^\top \boldsymbol{s}$ against the expected inventory cost $\boldsymbol{x}_m^\top \frac{\gamma}{2}\mathbf{1} + \frac{\gamma}{2}\boldsymbol{x}_m^\top(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{x}_m$ (Lemma 1). To obtain her optimal liquidity supply (5), a market maker maximizes a similar ratio $\frac{\boldsymbol{x}_m^\top(\boldsymbol{s}-\frac{\gamma}{2}\mathbf{1})}{\sqrt{\boldsymbol{x}_m^\top(\Sigma+\boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{x}_m}}$ then scales according to the inventory cost parameter γ.

As another connection, our equilibrium half spreads $\boldsymbol{s}$ are determined via market clearing (as in (2)), similar to how equilibrium expected returns $\boldsymbol{r}$ are determined in, e.g., CAPM. One key difference is that, in our model, market makers face spread-elastic liquidity demand ($\lambda_j(\cdot)$ is downward-sloping), while in CAPM, the assets are in inelastic fixed supplies. This is because whereas every agent in CAPM faces a portfolio problem, in our model, only the liquidity-supplying market makers do (the liquidity-demanding investors do not).[8]

3 Endogenous order flow segmentation

We next adapt our general model of liquidity provision to the setting of PFOF, highlighting how the economic forces that we model can cause order flow segmentation to arise endogenously.

3.1 Setup

To capture PFOF, we model two types of investors: institutional and retail. Institutional orders must be executed on-exchange; retail orders can be executed either on-exchange or off-exchange (albeit at a spread no worse than that on-exchange). This corresponds to a version of the model described in Section 2.1 with $J = 2$ marketplaces whose order flows are an *endogenous* mixture of retail and institutional orders.

[8] Another distinction applies to the version of the model that we subsequently consider in Sections 3–4. There, market makers are allowed to endogenously siphon order flows, which endogenizes the demand $\lambda_j(\cdot)$ in each marketplace, as well as the corresponding directionality characteristics $\boldsymbol{\mu}$ and Σ.

Investors and their liquidity demand. The two types of liquidity-demanding investors are labeled $k \in \{R, I\}$. Type-k liquidity demand arrives at the flow rate $\lambda_k(s)$, where for tractability we assume, for $k \in \{R, I\}$,

$$\lambda_k(s) = \max\{0, (\zeta - s)\omega_k\},$$

where $\omega_k > 0$ measures the magnitude of the type-k demand, and $\zeta > 0$ reflects the maximum acceptable trading cost—demand falls to zero if the half-spread s exceeds ζ. We also assume $\zeta > \frac{\gamma}{2}$ to guarantee trading in equilibrium.[9]

As before, we assume every order from a type-k investor is independently either a one-unit buy or a one-unit sell, with respective probabilities $\frac{1}{2}(1 + D_k)$ and $\frac{1}{2}(1 - D_k)$, where $D_k \in [-1, 1]$ captures the average direction of the type-k orders. For simplicity,[10] we let $\mathbb{E}[D_I] = \mathbb{E}[D_R] = 0$ and write $\Sigma_\circ = \text{var}\big[(D_I, D_R)^\top\big] = \begin{pmatrix} \sigma_I^2 & \rho\sigma_I\sigma_R \\ \rho\sigma_I\sigma_R & \sigma_R^2 \end{pmatrix}$. We assume that Σ_0 is positive-definite.

Remark 4. The linear demand can be microfounded, following the discussion on p. 9, by assuming that the investors' private value magnitudes are uniformly distributed on $[0, \zeta]$. Setting the same ζ for both $k \in \{R, I\}$ amounts to assuming that each investor type exhibits the same price elasticity, $\frac{\mathrm{d}\lambda_k/\lambda_k}{\mathrm{d}s/s} = -\frac{s}{\zeta - s}$, which is a natural benchmark case.

Marketplaces and routing. Aside from differences in the parameters (ω_R, σ_R) and (ω_I, σ_I), the other difference between the two investor types is that R-investors' orders can be siphoned off-exchange, while I-investors' orders must be executed on-exchange.[11] Formally, we define $J = 2$

[9] Indeed, (i) liquidity demand vanishes on all marketplaces j where $s_j > \zeta$; (ii) hence, market clearing requires zero liquidity provision ($x_{mj} = 0$) for such marketplaces and nonnegative liquidity provision ($x_{mj} \geq 0$) elsewhere; (iii) if $\zeta \leq \frac{\gamma}{2}$, then we have $s_j \leq \frac{\gamma}{2}$ for these other marketplaces; (iv) hence, according to (4), any such $\boldsymbol{x}_m \neq \mathbf{0}$ leads to a negative payoff, so that market makers optimally choose $\boldsymbol{x}_m = \mathbf{0}$.

[10] Setting $\mathbb{E}[D_I] = \mathbb{E}[D_R] = 0$ is with little loss of generality: as seen from Section 2.2 and Proposition 1, the mean vector $\boldsymbol{\mu}$ enters the analysis only via the matrix $\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top$, so that its effects can be equivalently attributed to Σ.

[11] We choose the labels R and I to evoke the practice of PFOF, whereby many retail orders are siphoned off-exchange while many institutional orders stay on-exchange. In practice, however, some institutional orders are also siphoned off-exchange, for example, via dark pools, crossing networks, and systematic internalizers. Likewise, some retail orders remain on-exchange, for example, those that contain instructions to be routed to a specific platforms (known as "directed orders").

marketplaces, labeled as 1 and 2, where 1 refers to on-exchange and 2 to off-exchange execution of retail orders. All I-orders are routed to marketplace 1. For R-orders, an endogenous fraction $\alpha \in [0,1]$ are routed to marketplace 2, and the remaining $1-\alpha$ to marketplace 1. The flow rate of liquidity demand on the two marketplaces therefore becomes

$$\lambda_1(s_1) = \lambda_I(s_1) + (1-\alpha)\lambda_R(s_1) \text{ and } \lambda_2(s_2) = \alpha\lambda_R(s_2).$$

Order flow on marketplace 1 is a mixture of I-orders with weight ω_I and R-orders with weight $(1-\alpha)\omega_R$. Order flow on marketplace 2 is purely type-R. Letting the weighting matrix be

$$\boldsymbol{F}(\alpha) = \begin{pmatrix} \frac{\omega_I}{\omega_I+(1-\alpha)\omega_R} & \frac{(1-\alpha)\omega_R}{\omega_I+(1-\alpha)\omega_R} \\ 0 & 1 \end{pmatrix}, \tag{6}$$

the order flow directionality vector $(D_1, D_2)^\top = \boldsymbol{F}(\alpha)(D_I, D_R)^\top$ is a function of α and therefore endogenous. We also compute $\Sigma = \mathrm{var}\big[(D_1, D_2)^\top\big] = \boldsymbol{F}(\alpha)\Sigma_\circ \boldsymbol{F}(\alpha)^\top$.

Liquidity supply. Market makers, their supply intensities, and their objective functions are modeled precisely as in Section 2.1.

Equilibrium definition. An equilibrium consists of liquidity supply intensities (x_{m1}, x_{m2}) for each market maker m, the fraction $\alpha \in [0,1]$ of R-orders that are siphoned off-exchange, and half spreads s_1 and s_2, such that (i) each market maker m's (x_{m1}, x_{m2}) maximizes her expected payoff (1), (ii) market clearing holds for each marketplace j, which, following (2) can be written

$$\int_0^M x_{m1}\mathrm{d}m = (\zeta - s_1)\big(\omega_I + (1-\alpha)\omega_R\big) \quad\text{and}\quad \int_0^M x_{m2}\mathrm{d}m = (\zeta - s_2)\alpha\omega_R,$$

and (iii) the spreads satisfy both of the following conditions:

$$s_1 \le s_2 \text{ if } \alpha < 1; \tag{7a}$$

$$s_1 \ge s_2 \text{ if } \alpha > 0. \tag{7b}$$

Conditions (7a) and (7b) represent the fiduciary duty of (unmodeled) retail brokers to their clients. For example, they together imply that if $\alpha \in (0,1)$, then $s_1 = s_2$. The intuition is that $\alpha \in (0,1)$

means R-orders are routed both off-exchange and on-exchange. This mixing behavior conforms with best-execution obligations only if both outlets offer identical execution quality, in the sense that $s_1 = s_2$. Similarly, if $\alpha = 1$ ($\alpha = 0$), so that all R-orders are routed off-exchange (on-exchange), then $s_1 \geq s_2$ ($s_1 \leq s_2$).[12]

Remark 5. In practice, when executing retail orders off-exchange, market makers typically charge a spread lower than the one prevailing on-exchange. The difference is called *price improvement*, which endogenously arises in our model whenever $s_1 - s_2 > 0$. Furthermore, when a retail order is siphoned off-exchange in practice, the retail investor's broker may receive an additional payment, known as *payment for order flow*, from the market maker who handles the order. This payment tends to be extremely small, only 0%–4.8% of the size of price improvement, depending on the retail broker, according to Table VII of Schwarz et al. (2022). Nor is it central to the economic mechanism we analyze. We have therefore opted to simplify the model by omitting brokers and ignoring payments they would receive.

3.2 The incentive to siphon retail orders off-exchange

Before characterizing equilibrium, we pause to examine market makers' incentives to siphon retail orders off-exchange in our model. To do so, we conjecture an equilibrium in which all R-orders are traded on the exchange (i.e., $\alpha = 0$), and we seek conditions under which such a no-siphoning equilibrium holds.

With $J = 2$ marketplaces, a market maker m's expected payoff (1) can in general be written

$$\pi_m = \left(s_1 - \frac{\gamma}{2}\right)x_{m1} + \left(s_2 - \frac{\gamma}{2}\right)x_{m2} - \frac{\gamma}{2}\left({\sigma_1}^2 x_{m1}^2 + 2r\sigma_1\sigma_2 x_{m1}x_{m2} + {\sigma_2}^2 x_{m2}^2\right), \tag{8}$$

where ${\sigma_1}^2$ and ${\sigma_2}^2$ are the diagonal elements of Σ and $r\sigma_1\sigma_2$ is the off-diagonal covariance,

[12] The access rule of Reg NMS (Rule 610) prohibits exchanges from differential treatment of orders based on the identity of the trader. Our model captures this through (7b): if $s_1 < s_2$, then all R-investors would prefer to trade on the exchange; by the access rule, they must be allowed to do so; we therefore obtain $\alpha = 0$. In contrast, off-exchange execution is *not* subject to the access rule, which is why I-investors trade on-exchange, even if $s_2 < s_1$.

with $r \in [-1, 1]$ as the correlation. Note that both σ_1 and r are functions of α via the weighting matrix $\boldsymbol{F}(\alpha)$, whereas $\sigma_2 = \sigma_R$ regardless of α (as marketplace 2 contains only R-orders). Below, we write $\sigma_1(\alpha)$ and $r(\alpha)$ to emphasize this dependence.

Suppose we are in an equilibrium with $\alpha = 0$. Market clearing therefore requires that $x_{m2} = 0$. Because the assumption $\zeta > \frac{\gamma}{2}$ rules out a no-trade equilibrium (*cf.* Footnote 9), we therefore have $x_{m1} > 0$. To sustain the conjectured equilibrium, the payoff from a marginal unit of x_{m2} must be negative:

$$\frac{\partial \pi_m}{\partial x_{m2}}\Big|_{x_{m2}=0} = \left(s_2 - \frac{\gamma}{2}\right) - \gamma r(0)\sigma_1(0)\sigma_R x_{m1} \le 0; \tag{9}$$

and the first-order condition with respect to x_{m1} must hold:

$$\frac{\partial \pi_m}{\partial x_{m1}}\Big|_{x_{m2}=0} = \left(s_1 - \frac{\gamma}{2}\right) - \gamma \sigma_1(0)^2 x_{m1} = 0. \tag{10}$$

Therefore, a no-siphoning equilibrium can obtain only if both (9) and (10) hold. Following (7a), $s_2 \ge s_1$, so that both can hold only if $\gamma r(0)\sigma_1(0)\sigma_R x_{m1} \ge \gamma \sigma_1(0)^2 x_{m1}$. Because γ, $\sigma_1(0)$, σ_R, and x_{m1} are all strictly positive, this requires $r(0) \ge \frac{\sigma_1(0)}{\sigma_R}$.[13]

Evaluating $r(0)$ and $\sigma_1(0)$ in terms of the primitive parameters, we find that $\text{sign}\left[r(0) - \frac{\sigma_1(0)}{\sigma_R}\right]$ matches the sign of the quantity

$$\Delta := \left(\sigma_R^2 \omega_R + \rho \sigma_I \sigma_R \omega_I\right) - \left(\sigma_I^2 \omega_I + \rho \sigma_I \sigma_R \omega_R\right), \tag{11}$$

which summarizes the differences between the two investor types that are relevant for siphoning decisions. To understand this expression, note that the first bracketed term represents the covariance of the market maker's existing portfolio with a marginal R-order: σ_R^2 is the covariance with another R-order, $\rho\sigma_I\sigma_R$ is the covariance with an I-order, and these are respectively weighted by ω_R and ω_I. The second term is the analogue for a marginal I-order. When the difference Δ is negative, a

[13] This condition becomes less likely to hold if R-orders become more attractive: either by virtue of becoming relatively less likely to exhibit significant directionality (i.e., small σ_R relative to $\sigma_1(0)$) or by better diversifying inventory risk from on-exchange orders (i.e., small $r(0)$). When R-orders become so attractive that the above condition fails, market makers have an incentive to siphon them off-exchange, destroying the putative no-siphoning equilibrium.

marginal R-order amplifies inventory risk by less (or mitigates it by more) than a marginal I-order, so that market makers have incentive to siphon them.[14] In fact, Δ is the key determinant of the equilibrium:

> **Proposition 2 (The PFOF equilibrium).** The three equilibrium objects—the fraction α of off-exchange R-orders, the liquidity supply $\boldsymbol{x}_m$, and the half spreads $\boldsymbol{s}$—are determined as follows.
>
> (i) If $\Delta < 0$, then there is a unique equilibrium in which $\alpha = 1$.
>
> (ii) If $\Delta > 0$, then there is a unique equilibrium in which $\alpha = 0$.
>
> (iii) If $\Delta = 0$, then there is an equilibrium for any $\alpha \in [0, 1]$.
>
> In all cases, $\boldsymbol{x}_m$ and $\boldsymbol{s}$ follow Proposition 1, with $\lambda_1(s) = \lambda_I(s) + (1 - \alpha)\lambda_R(s)$, $\lambda_2(s) = \alpha\lambda_R(s)$, $\boldsymbol{\mu} = \mathbf{0}$, and $\Sigma = \boldsymbol{F}(\alpha)\Sigma_\circ\boldsymbol{F}(\alpha)^\top$.

Surprisingly, the equilibrium fraction of off-exchange R-orders is a boundary value $\alpha \in \{0, 1\}$ (except when $\Delta = 0$). An intuition is the following. If $\Delta < 0$, then by previous analysis, each market maker wants to siphon at least some R-orders off-exchange. Once all market makers do this, however, on-exchange flow becomes more heavily composed of I-orders, so that each market maker can achieve her targeted mixture of type-R and type-I flow only by siphoning even more R-orders off-exchange. This reinforcing logic repeats itself until all R-orders have been siphoned off-exchange in equilibrium.[15]

Depending on parameters, the equilibrium might not feature on-exchange trading. This can

[14] Conversely, when $\Delta > 0$, market makers would have an incentive to siphon I-orders—if they could—leaving only R-orders on the exchange.

[15] Indeed when $\Delta < 0$, market makers always feel they have too few R-orders in the putative equilibrium involving any $\alpha < 1$. To see this, note that by market clearing, the equilibrium weight of R-orders in market makers' portfolio of order flows can in general be written

$$\frac{(1-\alpha)\lambda_R(s_1) + \alpha\lambda_R(s_2)}{\lambda_I(s_1) + (1-\alpha)\lambda_R(s_1) + \alpha\lambda_R(s_2)}.$$

For all $\alpha \in [0, 1)$, the above remains constant at $\omega_R/(\omega_I + \omega_R)$, because $s_1 = s_2$ for $\alpha \in (0, 1)$ following (7a) and (7b) and because s_2 does not enter when $\alpha = 0$. Focusing on the case of $\alpha = 0$, the in-text analysis showed that if $\Delta < 0$, this exposure to R-orders is too low: market makers want to siphon R-orders to increase their exposure. We have just shown that any $\alpha \in (0, 1)$ leads to the same exposure, hence the same incentive to siphon further. As this discussion suggests, one model change that might lead to an interior equilibrium value for α would be if R- and I-investors exhibited different elasticities of demand (*cf.* Remark 4).

happen because of off-exchange trading: When $\Delta < 0$, market makers siphon R-orders off-exchange and take on the resulting inventories. If $\rho > 0$, this raises the marginal inventory cost of I-orders, potentially to the point at which it is no longer profitable for market makers to provide liquidity on-exchange. Equilibria without on-exchange trading are, of course, inconsistent with the current reality. The following proposition characterizes the parametric conditions that ensure both on- and off-exchange trading in equilibrium. Our subsequent analysis focuses on the case in which these conditions hold.

Proposition 3 (Equilibrium with positive volume both on-exchange and off-exchange). There is a unique equilibrium with both $\int_0^M x_{m1} \mathrm{d}m > 0$ and $\int_0^M x_{m2} \mathrm{d}m > 0$ simultaneously holding if and only if both $\Delta < 0$ and

$$M > \left(\rho\sigma_I\sigma_R - \sigma_R^2\right)\omega_R\gamma \tag{12}$$

simultaneously hold.[16]

In summary, the two conditions $\Delta < 0$ and (12) jointly characterize where market makers supply liquidity in equilibrium. First, the sign of Δ determines whether market makers supply liquidity off-exchange by siphoning R-orders ($x_{m2} > 0$). Second, if they do provide liquidity off-exchange, do they still provide liquidity on-exchange ($x_{m1} > 0$)? The necessary and sufficient condition is (12). One intuitive interpretation of this condition is that it requires sufficiently many market makers, so that none takes on enough inventory risk via off-exchange liquidity supply to fully deter on-exchange liquidity supply.

3.3 Discussion of parameters

Reality features positive volume both on- and off-exchange. According to Proposition 3, this realistic outcome arises in the model when $\Delta < 0$ and (12) holds. Thus, a test of the model is to

[16] By Proposition 2, insisting on equilibrium uniqueness merely rules out the non-generic case of $\Delta = 0$. In this knife-edge case of $\Delta = 0$, there is an equilibrium for each $\alpha \in [0, 1]$, all of which—with the exception of $\alpha = 0$—entail $\int_0^M x_{m1} \mathrm{d}m > 0$ and $\int_0^M x_{m2} \mathrm{d}m > 0$ simultaneously holding.

see whether realistic parameter values are consistent with those conditions. This subsection argues that this is the case.

The fluctuation of order directionality, σ_R and σ_I. Order flow of either type, $k \in \{R, I\}$, can be viewed as a mixture of independent and coordinated trades, with the parameter σ_k capturing the composition of this mixture. For example, the extreme in which each k-order is independently either to buy or to sell (each with equal probability) is captured by $\sigma_k = 0$. And the extreme in which all k-orders are children of the same parent order (which is either to buy or to sell, each with equal probability) is captured by the parametrization $\sigma_k = 1$. Consistent with $\sigma_R < \sigma_I$, independent trades empirically feature much more heavily in retail order flow. Indeed, the SEC report quoted in Footnote 1 makes exactly this point. Likewise, Ken Griffin (CEO of Citadel) said in his 2021 Congressional testimony:

> the average retail order is much smaller in totality than the average order that goes onto an exchange [. . .] Because it's a small order, the amount of risk that we need to assume in managing that order is relatively small, as compared to an order that we have to manage from our on exchange trading. (Griffin, 2021)

More direct evidence is available from Jones et al. (2022), who use comprehensive account-level data from the Chinese stock market to directly compute signed order imbalance measures for both retail and institutional investors at the daily level.[17] They report in Panel C of their Table I that order imbalances are more tightly clustered around zero for retail than for institutional, consistent with $\sigma_R < \sigma_I$. The standard deviation of imbalances is 0.455 for institutional accounts, and it ranges from 0.171 to 0.352 for retail, depending on account size.

The correlation of order directionality, ρ. The aforementioned evidence from Jones et al. (2022) also speaks to the correlation of signed order flow imbalances. They report in Panel C of

[17] It is difficult to perform a similar exercise using standard, nonproprietary datasets—like TAQ and Refinitiv—because they lack retail or institutional trade identifiers. It remains possible to imperfectly classify these trades, e.g., using the algorithm proposed by Boehmer et al. (2021). Their algorithm is, however, not exact (Schwarz et al., 2022), which could lead to biased estimates.

their Table I that retail and institutional imbalances are negatively correlated, consistent with $\rho < 0$. The precise correlation ranges from -0.380 to -0.188, depending on the size of retail account. Negative correlation is explained by retail and institutional investors trading against each other. For example, retail investors short squeezed institutional investors in GameStop and other “meme stocks” in early 2021. More generally, Glossner et al. (2022) find that Robinhood investors tended to purchase stocks during the pandemic that institutions sold. Similarly, Figure 1A of Barardehi et al. (2022) shows that institutional order imbalance is negatively correlated with retail imbalance (as measured via the method of Boehmer et al., 2021).

The magnitude of liquidity demand, ω_R and ω_I. Although retail trading has been on the rise, the majority of trading in the U.S. equity market remains dominated by institutions. For example, Bloomberg (2022) reports that retail trading accounts for 17.5% of total trading volume in the second quarter of 2022, while non-bank buy-side institutions account for 34.8%. This suggests a ratio of $\omega_R/\omega_I \approx 1/2$.

Summary. Overall, the evidence suggests that $\sigma_R < \sigma_I$, $\rho < 0$, and $\omega_R < \omega_I$. These are sufficient to guarantee $\Delta < 0$ in our model, implying positive off-exchange volume (i.e., siphoning of R-orders) in equilibrium. Further, $\rho < 0$ is by itself sufficient for (12) to hold, implying also positive on-exchange volume. As such, our model yields realistic trading patterns under realistic parametrizations.

4 Predictions

This section continues with the PFOF application studied above and explores the model’s predictions. We study three equilibrium objects: bid-ask spreads in Section 4.1, market makers’ profitability in Section 4.2, and total welfare in Section 4.3.

We also address recent policy debates over a potential PFOF ban. A ban could take different forms: some have advocated banning market makers’ paying for retail orders, while other have advocated altogether banning the off-exchange siphoning of retail orders. Our approach concerns the latter. In this sense, when referring to a "PFOF ban," our results speak to the scenario in which both R- and I-orders must be routed to the exchange (marketplace 1), yielding an equilibrium characterized by the following corollary.

> **Corollary 1 (The equilibrium under PFOF ban).** Under an exogenous $\alpha = 0$, the equilibrium liquidity supply $\boldsymbol{x}_m$ and the half spreads $\boldsymbol{s}$ follow Proposition 1, with $\lambda_1(s) = \lambda_I(s) + \lambda_R(s)$, $\lambda_2(s) = 0$, $\boldsymbol{\mu} = \mathbf{0}$, and $\Sigma = \boldsymbol{F}(0)\Sigma_\circ \boldsymbol{F}(0)^\top$.

For the analysis below, we shall write the equilibrium objects under the ban with a subscript b. For example, under the ban, all trades happen on-exchange and there is only one spread, s_b.

A no-siphoning equilibrium ($\alpha = 0$) can endogenously arise under certain parameter values and, of course, a PFOF ban has no effect in this case. More interesting—and more relevant—are cases where a PFOF ban will bite. For our analysis in this section, we specialize to the set of relevant parameters, for which equilibrium features positive volume both on and off exchange. Using the characterization of Proposition 3, we therefore maintain the following assumption:

Assumption *(Relevant parameter values).* $\Delta < 0$ and condition (12) both hold.

4.1 Bid-ask spreads: trading costs

This subsection studies equilibrium spreads. Without a PFOF ban, we examine s_1 for on-exchange, s_2 for off-exchange, and $\bar{s}$ for volume-weighted average half spread, defined as[18]

$$\bar{s} = \frac{\lambda_I(s_1)s_1 + \lambda_R(s_2)s_2}{\lambda_I(s_1) + \lambda_R(s_2)}, \tag{13}$$

and the price improvement $s_1 - s_2$. With a PFOF ban, there is only one spread s_b.

[18] The volume weights for s_1 and s_2 in (13) are $\lambda_I(s_1)$ and $\lambda_R(s_2)$ respectively because, under the maintained assumption that $\Delta < 0$, without a PFOF ban all R-orders are siphoned off-exchange (i.e., $\alpha = 1$) in equilibrium, meaning that all I-investors pay s_1 and all R-investors s_2.

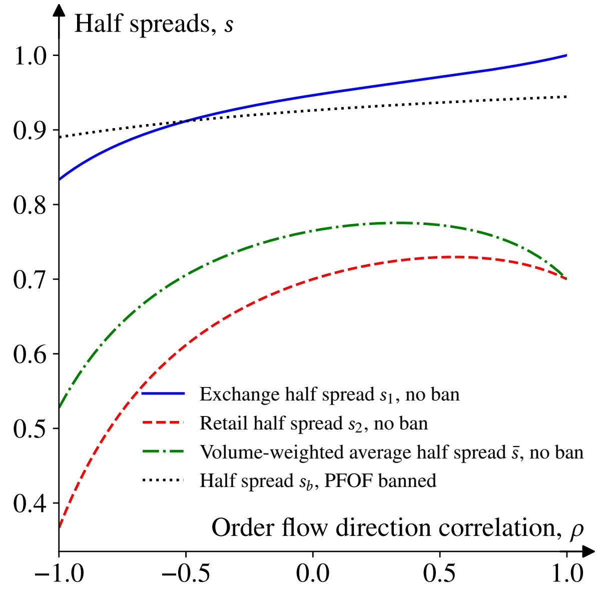


Figure 1: Bid-ask spreads: the effect of a PFOF ban. This figure shows how various (half) bid-ask spreads change when a PFOF ban is imposed. The order flow direction correlation ρ varies on the horizontal axis. The other parameters are set at $M = 3$, $\gamma = \zeta = 1$, $\omega_I = 100$, $\omega_R = 50$, $\sigma_I = 0.5$, $\sigma_R = 0.2$, and $\mu_I = \mu_R = 0$.

4.1.1 PFOF ban

We begin by examining a PFOF ban—more precisely, a ban on off-exchange retail trading. We find that R-investors are unambiguously harmed by a PFOF ban, in the sense that it causes them to pay a larger spread ($s_b > s_2$). In contrast, I-investors often benefit from a PFOF ban—although not always. In particular, a PFOF ban harms not only R-investors but also I-investors if and only if ρ is sufficiently negative (in a way made precise by Proposition 4 and as illustrated by Figure 1). Finally, such a ban unambiguously causes the volume-weighted average spread to increase ($s_b > \bar{s}$).

Proposition 4 (PFOF ban and spreads). $s_b > s_2$ and $s_b > \bar{s}$. Moreover, $s_b > s_1$ if and only if $\rho < -(M + \sigma_R^2 \omega_R \gamma)/(\sigma_I \sigma_R \omega_I)$.

Comparison with information-based theories of PFOF. We pause here to compare and contrast Proposition 4 with predictions from information-based theories of PFOF (e.g., Battalio and Holden, 2001). Under the natural assumption that R-orders are less informed than I-orders, those theories predict that R-investors would be harmed by a PFOF ban, while I-investors would benefit. Our model makes the same prediction regarding R-investors, but for an entirely different reason: R-investors pay a smaller spread when PFOF is allowed not because their orders are less informed but rather because their orders generate less inventory risk for market makers. In contrast to those theories, however, our model makes the novel prediction that I-investors might *also* pay a smaller spread when PFOF is allowed.

Intuition for why $s_2 < \bar{s}$. Given the maintained assumption that $\Delta < 0$, R-investors are less costly for market makers to intermediate than I-investors. When PFOF is allowed, R-investors therefore pay a smaller spread. That is, $s_2 < s_1$, implying $s_2 < \bar{s}$.

Intuition for why $\bar{s} < s_b$. Let f_R denote the fraction of volume due to R-orders. When PFOF is banned, volume is proportional to market size, so that $f_R^{\text{ban}} = \frac{\omega_R}{\omega_R+\omega_I}$. When PFOF is allowed, and given $\Delta < 0$, market makers siphon R-orders off-exchange, meaning that $f_R^{\text{no-ban}} > f_R^{\text{ban}}$. In either case, the volume-weighted average spread $\bar{s}$ is determined by the intersection of an average liquidity demand curve and an average liquidity supply curve:

> **Lemma 2 (Average liquidity demand and supply curves).** Investors' average liquidity demand and market makers' average liquidity supply curves are given, respectively, by
>
> $$\bar{s}(x; f_R) = \zeta - v(f_R)x \text{ and } \bar{s}(x; f_R) = \frac{\gamma}{2} + c(f_R)x,$$
>
> where x is the aggregate volume, and $v(\cdot)$ and $c(\cdot)$ are the respective curves' slopes:
>
> $$v(f_R) = \frac{(1-f_R)^2}{\omega_I} + \frac{f_R^2}{\omega_R} \text{ and } c(f_R) = \frac{\gamma}{M}\text{var}\big[(1-f_R)D_I + f_R D_R\big].$$

The conclusion $\bar{s} < s_b$ follows because demand is steeper and supply is flatter when PFOF is allowed than when it is banned.

- Demand is at its flattest when both investor types face the same pricing, as under a PFOF ban. Mathematically, $v(f_R)$ is minimized at $f_R^{\text{ban}} = \frac{\omega_R}{\omega_R+\omega_I}$. To understand the intuition, compare the following two extremes. On the one hand, a value $f_R = \frac{\omega_R}{\omega_R+\omega_I}$ obtains when both investor types are charged the same spread s. In that case, total inverse liquidity demand is $\lambda_R(s) + \lambda_I(s) = (\zeta - s)(\omega_R + \omega_I)$, implying a demand curve with slope $v\left(\frac{\omega_R}{\omega_R+\omega_I}\right) = \frac{1}{\omega_R+\omega_I}$. On the other hand, a value $f_R = 1$ obtains when all I-investors are priced out of the market. In that case, total inverse liquidity demand is entirely determined by the R-investors: $\lambda_R(s) = (\zeta - s)\omega_R$, implying $v(1) = \frac{1}{\omega_R}$. Being at its flattest when PFOF is banned, demand must be steeper when PFOF is allowed.

- Furthermore, supply is flatter when PFOF is allowed, in the sense that the ensuing increase in f_R causes a decrease in $c(f_R)$. The intuition is that when PFOF is allowed, market makers siphon R-orders, precisely because doing so reduces their inventory costs. Formally, $c(f_R^{\text{no-ban}}) < c(f_R^{\text{ban}})$ follows from two facts. First, $c(\cdot)$ is quadratic and convex. Indeed, we can compute $c(f_R) = \frac{\gamma}{M}\left[(1-f_R)^2\sigma_I^2 + 2f_R(1-f_R)\rho\sigma_I\sigma_R + f_R^2\sigma_R^2\right]$. Second, $f_R^{\text{ban}} < f_R^{\text{no-ban}} \leq \arg\min_{f_R} c(f_R)$. We have already seen that $f_R^{\text{ban}} < f_R^{\text{no-ban}}$. To see $f_R^{\text{no-ban}} \leq \arg\min_{f_R} c(f_R)$, consider a benchmark in which R- and I-orders could be procured at the same spread. In that benchmark, a market maker would optimally procure a portfolio in which the fraction of R-orders was $\arg\min_{f_R} c(f_R)$. But because $\Delta < 0$, we have $s_2 \leq s_1$ in the no-ban equilibrium, which makes R-orders (weakly) less attractive to market makers than they would be in the aforementioned benchmark. As a result, $f_R^{\text{no-ban}} \leq \arg\min_{f_R} c(f_R)$.

Intuition for why $s_1 < s_b$ is possible. Recall the maintained assumption $\Delta < 0$, which, following the discussion after (11), implies that R-orders exert a lower marginal inventory cost on market makers than I-orders. Under a PFOF ban, both types are pooled together and are priced at the volume-weighted average of their marginal costs. Once PFOF is allowed, each type is charged a spread equal to its own marginal cost, creating two effects:

- The first effect can be understood through what would happen if liquidity demand were perfectly inelastic. In that case, the less costly R-investors would be charged less ($s_2 < s_b$), and the more costly I-investors would be charged more ($s_1 > s_b$).

- Crucially, however, liquidity demand is not perfectly inelastic. If R-investors are charged less, then R-investor volume increases. How this affects the marginal inventory cost of I-orders depends on ρ. If $\rho > 0$ ($\rho < 0$), additional R-investor volume raises (lowers) the marginal cost of I-orders, reinforcing (counteracting) the first effect.

It follows that $s_1 < s_b$ if and only if ρ is sufficiently negative. Proposition 4 provides the exact condition.[19] The economic force discussed above echoes a concern raised by the SEC in its recently proposed Order Competition Rule (p. 298, SEC, 2022): "... a reduction in the volume of individual investor order flow internalized by wholesalers could increase wholesaler inventory risk, which in turn could cause wholesalers to reduce the liquidity they supply as exchange market makers or to institutional investors..."

Empirical predictions. Proposition 4 constitutes an empirical prediction regarding how off-exchange siphoning affects on-exchange liquidity. Much empirical literature has examined this question. On the one hand, on-exchange liquidity does not seem to have been harmed by off-exchange siphoning in certain settings (e.g., Battalio, 1997; Battalio, Greene, and Jennings, 1997; Garriott and Walton, 2018; Elsas, Johanning, and Theissen, 2022), which is consistent with our inventory-based theory but inconsistent with the information-based theories mentioned earlier. On the other hand, on-exchange liquidity does seem to have deteriorated in response to off-exchange siphoning in other settings (e.g., Degryse, de Jong, and van Kervel, 2015; Hatheway, Kwan, and Zheng, 2017; Comerton-Forde, Malinova, and Park, 2018; Hu and Murphy, 2022), which is consistent both with our inventory-based theory and with the information-based ones.

Discussion of Figure 1. Figure 1 illustrates bid-ask spreads as a function of ρ under parameters consistent with the discussion in Section 3.3: we set $\sigma_R = 0.2$ and $\sigma_I = 0.5$ to be consistent with the estimates of Jones et al. (2022), and we also set $\omega_R/\omega_I = 1/2$. The correlation parameter ρ is

[19] According to Proposition 4, $s_2 < s_b$ holds unambiguously, while $s_1 \lesseqgtr s_b$. This asymmetry between s_1 and s_2 arises due to the maintained assumption that $\Delta < 0$, which ensures that a marginal R-order is less costly than a marginal I-order, so that $s_2 < \bar{s} < s_1$. Previous analysis has shown $\bar{s} < s_b$, thus ensuring $s_2 < s_b$.

on the horizontal axis. As Proposition 4 states: (i) $s_b > s_2$ and $s_b > \bar{s}$ in the figure, regardless of ρ, and (ii) $s_b > s_1$ if and only if ρ is sufficiently negative, where for these parameters, the precise cutoff is $\rho = -1/2$. The estimates of Jones et al. (2022) indicate correlations not far from this cutoff, suggesting that $s_1 < s_b$ is in fact a realistic possibility.

The figure also indicates how spreads vary with ρ.[20] Roughly speaking, there are two effects. First, as ρ increases, R- and I-orders become less likely to offset. A given order portfolio therefore creates greater inventory risk, so market makers require larger spreads to compensate. This effect drives the initial increase of all spreads seen in Figure 1. Second, as spreads change, investors' participation decisions may change, and a market maker's order portfolio changes as well. For the parametrization of Figure 1, in the limit as $\rho \to 1$, $s_1 \to \zeta$, and I-investors stop participating altogether, which reduces the marginal cost of R-orders, driving the subsequent decrease of s_2.

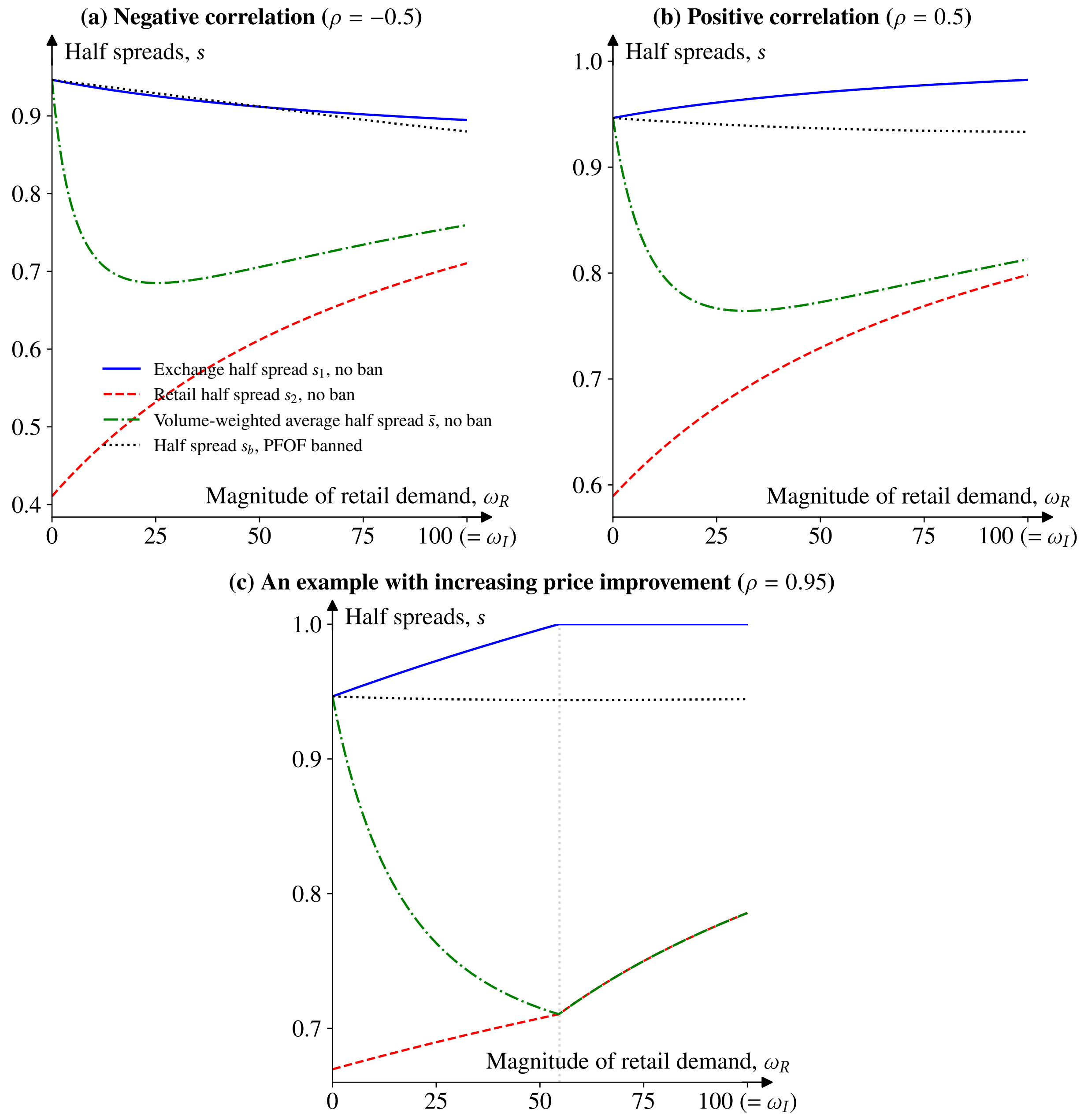
4.1.2 The rise of retail trading

Recent years have seen a rapid growth in retail trading activity. In the U.S. equity market, retail trading volume doubled from about \$15 billion per day before 2017 to about \$30 billion in 2022 (Mackintosh, 2022). This trend can be modeled as an increase in the parameter ω_R.

Proposition 5 (Rise of retail trading and spreads). As the magnitude of retail demand ω_R increases,

- the off-exchange half spread, s_2, monotonically increases;
- the on-exchange half spread, s_1, monotonically increases (decreases) if the order flow correlation $\rho > 0$ (< 0); and
- the price improvement, $s_1 - s_2$, monotonically decreases (increases) if $\rho < \hat{\rho}$ ($> \hat{\rho}$), where the threshold $\hat{\rho}$ is a function of other parameters (given in (22) in the proof) and is strictly positive.

[20] See Appendix C for a formal analysis of how spreads vary with ρ under the additional assumption that $\sigma_R = \sigma_I$.



(a) **Negative correlation** ($\rho = -0.5$)

(b) **Positive correlation** ($\rho = 0.5$)

(c) **An example with increasing price improvement** ($\rho = 0.95$)

Figure 2: Bid-ask spreads: the rise of retail trading. This figure shows how various (half) bid-ask spreads change as the retail trading demand increases. Panels (a) and (b) illustrate the cases of negative and positive order flow correlations ($\rho = -0.5$ and $\rho = 0.5$, respectively) between retail and institutional order flows. In both Panels (a) and (b), price improvement $s_1 - s_2$ is decreasing in ω_R; Panel (c) illustrates an example with increasing price improvement (with $\rho = 0.95$). The magnitude of retail demand ω_R varies on the horizontal axes. The other parameters are commonly set at $M = 3$, $\gamma = \zeta = 1$, $\omega_I = 100$, $\sigma_I = 0.5$, $\sigma_R = 0.2$, and $\mu_I = \mu_R = 0$. The vertical dotted line in (c) indicates the upper bound on ω_R, implied by (12): when ω_R exceeds that bound, no trading occurs on-exchange.

Figure 2 illustrates these effects. The off-exchange spread s_2 (the dashed line) rises with ω_R. This is because, rather intuitively, the increase in R-investor liquidity demand requires market makers to handle larger volumes, hence also higher inventory costs.

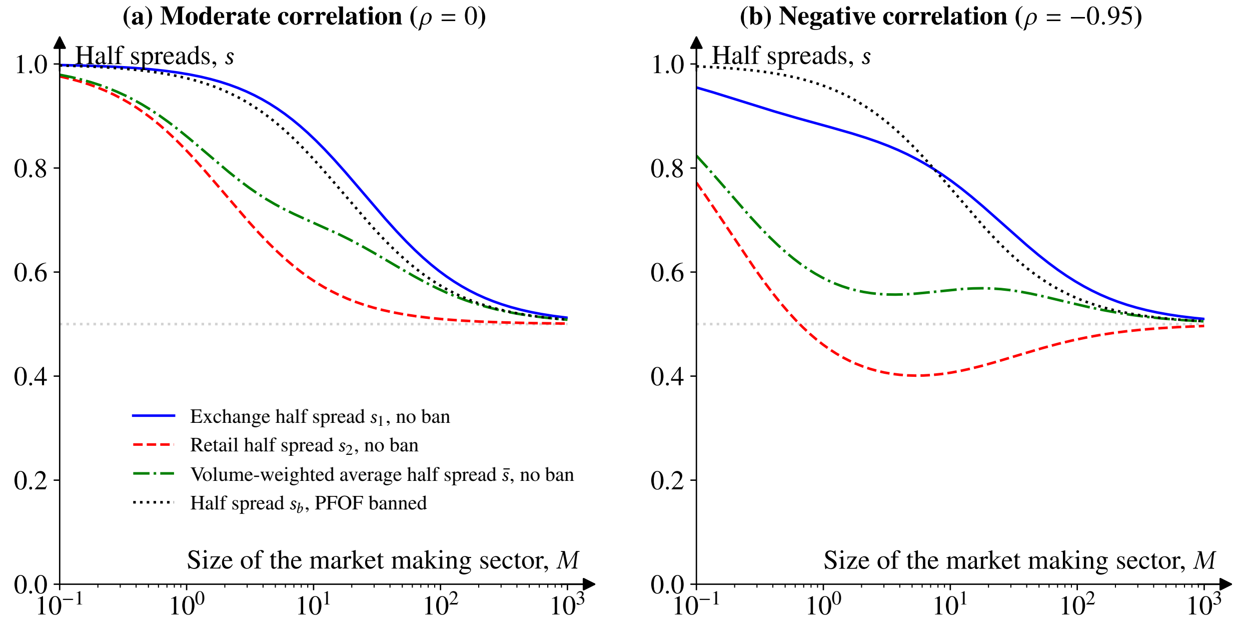
The effect on the on-exchange spread s_1 (the solid line) is more nuanced. As can be seen by comparing Panel (a) and (b), it depends on the order flow correlation ρ. If $\rho > 0$ (< 0), the increase in R-orders worsens (alleviates) the market makers' overall inventory costs through its correlation with the I-orders. In other words, as the "retail army" rises, it exerts a negative (positive) externality on other investors who on average trade in the same (opposite) direction.

The same force underlies how ω_R affects the price improvement $s_1 - s_2$. Clearly, when $\rho < 0$, $s_1 - s_2$ decreases with ω_R, because s_1 decreases and s_2 increases. If $\rho > 0$ instead, then the effect depends on the relative speed of the increases in s_1 and in s_2. As Proposition 5 states, s_1 is faster only if ρ is sufficiently positive. Intuitively, this is exactly the case where, as ω_R increases, R-investors' negative externality on I-investors is particularly strong, thus pushing up s_1 very quickly. Figure 2(c) depicts such an example.

Our predictions regarding how retail trading activity affects spreads and price improvement can be empirically tested. In particular, Proposition 5 predicts that ω_R affects the on-exchange spread s_1, and moreover that the direction of the effect depends on the sign of the order flow correlation ρ. These are novel predictions. To compare, under the information-based theories of PFOF, as long as all (uninformed) R-orders are siphoned off-exchange, the adverse-selection risk of the on-exchange I-orders is unaffected by ω_R, and so is s_1.

4.1.3 The size of the market making sector

Another focal point in debates over PFOF is the concentration of the market making sector (see, e.g., Hu and Murphy, 2022). Would the entry of additional market makers drive down investors' trading costs? Our model generates predictions along this line via comparative statics with respect to the size of the market-making sector, M. Perhaps surprisingly, the implications of an increase



(a) Moderate correlation ($\rho = 0$)

(b) Negative correlation ($\rho = -0.95$)

Figure 3: Bid-ask spreads: the size of the market making sector. This figure shows how various (half) bid-ask spreads change as the market making sector grows. Panels (a) and (b) illustrate the cases of moderate and negative order flow correlations, $\rho = 0$ and $\rho = -0.95$, respectively. The size of the market making sector M varies from 0.1 to 1,000 on the horizontal axes. The other parameters are set at $\gamma = \zeta = 1$, $\omega_I = 100$, $\omega_R = 50$, $\sigma_I = 0.5$, $\sigma_R = 0.2$, and $\mu_I = \mu_R = 0$.

in M are quite nuanced.

Figure 3 plots the various spreads against M. In the limit of $M \to \infty$, in both panels, the spreads converge to $\frac{\gamma}{2}$, which is 0.5 in the numerical illustration. This is because, in the limit, an infinite measure of market makers compete for a finite measure of orders, and, therefore, each market maker expects to receive at most one order. Hence, the limiting spread equals the marginal cost of one unit of inventory, which is $\frac{\gamma}{2}$.

However, the convergence to $\frac{\gamma}{2}$ is not always monotone. In particular, if the order flow correlation ρ is sufficiently negative, as in Panel (b), the off-exchange spread s_2 is U-shaped. Accordingly, the volume-weighted average spread $\bar{s}$ is also non-monotone. This implies that a larger market making sector, perhaps surprisingly, might actually raise investors' trading costs.

Proposition 6 (Size of the market making sector and spreads). As the size of the market making sector M increases,

- the on-exchange half spread, s_1, monotonically decreases;
- the off-exchange half spread, s_2, initially decreases but eventually increases (i.e., is U-shaped in M) if $\rho < -\frac{\sigma_R \omega_R}{\sigma_I \omega_I}$, and it monotonically decreases otherwise.

To see how this happens, note that the off-exchange spread s_2 may, in fact, drop below $\frac{\gamma}{2}$ under the parameterization of Figure 3(b). Were it not for I-orders, market makers would lose money when providing liquidity to R-investors at such a spread, following their payoff expression (4). Why are they willing to provide liquidity to R-orders at such a small spread? This is because the acquired R-orders are very useful in hedging I-orders, thanks to the negative order flow correlation ρ. In other words, the inventory cost savings from hedging I-orders subsidizes market makers' losses in providing liquidity to R-orders. Mapping to the real world, the above "subsidy" interpretation of our mechanism illuminates why market makers are willing to provide significant price improvements to their purchased retail orders (*cf.* Remark 5).

Let us now turn to the U-shape of s_2, driven by two countervailing effects of M. First is an intuitive "supply effect:" As M increases, the total liquidity supply $Mx_{m2}(\boldsymbol{s})$ to R-investors increases, and the market makers effectively walk down the decreasing demand curve given by $\lambda_R(s_2)$. Second, as explained in the previous paragraph, s_2 may decrease below $\frac{\gamma}{2}$ when ρ is sufficiently negative. In that case, s_2 must be eventually increasing in M, because, as we have previously observed, s_2 converges to $\frac{\gamma}{2}$ (the marginal cost of the first unit of inventory) as $M \to \infty$.

4.2 Market makers' profitability

We now turn to market makers' profitability, denoted π with PFOF and π_b under the ban. A key insight from the model is that the market makers might face a "prisoner's dilemma" in which each unilaterally wants to siphon R-orders off-exchange and yet, collectively they would be better off if

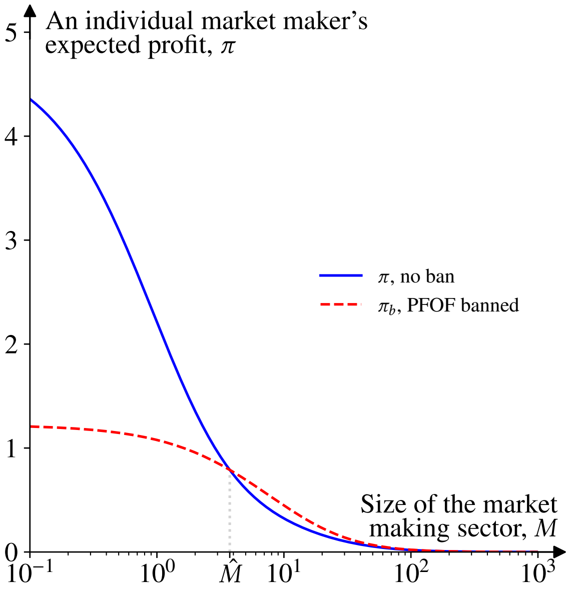


Figure 4: Market makers' profits. This figure shows how the size of the market making sector M affects a market maker's expected profit π with versus without a PFOF ban. The other parameters are set at $\gamma = \zeta = 1$, $\omega_I = 100$, $\omega_R = 50$, $\sigma_I = 0.5$, $\sigma_R = 0.2$, $\rho = -0.3$, and $\mu_I = \mu_R = 0$.

they all refrained from doing so. In such cases, a PFOF ban actually benefits market makers. Such a prisoner's dilemma can be seen in Figure 4 where the solid line falls below the dashed line—with sufficiently many market makers.

Proposition 7 (Market makers' profits). Market makers' profits are higher under the PFOF ban, i.e., $\pi > \pi_b$, if and only if $M < \hat{M}$, where $\hat{M}$ denotes the unique strictly positive root of a cubic polynomial given by equation (23) in the proof.

How does the prisoner's dilemma arise in this context? As we have seen in Section 3.2, under the maintained assumption that $\Delta < 0$, each market maker individually benefits from practicing PFOF, as it provides flexibility for tailoring exposure to R- and I-order flows. However, when all market makers collectively practice PFOF, that changes the order flow composition on-exchange (i.e., the $\boldsymbol{F}(\alpha)$ matrix) as well as the liquidity demand intensities $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$. Equilibrium spreads then also change—potentially in a way that harms market makers. Indeed, Proposition 4 has established that the volume-weighted average spread is lower when PFOF is allowed (i.e., $\bar{s} < s_b$).

A prisoner's dilemma emerges when the negative pecuniary externality (lowered spread revenues) outweighs the individual benefit of PFOF (flexibility for tailoring order flow exposures). Externalities often loom large when more parties are involved, so that, as indicated by the proposition, the prisoner's dilemma arises when M is sufficiently large.

One puzzle is why some market makers (e.g., Citadel as quoted in Footnote 3) have been quite open to a potential PFOF ban—after all, such a ban would undermine a core aspect of their current business model. Our analysis highlights a novel explanation for this puzzle: a PFOF ban might actually benefit market makers by resolving a prisoner's dilemma among them.

Comparison with information-based theories of PFOF. Inventory costs are central to this prisoner's dilemma. To see this, compare our model to Easley, Kiefer, and O'Hara (1996) and other existing models of PFOF, which do not feature inventory costs (but asymmetric information instead). In these models, market makers earn zero profits—both in the equilibrium with PFOF and in the equilibrium when PFOF is banned—so that this type of prisoner's dilemma cannot arise.

4.3 Welfare

Finally, let us examine implications on welfare. We first derive a general expression for total welfare—the sum of the investors' and the market makers' surplus. To do so, we first define x_I and x_R as the Poisson rates at which each individual market maker expects to receive R- and I-orders. Similarly, define s_I and s_R as the respective spreads charged to R- and I-orders.

Now we consider investor surplus. Under the maintained assumption of this section, $s_k \leq \zeta$ for $k \in \{I, R\}$ holds in equilibrium both with and without the PFOF ban. Then the type-k investors' inverse demand is $s(q) = \zeta - \frac{q}{\omega_k}$ and their surplus can be computed as

$$\int_0^{Mx_k} \left(\zeta - \frac{q}{\omega_k} - s_k\right) \mathrm{d}q = (Mx_k)\zeta - \frac{(Mx_k)^2}{2\omega_k} - (Mx_k)s_k.$$

An individual market maker's surplus is given by

$$x_I s_I + x_R s_R - \frac{\gamma}{2}(x_I + x_R) - \frac{\gamma}{2}(x_I^2\sigma_I^2 + 2\rho\sigma_I\sigma_R x_I x_R + x_R^2\sigma_R^2).$$

Summing the above, noting that there is a measure of M market makers in total, we obtain the welfare expression:

$$w(x_I, x_R) = \sum_{k\in\{I,R\}} \left[(Mx_k)\left(\zeta - \frac{\gamma}{2}\right) - \frac{(Mx_k)^2}{2\omega_k}\right] - \frac{M\gamma}{2}\left(x_I^2\sigma_I^2 + 2\rho\sigma_I\sigma_R x_I x_R + x_R^2\sigma_R^2\right). \tag{14}$$

By substituting the corresponding equilibrium supplies x_k, this welfare expression applies generally to any equilibrium. Let w and w_b respectively denote equilibrium welfare with PFOF and under the ban. In the model, a PFOF ban can only reduce welfare (i.e., $w_b < w$). This follows from a stronger result—that the equilibrium without a PFOF ban in fact leads to the welfare-maximizing (x_I, x_R). Figure 5 illustrates.

Proposition 8 (PFOF ban and welfare). Absent a PFOF ban, the equilibrium outcome maximizes total welfare, and, hence, $w > w_b$.

That the equilibrium without a PFOF ban leads to the welfare-maximizing outcome is essentially a consequence of the First Welfare Theorem. For example, our model features competitive pricing, no externalities (aside from pecuniary ones), and separate "prices" (spreads s_1 and s_2) for each and every different "good" (liquidity to R- and I-orders). With a PFOF ban in place, the First Welfare Theorem no longer applies, for we then have only a single "price" (s_b) for the two separate "goods." The reason for a strict (rather than weak) inequality in Proposition 8 is the maintained assumption $\Delta < 0$, which implies that the PFOF ban bites.

Our result adds to recent policy discussions about PFOF. Both in the U.S. and in Europe, financial market regulators have expressed concerns regarding PFOF as well as intentions to ban

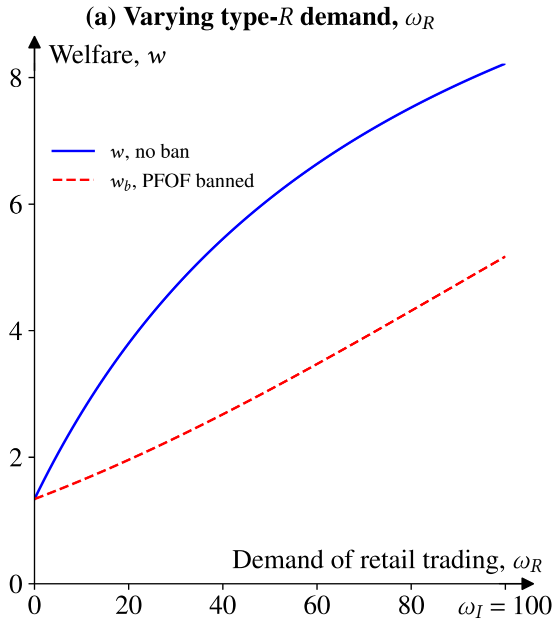


Figure 5: Welfare. This figure shows how total welfare is affected by the magnitude of retail liquidity demand in Panel (a) and the order flow correlation ρ in Panel (b). For Panel (a), $\rho = -0.3$; for Panel (b), $\omega_R = 50$. The other parameters are set at $M = 3$, $\gamma = \zeta = 1$, $\omega_I = 100$, $\sigma_I = 0.5$, $\sigma_R = 0.2$, and $\mu_I = \mu_R = 0$.

it. Their arguments largely refer to negative effects of PFOF that are not captured by our model. For example, some have argued that the practice poses conflicts of interest, as a broker would "choose the [market maker] offering the highest payment, rather than the best possible outcome for its clients" (ESMA, 2021). The SEC chair, Gary Gensler, also warns that via PFOF, "[market makers] get the data, they get the first look, they get to match off buyers and sellers out of that order flow" (Barron's, 2021). Our welfare result, as stated in Proposition 8 above, cuts in the opposite direction. We contribute to these discussions by highlighting the benefit of PFOF in a stylized setting featuring market makers' inventory concerns.

Comparison with information-based theories of PFOF. The first part of Proposition 8 says that, without a PFOF ban, the equilibrium outcome maximizes welfare. No analogous result typically holds in information-based theories of PFOF, for the reason that adverse selection generally invalidates the First Welfare Theorem. The second part of Proposition 8 says that a PFOF ban would reduce welfare. No analogous welfare comparison is made in many information-based models of PFOF (e.g., Easley, Kiefer, and O'Hara, 1996; Battalio and Holden, 2001), for the reason that uninformed investors are assumed to be price-inelastic in those models, so that a PFOF ban has no effect on welfare. Incorporating an elasticity into those models could, however, permit an analogous result in certain cases.

5 A dynamic extension

In the model, market makers' liquidity supply decisions can be interpreted as their limit orders, their ex-ante allocation of groundwork, or their negotiations with brokers (see Remark 3). In Sections 2 and 3, we assume that these decisions are made once, before trading starts. In this section, we relax this assumption and examine how market makers supply liquidity dynamically.

In Section 5.1, we describe the model setup. To keep the analysis tractable and to highlight novel insights, we also introduce a few simplifying assumptions. We then characterize the equilibrium in Section 5.2, showing that market makers' incentive to siphon R-orders remains intact. In fact, this dynamic extension highlights a novel effect that further incentivizes siphoning. This new mechanism arises from two unique features of the dynamic extension: (1) market makers accumulate inventories from previous periods, and (2) these inventories can be correlated with future orders because of autocorrelation in order flows.

5.1 Setup

Timing. There are two trading periods $t \in \{1, 2\}$. As in previous sections, each period lasts for one unit of time. As in Section 3, there are two marketplaces $j \in \{1, 2\}$, where 1 denotes on-exchange and 2 off-exchange trading. In each period t, each market maker $m \in [0, M]$ chooses her liquidity supplies for the two marketplaces $\boldsymbol{x}_m^{(t)} = \left(x_{m1}^{(t)}, x_{m2}^{(t)}\right)^\top$, where the superscript "(t)" indicates the period t. The period-t half spreads for these two marketplaces are denoted by $\boldsymbol{s}^{(t)} = \left(s_1^{(t)}, s_2^{(t)}\right)^\top$ and will be determined endogenously in equilibrium. A market maker's payoff is her spread revenue (across both marketplaces and both time periods) minus $\frac{\gamma}{2}$ times the square of her terminal inventory.

Liquidity demand. As in Section 3, there are two types of investors, $k \in \{R, I\}$. For simplicity, their liquidity demands are assumed to have time-invariant intensities, i.e., $\lambda_k^{(t)}(s) = \lambda_k(s) = \max\{0, (\zeta - s)\omega_k\}$, for both $t \in \{1, 2\}$. Whereas we permitted an arbitrary joint distribution for directionalities in previous sections, we impose additional structure here so as to introduce autocorrelation in a tractable way. Formally, we model directionalities as follows:

$$D_I^{(1)} = (-1)^{Y_I^{(1)}} \sigma_I, \; D_I^{(2)} = X_I D_I^{(1)} + (1 - X_I)(-1)^{Y_I^{(2)}} \sigma_I, \text{ and } D_R^{(1)} = D_R^{(2)} = 0, \tag{15}$$

where $\left\{Y_I^{(1)}, Y_I^{(2)}, X_I\right\}$ are independent Bernoulli draws with respective success rates $\left\{\frac{1}{2}, \frac{1}{2}, \phi_I\right\}$ with $\phi_I \in [0, 1)$ and $\sigma_I \in (0, 1]$. In words, for each period t, I-orders will realize a directionality $D_I^{(t)}$ of either $\pm\sigma_I$, yet are unconditionally balanced, with $\mathbb{E}\left[D_I^{(t)}\right] = 0$. The period-2 directionality $D_I^{(2)}$ remains equal to the period-1 directionality $D_I^{(1)}$ with probability ϕ_I and is an i.i.d. new draw with probability $1 - \phi_I$. In contrast, we assume that R-orders lack directionality altogether, which lends tractability while simultaneously encoding the realistic feature that R-orders are less directional than I-orders (as discussed in Section 3.3).[21]

Siphoning. In each period $t \in \{1, 2\}$, a fraction $\alpha^{(t)} \in [0, 1]$ of the R-orders is endogenously siphoned off-exchange, so that the marketplace-level directionalities are $\boldsymbol{D}^{(t)} = \boldsymbol{F}\left(\alpha^{(t)}\right)\left(D_I^{(t)}, D_R^{(t)}\right)^\top$, where the weighting matrix $\boldsymbol{F}(\cdot)$ remains as in (6). As in Section 3, brokers honor the best-execution requirement, so that $s_1^{(t)} \leq (\geq) s_2^{(t)}$ if $\alpha^{(t)} < 1 (> 0)$; see (7a)–(7b) and Remark 5.

[21] Our analysis can be generalized to the case where R-order directionalities take the same functional form as for I-orders: $D_R^{(1)} = (-1)^{Y_R^{(1)}} \sigma_R$, and $D_R^{(2)} = X_R D_I^{(1)} + (1 - X_R)(-1)^{Y_R^{(2)}} \sigma_R$, where $\left\{Y_R^{(1)}, Y_R^{(2)}, X_R\right\}$ are independent Bernoulli draws with respective success rates $\left\{\frac{1}{2}, \frac{1}{2}, \phi_R\right\}$ with $\phi_R \in [0, 1)$ and $\sigma_R \in (0, 1]$. In particular, the same key result (that retail orders are all siphoned off-exchange) will be obtained for sufficiently small $\sigma_R > 0$.

Trading outcomes from $t = 1$. After the $t = 1$ trading, a market maker m observes $\mathcal{I}_m^{(1)} = \{z_m^{(1)}, D_I^{(1)}, \bar{z}^{(1)}\}$, where $z_m^{(1)}$ is the market maker's own inventory at that time, and $\bar{z}^{(1)} := \frac{1}{M}\int_0^M z_m^{(1)} \mathrm{d}m$ is the average inventory across all market makers. Naturally, the market maker knows her own $z_m^{(1)}$. She can further infer $D_I^{(1)}$ and $\bar{z}^{(1)}$, for example, from a public data feed.[22]

Equilibrium definition. Analogous to Section 3.1, an equilibrium consists of, for both $t \in \{1, 2\}$, liquidity supply intensities $\boldsymbol{x}_m^{(t)}$ for each market maker m, siphoning fractions $\alpha^{(t)}$, and half spreads $\boldsymbol{s}^{(t)}$. In particular, for $t = 2$, a market maker m's supply $\boldsymbol{x}_m^{(2)}$ can depend on her own observation $\mathcal{I}_m^{(1)}$, while the market-wide variables, $\alpha^{(2)}$ and $\boldsymbol{s}^{(2)}$, can depend on $\cup_{m\in[0,M]}\mathcal{I}_m^{(1)}$. The equilibrium conditions determining these endogenous variables are: (i) each market maker m chooses her liquidity supply strategy $\{\boldsymbol{x}_m^{(1)}, \boldsymbol{x}_m^{(2)}\}$ to maximize her expected payoff—not only in the entire game but also in each $t = 2$ subgame (as in the spirit of subgame perfection); (ii) market clearing holds for each marketplace j (in every subgame), and (iii) the best-execution requirement is satisfied (in every subgame).

Comparison with existing literature. A key feature of our analysis is that we consider an environment with two marketplaces, so as to highlight endogenous order flow segmentation—the off-exchange siphoning of retail orders. In contrast, existing inventory-based models of dynamic liquidity supply typically consider only a single venue; see, e.g., Amihud and Mendelson (1980), Ho and Stoll (1981, 1983), and Hendershott and Menkveld (2014).[23]

[22] For example, such a data feed may contain the on-exchange half spread $s_1^{(1)}$, trading volume $V_1^{(1)} := \lambda_I(s_1^{(1)}) + (1 - \alpha^{(1)})\lambda_R(s_1^{(1)})$, and order imbalance $I_1^{(1)} := \lambda_I(s_1^{(1)})D_I^{(1)} + (1 - \alpha^{(1)})\lambda_R(s_1^{(1)})D_R^{(1)} = \lambda_I(s_1^{(1)})D_I^{(1)}$ (where the last equality holds because $D_R^{(1)} = 0$). Each market maker can infer $D_I^{(1)}$ by solving those two equations for the two unknowns $D_I^{(1)}$ and $\alpha^{(1)}$. The average inventory also follows $\bar{z}^{(1)} = -\frac{1}{M}I_1^{(1)}$ (note that the off-exchange imbalance $I_2^{(2)} := \alpha^{(1)}\lambda_R(s_2^{(1)})D_R^{(1)}$ is zero and hence does not affect $\bar{z}^{(1)}$).

[23] In addition to considerations related to inventory management, the literature has also highlighted several other dimensions of strategic behavior in dynamic liquidity provision. For example, Kyle (1985) demonstrates how competitive market makers dynamically supply liquidity in view of information asymmetry and its resolution over time. Bernhardt et al. (2004) show that dealers provide better liquidity to frequent customers to secure future business. Desgranges and Foucault (2005) show that repeated trading relationships can shield a dealer from being adversely selected by a possibly informed investor. Barbon et al. (2019) find evidence consistent with brokers leaking information about some of their clients' fire-selling orders to other clients.

5.2 Equilibrium

The equilibrium is solved backwards: We first derive the $t = 2$ equilibrium objects for any given $t = 1$ trading outcomes. This gives market makers' continuation values, with which we then solve for the $t = 1$ equilibrium objects.

5.2.1 Period 2

The analysis for period 2 is similar to that for the single-period model, with two key differences. First, given her observation $\mathcal{I}_m^{(1)}$ from $t = 1$, each market maker m possesses information about the $t = 2$ directionality of I-orders, $D_I^{(2)}$. In particular, all market makers observe the realized $t = 1$ directionality $D_I^{(1)}$, which following (15) is a sufficient statistic. For notational simplicity, we write $\mathbb{E}_1[\cdot] = \mathbb{E}\left[\,\cdot\,\middle|D_I^{(1)}\right]$ and $\mathrm{var}_1[\cdot] = \mathrm{var}\left[\,\cdot\,\middle|D_I^{(1)}\right]$. Therefore, by Bayes' rule, every market maker obtains the same posterior moments: $\mathbb{E}_1\left[D_I^{(2)}\right] = \phi_I D_I^{(1)}$, $\mathrm{var}_1\left[D_I^{(2)}\right] = (1 - \phi_I^2)\sigma_I^2$, and we of course also have $\mathbb{E}_1\left[D_R^{(2)}\right] = 0$ and $\mathrm{var}_1\left[D_R^{(2)}\right] = 0$. We write the posterior mean and variance of $\boldsymbol{D}^{(2)} = \boldsymbol{F}\left(\alpha^{(2)}\right) \cdot \left(D_I^{(2)}, D_R^{(2)}\right)^\top$ as $\boldsymbol{\mu}^{(2|1)}$ and $\Sigma^{(2|1)}$, respectively, and derive their expressions in the proof of Lemma 3.

Second, in the single-period model, all market makers begin with zero inventory, whereas now in $t = 2$, each market maker m inherits from her trading at $t = 1$ the inventory $z_m^{(1)}$. We show in the proof of Lemma 3 that her objective now becomes:

$$\begin{aligned}\pi_m^{(2)}\left(\boldsymbol{x}_m^{(2)}; z_m^{(1)}, D_I^{(1)}\right) = \boldsymbol{x}_m^{(2)\top}\left(\boldsymbol{s}^{(2)} - \frac{\gamma}{2}\mathbf{1}\right)&\\ - \frac{\gamma}{2}\left[\boldsymbol{x}_m^{(2)\top}\left(\Sigma^{(2|1)} + \boldsymbol{\mu}^{(2|1)}\boldsymbol{\mu}^{(2|1)\top}\right)\boldsymbol{x}_m^{(2)} - 2\boldsymbol{x}_m^{(2)\top}\boldsymbol{\mu}^{(2|1)}z_m^{(1)} + \left(z_m^{(1)}\right)^2\right],&\end{aligned} \quad (16)$$

where she takes as given the half spreads $\boldsymbol{s}^{(2)}$ and the siphoning fraction $\alpha^{(2)}$ (which determines $\Sigma^{(2|1)}$ and $\boldsymbol{\mu}^{(2|1)}$). Compared with the objective (4) in the single-period model, the last two terms in the squared-brackets are new. They arise from the market maker's existing inventory $z_m^{(1)}$: The expected inventory cost created by that existing inventory itself is proportional to $\left(z_m^{(1)}\right)^2$, and that created by its covariance with her $t = 2$ trading is proportional to $-\boldsymbol{x}_m^{(2)\top}\boldsymbol{\mu}^{(2|1)}z_m^{(1)}$.

We show in the proof of Lemma 3 that the objective (16) is strictly concave in $x_{m1}^{(2)}$ but linear in $x_{m2}^{(2)}$. Therefore, the first-order condition determines the optimal $x_{m1}^{(2)}$ (as a function of the on-exchange half spread $s_1^{(2)}$) and the off-exchange half spread $s_2^{(2)}$. Market clearing then determines the on-exchange half spread $s_1^{(2)}$ and the *aggregate* off-exchange liquidity supply. The equilibrium siphoning fraction $\alpha^{(2)}$ is determined by the best-execution requirement. The following lemma summarizes the results:

Lemma 3 (Period $t = 2$). Given period-1 outcomes $D_I^{(1)}$ and $\bar{z}^{(1)}$, the period-2 continuation game always has an equilibrium. All equilibria are characterized as follows. Define

$$\Delta^{(2)} := \frac{\phi_I D_I^{(1)} M \bar{z}^{(1)}}{\zeta - \frac{\gamma}{2}} - \sigma_I^2 \omega_I. \tag{17}$$

(i) If $\Delta^{(2)} < 0$, then $\alpha^{(2)} = 1$.

(ii) If $\Delta^{(2)} > 0$, then $\alpha^{(2)} = 0$.

(iii) If $\Delta^{(2)} = 0$, then $\alpha^{(2)}$ can take any value in [0,1].

In all cases, the equilibrium half spreads $s_1^{(2)}$ and $s_2^{(2)}$ are defined by equations (26)–(27); equilibrium on-exchange liquidity supply $x_{m1}^{(2)}$ is given by (28); and equilibrium off-exchange aggregate liquidity supply $\int_0^M x_{m2}^{(2)} \mathrm{d}m$ is given by (29) (but may be allocated arbitrarily among the individual market makers). Furthermore, when the market clears at $t = 1$ (as it would in equilibrium of the full game), we have $\alpha^{(2)} = 1$ (i.e., all R-orders are siphoned off-exchange in $t = 2$).

According to the last part of the lemma, in (the full game) equilibrium, all R-orders are siphoned off-exchange in $t = 2$. Two effects drive this equilibrium feature.

- First, suppose (contrary to the equilibrium) that all market makers were to enter period 2 without any inventory. This would imply that $\bar{z}^{(1)} = 0$, leading to $\Delta^{(2)} = -\omega_I \sigma_I^2 < 0$ by equation (17). Thus, even absent the inventories inherited from $t = 1$, R-orders would be siphoned off-exchange. The intuition is that R-orders, being balanced, would be cheaper than I-orders, being imbalanced, in terms of market makers' inventory costs. In fact, this first effect is implied by the analysis in Section 3, where R-orders are siphoned off-exchange if $\Delta < 0$.[24] In other words, the siphoning incentive identified in the single-period case remains robust in the dynamic extension.

- Second, the inventory that market makers bring into period 2 leads to an extra term in the expression for $\Delta^{(2)}$, namely $\phi_I D_I^{(1)} M \bar{z}^{(1)} / \big(\zeta - \frac{\gamma}{2}\big)$. Equilibrium requires market clearing at $t = 1$ so that $M\bar{z}^{(1)} = -\lambda_I\big(s_1^{(1)}\big) D_I^{(1)}$. Plugging this in, the extra term becomes $-\phi_I \lambda_I \big(s_1^{(1)}\big) \sigma_I^2 / \big(\zeta - \frac{\gamma}{2}\big)$, which makes $\Delta^{(2)}$ even more negative. The intuition is that autocorrelation in I-investor order flow implies that I-orders in $t = 2$ are on average expected to exacerbate market makers' inventories inherited from $t = 1$: I-investors are expected to buy (sell) in $t = 2$ precisely when market makers are short (long) on average, owing to I-orders from $t = 1$. This makes I-orders even less attractive relative to R-orders, further incentivizing market makers to siphon R-orders off-exchange.

The second effect above is new in the dynamic model. It arises only if I-orders are autocorrelated. Indeed, if $\phi_I = 0$, then in expectation, the average market maker's inventory $\bar{z}^{(1)}$ is no longer exacerbated by the $t = 2$ I-orders.

5.2.2 Period 1

The equilibrium analysis for period $t = 1$ proceeds similarly. We sketch the steps below and defer details to the proof of Proposition 9. First, taking the spreads $\boldsymbol{s}^{(1)}$ and siphoning $\alpha^{(1)}$ as given, a

[24] In Section 3, to simplify notation, we assumed $\mathbb{E}[D_I] = \mathbb{E}[D_R] = 0$ (*cf.* Footnote 10), so that $\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top$ reduces to Σ. Without such a simplification, it can be shown that the definition of Δ generalizes from (11) to $\Delta := \big(\mathbb{E}[D_R^2]\omega_R + \mathbb{E}[D_R D_I]\omega_I\big) - \big(\mathbb{E}[D_I^2]\omega_I + \mathbb{E}[D_R D_I]\omega_R\big)$; that is, $\sigma_k^2 = \mathrm{var}[D_k]$ is replaced by $\mathbb{E}[D_k^2]$ and $\rho\sigma_R\sigma_I = \mathrm{cov}[D_R, D_I]$ by $\mathbb{E}[D_R D_I]$. To adapt this generalized definition of Δ to the second-period phase of our dynamic model (for the case of $\bar{z}^{(1)} = 0$), we simply use the conditional expectation $\mathbb{E}_1[\cdot]$ and set $\sigma_R = 0$ to obtain

$$\Delta = \left(\underbrace{\mathbb{E}_1\left[\big(D_R^{(2)}\big)^2\right]}_{=0}\omega_R + \underbrace{\mathbb{E}_1\left[D_R^{(2)}D_I^{(2)}\right]}_{=0}\omega_I\right) - \left(\underbrace{\mathbb{E}_1\left[\big(D_I^{(2)}\big)^2\right]}_{=\sigma_I^2}\omega_I + \underbrace{\mathbb{E}_1\left[D_R^{(2)}D_I^{(2)}\right]}_{=0}\omega_R\right) = -\omega_I\sigma_I^2 < 0,$$

which exactly coincides with the result of setting $\bar{z}^{(1)} = 0$ in the expression for $\Delta^{(2)}$ given in (17).

market maker m chooses her optimal liquidity supply $\boldsymbol{x}_m^{(1)}$ to maximize

$$\pi_m^{(1)}\big(\boldsymbol{x}_m^{(1)}\big) = \boldsymbol{x}_m^{(1)\top}\boldsymbol{s}^{(1)} + \mathbb{E}\Big[\pi_m^{(2)}\big(\boldsymbol{x}_m^{(2)}; z_m^{(1)}, D_I^{(1)}\big)\Big],$$

which is the sum of her expected spread revenue from $t = 1$ and her expected continuation value $\pi_m^{(2)}(\cdot)$ as given in (16). Note that the choice variable $\boldsymbol{x}_m^{(1)}$ affects the continuation value $\mathbb{E}\big[\pi_m^{(2)}(\cdot)\big]$ for it affects the distribution of $z_m^{(1)}$, the inventory that the market maker will acquire in $t = 1$. We derive the expression of $\pi_m^{(1)}$ in the proof of Proposition 9 and show that, analogous to $\pi_m^{(2)}$, it is strictly concave in $x_{m1}^{(1)}$ but linear in $x_{m2}^{(1)}$. We then proceed exactly as for $t = 2$. The following proposition summarizes the equilibrium.

Proposition 9 (Equilibrium). An equilibrium for the full game always exists. All equilibria are characterized as follows. In period 1, all R-orders are siphoned off-exchange, i.e., $\alpha^{(1)} = 1$; the equilibrium half spreads $s_1^{(1)}$ and $s_2^{(1)}$ are defined by (40) and (38); equilibrium on-exchange liquidity supply $x_{m1}^{(1)}$ is given by (39); and equilibrium off-exchange aggregate liquidity supply $\int_0^M x_{m2}^{(2)} \mathrm{d}m$ is given by (41) (but may be allocated arbitrarily among the individual market makers). The period-2 equilibrium objects are as described in Lemma 3.

Although equilibrium (for the full game) is not unique, equilibrium does make a unique prediction regarding the spreads.[25] For period 1, these unique spreads entail $s_1^{(1)} > s_2^{(1)}$, consistent with how, according to the proposition, all R-orders are siphoned off-exchange in $t = 1$. This is because, as before, the R-orders are balanced and, thus, are relatively cheaper than the directional I-orders in terms of inventory costs. And as previously discussed, on the equilibrium path, all R-orders are also siphoned off-exchange in $t = 2$. Therefore, we conclude that the economic forces that drive siphoning are robust to the dynamic considerations we have modeled here.[26]

[25] The (full game) equilibrium is unique up to (i) how the aggregate off-exchange liquidity supply is allocated across market makers in each period, and (ii) how $\alpha^{(2)}$ is specified for period-2 subgames in which $\Delta^{(2)} = 0$.

[26] Interestingly, our model also entails a prediction on spread dynamics: It can be shown that for both marketplaces j, $s_j^{(1)} < s_j^{(2)}$. Intuitively, this is because the dynamic model allows market makers to flexibly adjust their liquidity supply over time. In fact, they have an incentive to frontload their liquidity supplies, because doing so resolves uncertainty while they still have time to react, which permits them to reduce the variance of their terminal inventory. For example,

6 Conclusion

This paper studies order flow segmentation from the novel perspective of market makers' inventory management. In isolation, a given source of orders is riskier for market makers to intermediate if it is more likely to exhibit significant directionality. Yet, market makers typically intermediate order flow from several sources, whose directionalities are potentially correlated. These considerations incentivize market makers to form portfolios of liquidity supply to different order flows, so as to optimally balance spread revenues against overall inventory costs. While the portfolio perspective on inventory management *across assets* has been previously examined in the literature (Stoll, 1978; Ho and Stoll, 1983), our portfolio perspective on inventory management *across order flows* of the same asset is novel.

In a setting tailored to PFOF, we show that siphoning specific types of orders off-exchange may be a part of portfolio-based inventory management by market makers. That is, PFOF can endogenously emerge out of inventory considerations. Our inventory perspective on PFOF makes novel predictions about consequences of regulations that ban order flow segmentation.

if a market maker received positive inventory after $t = 1$ and if she expects the $t = 2$ order flow to be buying (selling), then she can scale up (down) her supply intensity in $t = 2$. As liquidity supply is frontloaded, spreads are narrower in $t = 1$ than in $t = 2$.

Appendix

A Table of notation

Notation used in Section 2		
Exogenous parameters		
J	number of marketplaces	
$\lambda_j(\cdot)$	Poisson arrival rate of liquidity-demanding orders on marketplace j	
D_j	(random) directionality of liquidity-demanding orders on marketplace j	Endogenized in Section 3
$\boldsymbol{\mu}$	$\mathbb{E}[(D_1, D_2, \ldots, D_J)^\top]$	
Σ	$\mathrm{var}[(D_1, D_2, \ldots, D_J)^\top]$	
M	measure of market makers	
γ	parametrization of market maker inventory costs	
Endogenous variables		
s_j	half bid-ask spread of marketplace j	
x_{mj}	Poisson intensity of liquidity provision for market maker m on marketplace j	
Q_{mj}	(random) volume of market maker m on marketplace j	
Z_{mj}	(random) net inventory of market maker m on marketplace j	
π_m	expected profit of market maker m	
Additional notation used in Sections 3–4		
Exogenous parameters		
$\lambda_k(\cdot)$	Poisson arrival rate of type-k liquidity-demanding orders	
ζ	maximum acceptable half-spread	
ω_k	magnitude of type-k liquidity demand	
D_k	(random) directionality of type-k liquidity demand	
$\Sigma_\circ$	$\mathrm{var}\left[(D_I, D_R)^\top\right]$	
σ_k	$\mathrm{sd}(D_k)$	
ρ	$\mathrm{corr}(D_I, D_R)$	
Δ	$\left(\sigma_R^2\omega_R + \rho\sigma_I\sigma_R\omega_I\right) - \left(\sigma_I^2\omega_I + \rho\sigma_I\sigma_R\omega_R\right)$	
Endogenous variables		
α	fraction of R-orders routed off-exchange (i.e., to marketplace 2)	
$\boldsymbol{F}(\alpha)$	order flow weighting matrix	
$\bar{s}$	volume-weighted average half spread	
s_b	half spread when PFOF is banned	
w	total welfare	
Additional notation used in Section 5		
Exogenous parameters		
t	time period ($t \in \{1, 2\}$); superscripted on other variables as "(t)"	
ϕ_k	autocorrelation of type-k order flow	
Endogenous variables		
$\pi_m^{(t)}$	a market maker m's expected payoff, before period t trading	
$z_m^{(t)}$	a market maker m's inventory after period t trading	
$\bar{z}^{(t)}$	market makers' average inventory after period t trading	
$\Delta^{(2)}$	$\left(\phi_I D_I^{(1)} M \bar{z}^{(1)}\right)/\left(\zeta - \frac{\gamma}{2}\right) - \sigma_I^2\omega_I$	

B Proofs

Lemma 1

Proof. Consider first Z_{mj}, market maker m's net inventory from marketplace j. Conditional on the realizations of $(Q_{mj})_{j=1}^{J}$, we have

$$\begin{aligned}
\mathbb{E}\left[Z_{mj}^2 \middle| Q_{mj}\right] &= \mathbb{E}\left[\sum_{i=1}^{Q_{mj}}(-1)^{2B_{mji}} + 2\sum_{i\neq i'}(-1)^{B_{mji}}(-1)^{B_{mji'}}\right] \\
&= Q_{mj} + (Q_{mj}-1)Q_{mj}\mathbb{E}\left[\mathbb{E}\left[(-1)^{B_{mji}}(-1)^{B_{mji'}} \middle| D_j\right]\right] \\
&= Q_{mj} + (Q_{mj}-1)Q_{mj}\mathbb{E}\left[D_j^2\right] = Q_{mj} + (Q_{mj}-1)Q_{mj}\cdot(\sigma_j^2+\mu_j^2),
\end{aligned}$$

where (B_{mji}) are i.i.d. Bernoulli draws with success rate $\frac{1}{2}(1+D_j)$, μ_j is the j-th element of $\boldsymbol{\mu}$, and σ_j^2 is the j-th diagonal element of Σ; and, for $j \neq j'$,

$$\begin{aligned}
\mathbb{E}\left[Z_{mj}Z_{mj'} \middle| Q_{mj}, Q_{mj'}\right] &= \sum_{i=1}^{Q_{mj}}\sum_{i'=1}^{Q_{mj'}}\mathbb{E}\left[(-1)^{B_{mji}}(-1)^{B_{mji'}}\right] = Q_{mj}Q_{mj'}\mathbb{E}\left[\mathbb{E}\left[(-1)^{B_{mji}}(-1)^{B_{mji'}}\middle| D_j, D_{j'}\right]\right] \\
&= Q_{mj}Q_{mj'}\mathbb{E}\left[D_jD_{j'}\right] = Q_{mj}Q_{mj'}\cdot\left(\rho_{jj'}\sigma_j\sigma_{j'}+\mu_j\mu_{j'}\right),
\end{aligned}$$

where $\rho_{jj'}$ is the correlation between D_j and $D_{j'}$. The market maker's total net inventory is $Z_m := \sum_{j=1}^{J} Z_{mj}$ and,

$$\begin{aligned}
\mathbb{E}\left[Z_m^2 \middle| Q_{m1}, \ldots, Q_{mJ}\right] &= \sum_{j=1}^{J}\mathbb{E}\left[Z_{mj}^2 \middle| Q_{mj}\right] + \sum_{j\neq j'}\mathbb{E}\left[Z_{mj}Z_{ij'} \middle| Q_{mj}, Q_{mj'}\right] \\
&= \sum_{j=1}^{J}\Big(Q_{mj} + (Q_{mj}-1)Q_{mj}(\sigma_j^2+\mu_j^2)\Big) + \sum_{j\neq j'}Q_{mj}Q_{mj'}\left(\rho_{jj'}\sigma_j\sigma_{j'}+\mu_j\mu_{j'}\right) \\
&= \sum_{j=1}^{J}Q_{mj} + \sum_{j=1}^{J}(Q_{mj}^2 - Q_{mj})(\sigma_j^2+\mu_j^2) + \sum_{j\neq j'}Q_{mj}Q_{mj'}\left(\rho_{jj'}\sigma_j\sigma_{j'}+\mu_j\mu_{j'}\right)
\end{aligned}$$

Finally, take the unconditional expectation to get

$$\begin{aligned}\mathbb{E}\big[Z_m^2\big] &= \sum_{j=1}^{J} x_{mj} + \sum_{j=1}^{J}(x_{mj}^2 + x_{mj} - x_{mj})(\sigma_j^2 + \mu_j^2) + \sum_{j\neq j'} x_{mj}x_{mj'}\big(\rho_{jj'}\sigma_j\sigma_{j'} + \mu_j\mu_{j'}\big) \\ &= \underbrace{\sum_{j=1}^{J} x_{mj}}_{=\boldsymbol{x}_m^\top \mathbf{1}} + \underbrace{\sum_{j=1}^{J} x_{mj}^2\Big(\sigma_j^2 + \mu_j^2\Big) + \sum_{j\neq j'} x_{mj}x_{mj'}\big(\rho_{jj'}\sigma_j\sigma_{j'} + \mu_j\mu_{j'}\big)}_{=\boldsymbol{x}_m^\top(\Sigma+\boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{x}_m}. \qquad \square\end{aligned}$$

Proposition 1

Proof. Given the optimal supply (5), the market-clearing conditions (2) become

$$\frac{M}{\gamma}\big(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\big)^{-1}\Big(\boldsymbol{s} - \frac{\gamma}{2}\mathbf{1}\Big) = \big(\lambda_1(s_1), \ldots, \lambda_J(s_J)\big)^\top. \tag{18}$$

Below we proceed to show the existence and the uniqueness of a solution $\boldsymbol{s}$ to the above.

Existence: Notice that a solution to (18) is equivalent to a fixed point of the function

$$\boldsymbol{G}(\boldsymbol{s}) = \frac{\gamma}{M}\big(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\big)\big(\lambda_1(s_1), \ldots, \lambda_J(s_J)\big)^\top + \frac{\gamma}{2}\mathbf{1}.$$

Because each $\lambda_j(\cdot)$ is nonincreasing and nonnegative, the range of $\big(\lambda_1(s_1), \ldots, \lambda_J(s_J)\big)^\top$ is the compact, convex set $\prod_{j=1}^{J}[0, \lambda_j(0)]$. The range of $\boldsymbol{G}(\boldsymbol{s})$ is a linear transformation of that set, so is also compact and convex. Moreover, $\boldsymbol{G}$ is continuous. It therefore follows from Brouwer's fixed-point theorem that $\boldsymbol{G}$ has a fixed point, and hence that (18) has a solution.

Uniqueness: Suppose there are two solutions to (18), denoted by $\boldsymbol{s}$ and $\boldsymbol{s}' = \boldsymbol{s} + \boldsymbol{\delta}$, where the vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_J)^\top$ is the difference of the two solutions. For notational simplicity, write $\boldsymbol{\lambda}$ as the right-hand side of (18) under $\boldsymbol{s}$ and $\boldsymbol{\lambda}'$ for that under $\boldsymbol{s}'$. Difference the two market-clearing conditions and then left-multiply both sides with $\boldsymbol{\delta}^\top$ to get $\boldsymbol{\delta}^\top\Big(\frac{M}{\gamma}\big(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\big)^{-1}\Big)\boldsymbol{\delta} = \boldsymbol{\delta}^\top(\boldsymbol{\lambda}' - \boldsymbol{\lambda})$, where if $\boldsymbol{\delta} \neq \mathbf{0}$, the left-hand side is positive, because $\big(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top\big)^{-1}$ is positive-definite. Suppose $\delta_j > 0$ (< 0). Then, since the demand functions are monotonically weakly decreasing, $\lambda_j(s_j + \delta_j) - \lambda_j(s_j) \leq 0$ (≥ 0), and the right-hand side above is weakly negative. Therefore, the two solutions $\boldsymbol{s}$ and $\boldsymbol{s}'$ must collapse with $\boldsymbol{\delta} = \mathbf{0}$. □

Proposition 2

Proof. We consider the three cases of $\alpha = 1$, $\alpha = 0$, and $\alpha \in (0, 1)$ separately, following conditions (7a)–(7b).

Consider first the case of $\alpha = 1$. Then Proposition 1 applies with $\lambda_1(s_1) = \lambda_I(s_1)$, $\lambda_2(s_2) = \lambda_R(s_2)$, and weighting matrix $\boldsymbol{F}(1)$. To verify that this outcome satisfies the notion of equilibrium defined in Section 3.1, it remains only to check (7b), which requires $s_1 \geq s_2$. Recall that the demand $\lambda_k(s)$ exhibits a kink at $s = \zeta$. We then have three subcases depending on the ranking of s_1, s_2, and ζ.

- If $\zeta > s_1 \geq s_2$, then $\lambda_1(s_1) = (\zeta - s_1)\omega_I$ and $\lambda_2(s_2) = (\zeta - s_2)\omega_R$. Jointly solving the two market-clearing conditions $Mx_{mj} = \lambda_j(s_j)$ for s_1 and s_2, we obtain $s_j = \frac{\gamma}{2} + \left(\zeta - \frac{\gamma}{2}\right)\beta_j$, where

$$\beta_1 = \frac{(\sigma_I^2\omega_I + \rho\sigma_I\sigma_R\omega_R)\gamma M + (1-\rho^2)\sigma_I^2\sigma_R^2\gamma^2\omega_I\omega_R}{M^2 + (\sigma_I^2\omega_I + \sigma_R^2\omega_R)\gamma M + (1-\rho^2)\sigma_I^2\sigma_R^2\gamma^2\omega_I\omega_R}; \text{ and} \tag{19}$$

$$\beta_2 = \frac{(\sigma_R^2\omega_R + \rho\sigma_I\sigma_R\omega_I)\gamma M + (1-\rho^2)\sigma_I^2\sigma_R^2\gamma^2\omega_I\omega_R}{M^2 + (\sigma_I^2\omega_I + \sigma_R^2\omega_R)\gamma M + (1-\rho^2)\sigma_I^2\sigma_R^2\gamma^2\omega_I\omega_R}. \tag{20}$$

  It can be seen that $\zeta > s_1$ (equivalently, $1 > \beta_1$) is satisfied if and only if (12) holds. In addition, $s_1 \geq s_2$ (equivalently, $\beta_1 \geq \beta_2$) is satisfied if and only if $\Delta \leq 0$.

- If $s_1 \geq \zeta > s_2$, then $\lambda_1(s_1) = 0$ and $\lambda_2(s_2) = (\zeta - s_2)\omega_R$. The market-clearing conditions then yield $s_j = \frac{\gamma}{2} + \left(\zeta - \frac{\gamma}{2}\right)\beta_j$, where

$$\beta_1 = \frac{\rho\sigma_I\sigma_R\omega_R\gamma}{M + \sigma_R^2\omega_R\gamma} \text{ and } \beta_2 = \frac{\sigma_R^2\omega_R\gamma}{M + \sigma_R^2\omega_R\gamma}.$$

  This solution is consistent with $s_1 \geq \zeta > s_2$ (equivalently, $\beta_1 \geq 1 > \beta_2$) if and only if (12) fails. We also note that the failure of (12) necessarily implies $\Delta < 0$. To see this, suppose the opposite is true, i.e., (12) fails and that $\left(\sigma_R^2\omega_R + \rho\sigma_I\sigma_R\omega_I\right) - \left(\sigma_I^2\omega_I + \rho\sigma_I\sigma_R\omega_R\right) \geq 0$, the latter implying

$$\left(-\sigma_I^2 + \rho\sigma_I\sigma_R\right)\omega_I \geq (\rho\sigma_I - \sigma_R)\sigma_R\omega_R \implies \rho \geq \frac{(\rho\sigma_I - \sigma_R)\omega_R}{\sigma_I\omega_I} + \frac{\sigma_I}{\sigma_R}.$$

  Note that for (12) to fail, we must have $\rho\sigma_I > \sigma_R$, for otherwise the right-hand side of (12) is weakly negative, which would mean that (12) must hold, because $M > 0$. It then follows that $\rho\sigma_I - \sigma_R > 0$ and $\sigma_R < \rho\sigma_I < \sigma_I \implies \frac{\sigma_I}{\sigma_R} > 1$. Therefore, we obtain from the above inequality that $\rho > 1$, a contradiction.

- Finally, if $s_1 \geq s_2 \geq \zeta$, then $\lambda_1(s_1) = \lambda_2(s_2) = 0$. But then using the optimal demand (5), market clearing requires $s_1 = s_2 = \frac{\gamma}{2} < \zeta$. Hence, this cannot be an equilibrium.

Hence, there is an equilibrium with $\alpha = 1$ if and only if $\Delta \leq 0$. Furthermore, this equilibrium is such that $s_1 < \zeta$ if and only if (12) also holds.

Next, suppose $\alpha = 0$. Then Proposition 1 applies with $\lambda_1(s_1) = \lambda_I(s_1) + \lambda_R(s_1)$, $\lambda_2(s_2) = 0$, and weighting matrix $\boldsymbol{F}(0)$. To verify that this outcome satisfies the notion of equilibrium defined in Section 3.1, it remains only to check (7a), which requires $s_1 \leq s_2$. Because there is no demand in $j = 2$ in this case, we only need to discuss two subcases.

- If $s_1 < \zeta$, then $\lambda_1(s_1) = (\zeta - s_1)(\omega_I + \omega_R)$ and $\lambda_2(s_2) = 0$. By market clearing, we obtain $s_j = \frac{\gamma}{2} + \big(\zeta - \frac{\gamma}{2}\big)\beta_j$, where

$$\begin{aligned}\beta_1 &= \frac{(\sigma_I^2\omega_I^2 + 2\rho\sigma_I\sigma_R\omega_I\omega_R + \sigma_R^2\omega_R^2)\gamma}{M(\omega_I + \omega_R) + (\sigma_I^2\omega_I^2 + 2\rho\sigma_I\sigma_R\omega_I\omega_R + \sigma_R^2\omega_R^2)\gamma}; \text{ and} \\ \beta_2 &= \frac{(\rho\sigma_I\omega_I\sigma_R + \sigma_R\omega_R^2)(\omega_I + \omega_R)\gamma}{M(\omega_I + \omega_R) + (\sigma_I^2\omega_I^2 + 2\rho\sigma_I\sigma_R\omega_I\omega_R + \sigma_R^2\omega_R^2)\gamma}.\end{aligned} \tag{21}$$

  It is clear that $\beta_1 < 1$, hence also $s_1 < \zeta$, is guaranteed. Also, $s_1 \leq s_2$ (equivalently, $\beta_1 \leq \beta_2$) is satisfied if and only if $\Delta \geq 0$.

- If $s_1 \geq \zeta$, then $\lambda_1(s_1) = \lambda_2(s_2) = 0$. But then market clearing requires $s_1 = s_2 = \frac{\gamma}{2} < \zeta$. Hence, this cannot be an equilibrium.

Hence, there is an equilibrium with $\alpha = 0$ if and only if $\Delta \geq 0$.

Finally, suppose $\boldsymbol{\alpha \in \big(0, 1\big)}$. Then Proposition 1 applies with $\lambda_1(s_1) = \lambda_I(s_1) + (1 - \alpha)\lambda_R(s_1)$, $\lambda_2(s_2) = \alpha\lambda_R(s_2)$, and weighting matrix $\boldsymbol{F}(\alpha)$. To verify that this outcome satisfies the notion of equilibrium defined in Section 3.1, it remains only to check (7a) and (7b), which jointly require $s_1 = s_2 = s$. There are then two subcases.

- If $s < \zeta$, then $\lambda_1(s) = (\zeta - s)\big(\omega_I + (1-\alpha)\omega_R\big)$ and $\lambda_2(s) = (\zeta - s)\alpha\omega_R$. The two remaining unknowns s and α are pinned down by the market-clearing conditions, which yield $s = \frac{\gamma}{2} + (\zeta - \frac{\gamma}{2})\beta$ with

$$\beta = \frac{(1 - \rho^2)\sigma_I^2\sigma_R^2(\omega_I + \omega_R)\gamma}{M(\sigma_I^2 - 2\rho\sigma_I\sigma_R + \sigma_R^2) + (1 - \rho^2)\sigma_I^2\sigma_R^2(\omega_I + \omega_R)\gamma};$$

  and for α:

$$(\omega_I + \omega_R - \alpha\omega_R)\Big[\underbrace{\big(\sigma_R^2\omega_R + \rho\sigma_I\sigma_R\omega_I\big) - \big(\sigma_I^2\omega_I + \rho\sigma_I\sigma_R\omega_R\big)}_{=\Delta}\Big] = 0,$$

  which, for any $\alpha \in (0, 1)$, holds if and only if $\Delta = 0$. Note that $\beta < 1$ always holds, guaranteeing $s < \zeta$.

- If $s \geq \zeta$, then $\lambda_1(s) = \lambda_2(s) = 0$. But then market clearing requires $s = \frac{\gamma}{2} < \zeta$. Hence, this cannot be an equilibrium.

Hence, for any $\alpha \in (0, 1)$, there is a corresponding equilibrium if and only if $\Delta = 0$. □

Corollary 1

Proof. A PFOF ban exogenously forces $\alpha = 0$. Then Proposition 1 applies, with $\lambda_1(s_b) = \lambda_I(s_b) + \lambda_R(s_b)$, $\lambda_2(s) = 0$, and weighting matrix $\boldsymbol{F}(0)$. Although this completes the proof, it is useful for subsequent proofs to derive an expression for s_b, the spread prevailing in this equilibrium. We conjecture (and subsequently verify) that $s_b < \zeta$. Under this conjecture, liquidity demand is $\lambda_1(s_b) = (\zeta - s_b)(\omega_I + \omega_R)$, so that market-clearing implies $s_b = \frac{\gamma}{2} + \left(\zeta - \frac{\gamma}{2}\right)\beta_b$, where β_b has the same expression as (21). Clearly, $\beta_b < 1$ and this guarantees $s_b < \zeta$. □

Proposition 3

Proof. Equilibrium involves positive volume on-exchange, i.e., $\int_0^M x_{m1} \mathrm{d}m > 0$, if and only if $s_1 < \zeta$. And equilibrium involves positive volume off-exchange, i.e., $\int_0^M x_{m2} \mathrm{d}m > 0$, if and only if both $\alpha > 0$ and $s_2 < \zeta$. According to Proposition 2, there is a unique equilibrium involving $\alpha > 0$ if and only if $\Delta < 0$. The proof of Proposition 2 establishes that when $\Delta < 0$, the equilibrium is guaranteed to feature $s_2 < \zeta$. Finally, the proof of Proposition 2 additionally establishes that this equilibrium also features $s_1 < \zeta$ if and only if (12) also holds. □

Lemma 2

Proof. Consider the average demand curve. With PFOF, $\bar{s}$ as defined in (13) can be rewritten as

$$\bar{s} = \zeta - \frac{\lambda_I(s_1)^2/\omega_I + \lambda_R(s_2)^2/\omega_R}{\lambda_I(s_1) + \lambda_R(s_2)} = \zeta - \frac{(Mx_{m1})^2/\omega_I + (Mx_{m2})^2/\omega_R}{Mx_{m1} + Mx_{m2}} = \zeta - \left(\frac{(1-f_R)^2}{\omega_I} + \frac{f_R^2}{\omega_R}\right)x,$$

where the first equality follows from the individual liquidity demand curves of the two investor types: $\lambda_I(s_1) = (\zeta - s_1)\omega_I$ and $\lambda_R(s_2) = (\zeta - s_2)\omega_R$; the second equality follows from market clearing; and the third equality uses $f_R := \frac{x_{m2}}{x_{m2}+x_{m1}}$ and $x := (x_{m1} + x_{m2})M$. The expression also applies when PFOF is banned, in which case $f_R = \frac{\omega_R}{\omega_I+\omega_R}$, implying $s_b = \bar{s} = \zeta - x/(\omega_I + \omega_R)$.

Consider next the average supply curve. With PFOF, market makers' equilibrium supply quantities satisfy the first-order conditions to (8):

$$s_1 - \frac{\gamma}{2} - \gamma \cdot \left(\sigma_I^2 x_{m1} + \rho\sigma_I\sigma_R x_{m2}\right) = 0; \text{ and } s_2 - \frac{\gamma}{2} - \gamma \cdot \left(\sigma_R^2 x_{m2} + \rho\sigma_I\sigma_R x_{m1}\right) = 0.$$

Multiply the first with x_{m1} and the second with x_{m2}, add them up, and finally divide the sum by $x_{m1} + x_{m2}$ to get

$$\begin{aligned}\bar{s} = \frac{x_{m1}s_1 + x_{m2}s_2}{x_{m1} + x_{m2}} &= \frac{\gamma}{2} + \frac{\sigma_I^2 x_{m1}^2 + 2\rho\sigma_I\sigma_R x_{m1}x_{m2} + \sigma_R^2 x_{m2}^2}{x_{m1} + x_{m2}}\gamma = \frac{\gamma}{2} + \frac{\gamma \mathrm{var}[x_{m1}D_I + x_{m2}D_R]}{x_{m1} + x_{m2}} \\ &= \frac{\gamma}{2} + \left(\frac{\gamma}{M}\mathrm{var}[(1-f_R)D_I + f_R D_R]\right)x,\end{aligned}$$

where the last equality uses $f_R := \frac{x_{m2}}{x_{m2}+x_{m1}}$ and $x := (x_{m1} + x_{m2})M$. When PFOF is banned, the above expression also applies, with $f_R = \frac{\omega_R}{\omega_I+\omega_R}$. □

Proposition 4

Proof. Under $\Delta < 0$ and under (12), the equilibrium features both on-exchange and off-exchange volume. Hence, following the proofs of Proposition 2 and Corollary 1, $s_j = \frac{\gamma}{2} + \left(\zeta - \frac{\gamma}{2}\right)\beta_j$ for $j \in \{1, 2, b\}$, where β_1, β_2, and β_b are given by (19), (20), and (21), respectively. Compare first s_b and s_2. Direct calculation shows that $\text{sign}[s_2 - s_b] = \text{sign}[\beta_2 - \beta_b] = \text{sign}\left[(\sigma_R^2\omega_R + \rho\sigma_I\sigma_R\omega_I) - (\sigma_I^2\omega_I + \rho\sigma_I\sigma_R\omega_R)\right]\text{sign}[M + \gamma\sigma_I^2\omega_I + \gamma\rho\sigma_I\sigma_R\omega_R]$. The first factor is exactly $\text{sign}[\Delta]$, which is negative, as assumed. For the second factor, note that $M + \gamma\sigma_I^2\omega_I + \gamma\rho\sigma_I\sigma_R\omega_R$ is increasing in ρ. We therefore examine this factor at the minimum value of ρ that is jointly permitted by the assumptions $\Delta < 0$ and (12). To begin, (12) implies no lower bound for ρ (only the upper bound $\rho < \frac{\sigma_R}{\sigma_I} + \frac{M}{\gamma\sigma_I\sigma_R\omega_R}$), so it suffices to consider only the implications of $\Delta < 0$. On the one hand, suppose $\Delta < 0$ implies no lower bound for ρ, meaning that $0 \geq \lim_{\rho\to-1}\Delta(\rho) = (\sigma_I + \sigma_R)(\sigma_I\omega_I - \sigma_R\omega_R)$, and hence $\sigma_I\omega_I - \sigma_R\omega_R \geq 0$. Then $M + \gamma\sigma_I^2\omega_I + \gamma\rho\sigma_I\sigma_R\omega_R > M + \gamma\sigma_I(\sigma_I\omega_I - \sigma_R\omega_R) > 0$, so that the second factor is positive. On the other hand, suppose $\Delta < 0$ does imply a lower bound for ρ: $\rho \geq \frac{\sigma_I^2\omega_I - \sigma_R^2\omega_R}{(\omega_I - \omega_R)\sigma_I\sigma_R} \geq -1$. Note that this can be the case only if both $\omega_I < \omega_R$ and $\sigma_R\omega_R > \sigma_I\omega_I$. At this constrained lower bound, $M + \gamma\sigma_I^2\omega_I + \gamma\rho\sigma_I\sigma_R\omega_R = M + \gamma(\sigma_I\omega_I + \sigma_R\omega_R)(\sigma_I\omega_I - \sigma_R\omega_R)/(\omega_I - \omega_R) > 0$, so that the second factor is positive. In either case, we conclude $s_2 < s_b$.

Next, compare s_b and s_1. Direct calculation shows that $\text{sign}[s_1 - s_b] = \text{sign}[\beta_1 - \beta_b] = \text{sign}[-\Delta]\text{sign}[M + \gamma\sigma_R^2\omega_R + \gamma\rho\sigma_I\sigma_R\omega_I]$. Since $\Delta < 0$, it remains to sign the second part. It is negative if and only if $\rho < -(M + \gamma\sigma_R^2\omega_R)/(\sigma_R\sigma_I\omega_I)$.

Finally, we compare s_b and $\bar{s}$. To do so, we first define $f_R \in [0, 1]$ as the fraction of R-orders in a market maker's portfolio of orders. When PFOF is banned, both investor types pay the same spread, so we have $f_R^{\text{ban}} = \frac{\omega_R}{\omega_R + \omega_I}$. Let $f_R^{\text{no-ban}}$ denote the equilibrium value for f_R when PFOF is allowed. Given that $\Delta < 0$, market makers want to siphon R-orders off-exchange, so that $f_R^{\text{no-ban}} > \frac{\omega_R}{\omega_R + \omega_I}$. Specifically, $f_R^{\text{no-ban}} = \frac{(\zeta - s_2)\omega_R}{(\zeta - s_2)\omega_R + (\zeta - s_1)\omega_I}$, which, following (19) and (20), becomes

$$f_R^{\text{no-ban}} = \frac{\omega_R + (\sigma_I^2 - \rho\sigma_I\sigma_R)\omega_I\omega_R\gamma/M}{(\omega_I + \omega_R) + (\sigma_I^2 - 2\rho\sigma_I\sigma_R + \sigma_R^2)\omega_I\omega_R\gamma/M}.$$

Let $\bar{s}(f_R)$ denote the volume-weighted average spread as a function of f_R, defined as the intersection of the two curves from Lemma 2: the average liquidity demand curve $\bar{s}(f_R) = \zeta - v(f_R)x$ and the average liquidity supply curve $\bar{s}(f_R) = \frac{\gamma}{2} + c(f_R)x$. Thus, $\bar{s} < s_b$ will follow if we show that $\bar{s}(f_R^{\text{ban}}) > \bar{s}(f_R^{\text{no-ban}})$. To do so, we solve for $\bar{s}(f_R)$, by eliminating the aggregate volume x:

$$\bar{s}(f_R) = \frac{\frac{\gamma}{2}v(f_R) + \zeta c(f_R)}{v(f_R) + c(f_R)} = \frac{\gamma}{2} + \left(\zeta - \frac{\gamma}{2}\right)\frac{c(f_R)}{v(f_R) + c(f_R)} = \frac{\gamma}{2} + \left(\zeta - \frac{\gamma}{2}\right)\frac{1}{1 + v(f_R)/c(f_R)}.$$

As observed in the text (in the two bullet points following Lemma 2), we have both $v(f_R^{\text{ban}}) <$

$v(f_R^{\text{no-ban}})$ and $c(f_R^{\text{ban}}) > c(f_R^{\text{no-ban}})$, which together imply $\bar{s}(f_R^{\text{ban}}) > \bar{s}(f_R^{\text{no-ban}})$, as desired. □

Proposition 5

Proof. Under $\Delta < 0$ and under (12), the equilibrium features both on-exchange and off-exchange volume. Hence, following the proof of Proposition 2, $s_j = \frac{\gamma}{2} + \left(\zeta - \frac{\gamma}{2}\right)\beta_j$ for $j \in \{1, 2\}$, where β_1 and β_2 are given by (19) and (20), respectively. Hence, $\text{sign}\left[\frac{\mathrm{d}s_2}{\mathrm{d}\omega_R}\right] = \text{sign}\left[\frac{\mathrm{d}\beta_2}{\mathrm{d}\omega_R}\right] = \text{sign}[M + (\sigma_I^2 - \rho\sigma_I\sigma_R)\gamma\omega_I]$. Using (12), we have $\text{sign}[M + (\sigma_I^2 - \rho\sigma_I\sigma_R)\gamma\omega_I] \geq \text{sign}[\sigma_I^2\omega_I - (\omega_I - \omega_R)\sigma_I\sigma_R\rho - \sigma_R^2\omega_R] = \text{sign}[-\Delta] > 0$. Therefore, $\frac{\mathrm{d}s_2}{\mathrm{d}\omega_R} > 0$.

Likewise, $\text{sign}\left[\frac{\mathrm{d}s_1}{\mathrm{d}\omega_R}\right] = \text{sign}\left[\frac{\mathrm{d}\beta_1}{\mathrm{d}\omega_R}\right] = \text{sign}[\rho]\text{sign}[M + (\sigma_I^2 - \rho\sigma_I\sigma_R)\gamma\omega_R]$. The second factor was shown to be positive above. Hence, $\text{sign}\left[\frac{\mathrm{d}s_1}{\mathrm{d}\omega_R}\right] = \text{sign}[\rho]$.

Finally, $\text{sign}\left[\frac{\mathrm{d}(s_1 - s_2)}{\mathrm{d}\omega_R}\right] = \text{sign}\left[\frac{\mathrm{d}(\beta_1 - \beta_2)}{\mathrm{d}\omega_R}\right] = \text{sign}\left[\gamma\sigma_I^2\sigma_R\omega_I\rho^2 + M\sigma_I\rho - (M + \gamma\sigma_I^2\omega_I)\sigma_R\right]$. This quadratic expression in $\rho \in [-1, 1]$ is convex, is strictly negative at $\rho = 0$, and is strictly increasing at $\rho = 0$. Note also that at $\rho = -1$, it becomes $-(\sigma_I + \sigma_R)M < 0$, implying that $\frac{\mathrm{d}(s_1 - s_2)}{\mathrm{d}\omega_R} < 0$ for all $\rho \in [-1, 0]$. Therefore, there exists a unique threshold $\hat{\rho} > 0$,

$$\hat{\rho} = \frac{1}{2\gamma\sigma_I\sigma_R\omega_I}\left(-M + \sqrt{M^2 + 4M\gamma\sigma_R^2\omega_I + 4\gamma^2\sigma_I^2\sigma_R^2\omega_I^2}\right), \tag{22}$$

which is the positive root of the above quadratic expression, such that $\frac{\mathrm{d}(s_1 - s_2)}{\mathrm{d}\omega_R} > 0$ for $\rho > \hat{\rho}$. (Note that the threshold $\hat{\rho}$ may or may not lie within the domain of ρ, i.e., $\hat{\rho}$ can be $\lesseqgtr 1$.) Summing up, the necessary and sufficient condition for $\frac{\mathrm{d}(s_1 - s_2)}{\mathrm{d}\omega_R} < 0$ is $\rho < \hat{\rho}$. □

Proposition 6

Proof. Under $\Delta < 0$ and under (12), the equilibrium features both on-exchange and off-exchange volume. Hence, following the proof of Proposition 2, $s_j = \frac{\gamma}{2} + \left(\zeta - \frac{\gamma}{2}\right)\beta_j$ for $j \in \{1, 2\}$, where β_1 and β_2 are given by (19) and (20), respectively. Hence, $\text{sign}\left[\frac{\mathrm{d}s_1}{\mathrm{d}M}\right] = \text{sign}\left[\frac{\mathrm{d}\beta_1}{\mathrm{d}M}\right] = \text{sign}\Big[(1 - \rho^2)(\rho\sigma_I - \sigma_R)\sigma_I\sigma_R^3\omega_I\omega_R^2\gamma^2 - 2(1 - \rho^2)M\gamma\sigma_I\sigma_R^2\omega_I\omega_R - (\sigma_I\omega_I + \rho\sigma_R\omega_R)M^2\Big]$. That is, we need to evaluate the sign of the above quadratic expression in M. There are two cases, depending on whether M is constrained by (12).

- Suppose (12) is not binding, i.e., $\rho\sigma_I\sigma_R - \sigma_R^2 \leq 0$. First, we note that this implies that the intercept of the quadratic expression is negative. Second, we show that the quadratic expression in M must be (weakly) concave. To do so, we assume the opposite, i.e., the coefficient on M^2 is positive, i.e., $\sigma_I\omega_I + \rho\sigma_R\omega_R < 0$. Next, $\Delta < 0$ implies, after some rearranging, that $(\sigma_R\omega_R + \rho\sigma_I\omega_I)\sigma_R <$

$(\sigma_I\omega_I + \rho\sigma_R\omega_R)\sigma_I$. Hence, $\sigma_R\omega_R + \rho\sigma_I\omega_I < 0$. Summing $\sigma_I\omega_I + \rho\sigma_R\omega_R < 0$ and $\sigma_R\omega_R + \rho\sigma_I\omega_I < 0$ implies that $(1+\rho)(\sigma_I\omega_I + \sigma_R\omega_R) < 0$, which is a contradiction because $\rho > -1$. Third, in the limit of $M \to 0$, the slope of the quadratic expression is negative.

- Suppose (12) is binding, i.e., $\rho\sigma_I\sigma_R - \sigma_R^2 > 0$. First, we note that this implies $\rho > 0$ and, hence, the coefficient on M^2 in the above quadratic expression is strictly negative. Second, just as in the previous case, the slope of the quadratic expression is negative at $M = 0$. Note, however, that the relevant domain for M is now bounded away from 0; rather, (12) implies $M > (\rho\sigma_I\sigma_R - \sigma_R^2)\omega_R\gamma$. Third, in the limit as $M \to (\rho\sigma_I\sigma_R - \sigma_R^2)\omega_R\gamma$, the quadratic expression evaluates to $\gamma^3\rho\sigma_I\sigma_R(\rho\sigma_I\sigma_R - \sigma_R^2)\omega_R^2\Delta < 0$.

Summing up, in either case, the quadratic expression above is negative for all M on the relevant domain. Therefore, β_1 (hence also s_1) is always decreasing in M.

Likewise, $\text{sign}\left[\frac{\mathrm{d}s_2}{\mathrm{d}M}\right] = \text{sign}\left[\frac{\mathrm{d}\beta_2}{\mathrm{d}M}\right]$. We first show that β_2 is quasi-convex in M. Direct evaluation shows

$$\frac{\mathrm{d}\beta_2}{\mathrm{d}M} = \frac{\gamma\sigma_R}{h(M)}\Big(-(1-\rho^2)(\sigma_I - \rho\sigma_R)\sigma_I^3\sigma_R\omega_I^2\omega_R\gamma^2 - 2(1-\rho^2)\gamma\sigma_I^2\sigma_R\omega_I\omega_R M - (\rho\sigma_I\omega_I + \sigma_R\omega_R)M^2\Big),$$

where $h(M)$ is some strictly positive 4th-order polynomial in M, not affecting the sign of $\frac{\mathrm{d}\beta_2}{\mathrm{d}M}$. Consider a stationary point denoted by M^*. We have

$$\begin{aligned}\frac{\mathrm{d}^2\beta_2}{\mathrm{d}M^2}\bigg|_{M=M^*} &= \frac{2\gamma\sigma_R}{h(M^*)}\Big(-(1-\rho^2)\gamma\sigma_I^2\sigma_R\omega_I\omega_R - (\rho\sigma_I\omega_I + \sigma_R\omega_R)M^*\Big) \\ &= \frac{2\gamma^2(1-\rho^2)}{h(M^*)M^*}\sigma_I^2\sigma_R^2\omega_I\omega_R\Big(M^* + \gamma(\sigma_I^2 - \rho\sigma_I\sigma_R)\omega_I\Big) > \frac{2\gamma^2(1-\rho^2)}{h(M^*)M^*}\sigma_I^2\sigma_R^2\omega_I\omega_R \cdot (-\gamma\Delta) > 0,\end{aligned}$$

where the second line follows from $\frac{\mathrm{d}\beta_2}{\mathrm{d}M}\Big|_{M=M^*} = 0$ by substituting $(\rho\sigma_I\omega_I + \sigma_R\omega_R)M^* = -\frac{1}{M^*}(1 - \rho^2)(\sigma_I - \rho\sigma_R)\sigma_I^3\sigma_R\omega_I^2\omega_R\gamma^2 - 2(1-\rho^2)\gamma\sigma_I^2\sigma_R\omega_I\omega_R$, the first inequality follows from (12), and the second from $\Delta < 0$. That is, at all stationary points (if any exist), β_2 is strictly convex, and, hence, β_2 is quasi-convex on the domain $M > 0$. We then examine the limits of β_2. Direct computation shows that if M is not constrained by (12), $\lim_{M\downarrow 0}\beta_2 = 1$; that if M is constrained by (12), $\lim_{M\downarrow(\rho\sigma_I\sigma_R - \sigma_R^2)\omega_R\gamma}\beta_2 = \sigma_R/(\rho\sigma_I)$ (note that $\rho > 0$ in this case); and that $\lim_{M\to\infty}\beta_2 = 0$. That is, the left limit of β_2, irrespective of whether M is constrained, is always strictly larger than its right limit. Together with the quasi-convexity, it follows that β_2 (hence also s_2) is either monotonically decreasing or U-shaped in M, depending on $\lim_{M\to\infty}\text{sign}\left[\frac{\mathrm{d}\beta_2}{\mathrm{d}M}\right]$. Since $\text{sign}\left[\frac{\mathrm{d}\beta_2}{\mathrm{d}M}\right] = \text{sign}\left[M^2\frac{\mathrm{d}\beta_2}{\mathrm{d}M}\right]$, we directly compute $\lim_{M\to\infty}\text{sign}\left[M^2\frac{\mathrm{d}\beta_2}{\mathrm{d}M}\right] = \text{sign}[-(\rho\sigma_I\omega_I + \sigma_R\omega_R)]$, recalling that $h(M)$ is a 4th-order polynomial in M. Therefore, if $\rho < -\frac{\sigma_R\omega_R}{\sigma_I\omega_I}$, then the off-exchange spread s_2 is U-shaped in M; or else, s_2 is also monotonically decreasing in M. □

Proposition 7

Proof. We first derive a market maker's equilibrium expected payoff π without the ban and π_b with the ban. To do so, we calculate market makers' aggregate surplus as the area of the triangle, in a supply-demand graph, formed by the vertical (price) axis, the aggregate supply in marketplace j, and the horizontal line at s_j. The per capita surplus, therefore, amounts to $\frac{1}{2M}\big(s_j - \frac{\gamma}{2}\big)(\zeta - s_j)\omega_j$, where $\omega_j = \omega_I$ if $j = 1$, $\omega_j = \omega_R$ if $j = 2$, and $\omega_j = (\omega_I + \omega_R)$ if $j = b$. We then further plug in $s_j = \frac{\gamma}{2} + (\zeta - \frac{\gamma}{2})\beta_j$ for $j \in \{1, 2, b\}$, where the equilibrium values for β_1, β_2, and β_b are given by (19), (20), and (21), respectively. This gives $\pi = \frac{(\zeta-\gamma/2)^2}{2M}\big((1-\beta_1)\beta_1\omega_I + (1-\beta_2)\beta_2\omega_R\big)$; and $\pi_b = \frac{(\zeta-\gamma/2)^2}{2M}(\omega_I + \omega_R)(1-\beta_b)\beta_b$. Directly evaluating $\pi - \pi_b$ yields that sign$[\pi - \pi_b]$ is equivalent to the sign of the following cubic polynomial in M:

$$\begin{aligned}&\underbrace{-2(\omega_I + \omega_R)}_{<0} M^3 + \underbrace{\big(2\rho\sigma_I\sigma_R\omega_I\omega_R + \sigma_I^2\omega_I(2\omega_I + \omega_R) + \sigma_R^2\omega_R(\omega_I + 2\omega_R)\big)\gamma}_{>0} M^2 \\ &\quad + \underbrace{\gamma^3\Big(1-\rho^2\Big)\sigma_I^2\sigma_R^2\omega_I\omega_R\Big(2\rho\sigma_I\sigma_R\omega_I\omega_R + \sigma_I^2\omega_I^2 + \sigma_R^2\omega_R^2\Big)}_{>0}.\end{aligned} \tag{23}$$

Clearly, its derivative is a concave quadratic function of M, with one strictly negative root and the other root at zero. That is, for all $M > 0$, the cubic expression is strictly decreasing in M. Since the intercept is positive, the cubic polynomial always has a unique positive root $\hat{M} > 0$. □

Proposition 8

Proof. Under $\Delta < 0$ and under (12), the equilibrium without a PFOF ban involves all R-orders being siphoned off-exchange, as well as positive volume both on-exchange and off-exchange. Hence, (x_I, x_R, s_I, s_R) are characterized by the following two conditions, which respectively capture market maker optimization (following (5)) and market-clearing:

$$\begin{pmatrix} x_I \\ x_R \end{pmatrix} = \frac{1}{\gamma}\Sigma_0^{-1}\begin{pmatrix} s_I - \frac{\gamma}{2} \\ s_R - \frac{\gamma}{2} \end{pmatrix} \qquad M\begin{pmatrix} x_I \\ x_R \end{pmatrix} = \begin{pmatrix} (\zeta - s_I)\omega_I \\ (\zeta - s_R)\omega_R \end{pmatrix}.$$

Eliminating (s_I, s_R), we obtain a condition involving x_I and x_R only:

$$\gamma\Sigma_0\begin{pmatrix} x_I \\ x_R \end{pmatrix} = \begin{pmatrix} \zeta - \frac{\gamma}{2} - \frac{M}{\omega_I}x_I \\ \zeta - \frac{\gamma}{2} - \frac{M}{\omega_R}x_R \end{pmatrix}.$$

It is straightforward to verify that this condition coincides with the first-order conditions of the expression for welfare given by (14). Because (14) is concave, it follows that, as claimed in the text, the equilibrium without a PFOF ban leads to the welfare-maximizing choices of (x_I, x_R). It follows trivially that $w_b < w$. □

Lemma 3

Proof. We first derive the posterior order flow characteristics in each marketplace j. In equilibrium, a fraction $\alpha^{(2)} \in [0,1]$ of R-orders are siphoned off-exchange. Hence, $D_2^{(2)} = D_R^{(2)} = 0$ as all off-exchange orders are R-type, and $D_1^{(2)} = w_I^{(2)} D_I^{(2)} + \big(1 - w_I^{(2)}\big) D_R^{(2)} = w_I^{(2)} D_I^{(2)}$, where $w_I^{(2)} := \frac{\omega_I}{\omega_I + (1-\alpha^{(2)})\omega_R}$ is the relative weight of I-orders on-exchange. Given $\mathbb{E}_1\big[D_I^{(2)}\big] = \phi_I D_I^{(1)}$, $\mathrm{var}_1\big[D_I^{(2)}\big] = (1-\phi_I^2)\sigma_I^2$, $\mathbb{E}_1\big[D_R^{(2)}\big] = 0$, and $\mathrm{var}_1\big[D_R^{(2)}\big] = 0$, we obtain[27]

$$\boldsymbol{\mu}^{(2|1)} = \begin{pmatrix} w_I^{(2)} \phi_I D_I^{(1)} \\ 0 \end{pmatrix} \text{ and } \Sigma^{(2|1)} = \begin{pmatrix} \big(w_I^{(2)}\big)^2 (1-\phi_I^2)\sigma_I^2 & 0 \\ 0 & 0 \end{pmatrix}. \tag{24}$$

Next, we derive the objective function (16) of a market maker m, who enters period 2 with inventory $z_m^{(1)}$. Suppose her liquidity supply is $\boldsymbol{x}_m^{(2)} = \big(x_{m1}^{(2)}, x_{m2}^{(2)}\big)^\top$. Then her number of trades $Q_{mj}^{(2)}$ in each marketplace j is Poisson distributed with intensity $x_{mj}^{(2)}$. Her expected spread revenue from marketplace j remains $\mathbb{E}_1\big[Q_{mj}^{(2)} s_j^{(2)}\big] = x_{mj}^{(2)} s_j^{(2)}$. Denote her inventory from marketplace j by $z_{mj}^{(2)}$. At the beginning of period 2, her expectation of her terminal squared inventory is $\mathbb{E}_1\Big[\big(z_m^{(1)} + \sum_{j=1}^2 z_{mj}^{(2)}\big)^2\Big] = \big(z_m^{(1)}\big)^2 + 2 z_m^{(1)} \mathbb{E}_1\big[\sum_{j=1}^2 z_{mj}^{(2)}\big] + \mathbb{E}_1\Big[\big(\sum_{j=1}^2 z_{mj}^{(2)}\big)^2\Big]$. The third term, following Lemma 1, is $\mathbb{E}_1\Big[\big(\sum_{j=1}^2 z_{mj}^{(2)}\big)^2\Big] = \boldsymbol{x}_m^{(2)\top} \mathbf{1} + \boldsymbol{x}_m^{(2)\top} \big(\Sigma^{(2|1)} + \boldsymbol{\mu}^{(2|1)} \boldsymbol{\mu}^{(2|1)\top}\big) \boldsymbol{x}_m^{(2)}$. Directly evaluating the expectation in the second term yields

$$\begin{aligned}
\mathbb{E}_1\Bigg[\sum_{j=1}^2 z_{mj}^{(2)}\Bigg] &= \sum_{j=1}^2 \mathbb{E}_1\Big[\mathbb{E}_1\Big[z_{mj}^{(2)} \Big| Q_{mj}^{(2)}\Big]\Big] = \sum_{j=1}^2 \mathbb{E}_1\Bigg[\mathbb{E}_1\Bigg[\sum_{i=1}^{Q_{mj}^{(2)}} (-1)^{B_{mji}}\Bigg]\Bigg] \\
&= -\sum_{j=1}^2 \mathbb{E}_1\Big[\mathbb{E}_1\Big[Q_{mj}^{(2)} D_j^{(2)}\Big]\Big] = -\sum_{j=1}^2 x_{mj}^{(2)} \mu^{(2|1)} = -\boldsymbol{x}_m^{(2)\top} \boldsymbol{\mu}^{(2|1)},
\end{aligned}$$

where $\big(B_{mji}\big)_{i=1}^{Q_{mj}^{(2)}}$ are i.i.d. Bernoulli draws with success rate $\frac{1}{2}\big(1 + D_j^{(2)}\big)$. Combining the above results, we obtain the objective $\pi_m^{(2)}$ as given in (16).

The first-order condition regarding $x_{m1}^{(2)}$ uniquely solves the optimal $x_{m1}^{(2)}$ (and the second-order

[27] Note that, because we assume $D_R^{(2)} = 0$ is a constant, $\mathrm{var}\big[D_R^{(2)}\big] = 0$, and $\Big(\Sigma^{(2|1)} + \boldsymbol{\mu}^{(2|1)} \boldsymbol{\mu}^{(2|1)\top}\Big)$ is no longer invertible. Hence, unlike in (5), for example, an individual market maker's optimal liquidity supply $\boldsymbol{x}_m^{(2)}$ cannot be uniquely determined. Nevertheless, as the proof shows below, the on-exchange liquidity supply, the *aggregate* off-exchange liquidity supply, and the half spreads $\boldsymbol{s}^{(2)}$ all remain unique in equilibrium.

condition clearly holds):

$$\frac{\mathrm{d}\pi_m^{(2)}}{\mathrm{d}x_{m1}^{(2)}} = s_1^{(2)} - \frac{\gamma}{2} + w_I^{(2)}\gamma\phi_I D_I^{(1)} z_m^{(1)} - \gamma\left(w_I^{(2)}\sigma_I\right)^2 x_{m1}^{(2)} = 0$$

$$\implies x_{m1}^{(2)} = \frac{s_1^{(2)} - \frac{\gamma}{2} + w_I^{(2)}\gamma\phi_I D_I^{(1)} z_m^{(1)}}{\gamma\left(w_I^{(2)}\sigma_I\right)^2}. \tag{25}$$

However, $\frac{\mathrm{d}\pi_m^{(2)}}{\mathrm{d}x_{m2}^{(2)}} = s_2^{(2)} - \frac{\gamma}{2}$ does *not* depend on $\boldsymbol{x}_m^{(2)}$; in other words, the objective $\pi_m^{(2)}$ is linear in $x_{m2}^{(2)}$. Therefore, to satisfy the first-order condition $\frac{\mathrm{d}\pi_m^{(2)}}{\mathrm{d}x_{m2}^{(2)}} = 0$, the equilibrium off-exchange half spread must be

$$s_2^{(2)} = \frac{\gamma}{2}. \tag{26}$$

The on-exchange market-clearing condition is $\int_0^M x_{m1}^{(2)}\mathrm{d}m = \lambda_I\big(s_1^{(2)}\big) + (1-\alpha^{(2)})\lambda_R\big(s_1^{(2)}\big)$, or

$$\frac{M}{\gamma\left(w_I^{(2)}\sigma_I\right)^2}\left(s_1^{(2)} - \frac{\gamma}{2} + w_I^{(2)}\gamma\phi_I D_I^{(1)}\bar{z}^{(1)}\right) = \max\left\{0, \zeta - s_1^{(2)}\right\}\frac{\omega_I}{w_I^{(2)}},$$

using (25) and $\lambda_k(s) = \max\{0, \zeta - s\}\omega_k$. Define

$$\tau^{(2)} := \zeta - \frac{\gamma}{2} + \gamma w_I^{(2)}\phi_I D_I^{(1)}\bar{z}^{(1)}.$$

We then obtain the equilibrium on-exchange half spread

$$s_1^{(2)} = \begin{cases} \frac{\gamma}{2}\frac{M+2w_I^{(2)}\left(\zeta\omega_I\sigma_I^2-\phi_I D_I^{(1)}\bar{z}^{(1)}M\right)}{M+w_I^{(2)}\gamma\omega_I\sigma_I^2} & \text{if } \tau^{(2)} > 0; \\ \frac{\gamma}{2}\left(1 - 2w_I^{(2)}\phi_I D_I^{(1)}\bar{z}^{(1)}\right) & \text{if } \tau^{(2)} \leq 0. \end{cases} \tag{27}$$

Plugging (27) into (25), an individual market maker m's equilibrium on-exchange supply is

$$x_{m1}^{(2)} = \begin{cases} \frac{\left(z_m^{(1)}-\bar{z}^{(1)}\right)M\phi_I D_I^{(1)} + \left(\zeta-\frac{\gamma}{2}+w_I^{(2)}\gamma\phi_I D_I^{(1)} z_m^{(1)}\right)\sigma_I^2\omega_I}{Mw_I^{(2)}\sigma_I^2+\gamma\left(w_I^{(2)}\sigma_I^2\right)^2\omega_I} & \text{if } \tau^{(2)} > 0; \\ \frac{1}{w_I^{(2)}\sigma_I^2}\left(z_m^{(1)} - \bar{z}^{(1)}\right)\phi_I D_I^{(1)} & \text{if } \tau^{(2)} \leq 0. \end{cases} \tag{28}$$

Similarly, the off-exchange market-clearing condition is $\int_0^M x_{m2}^{(2)}\mathrm{d}m = \alpha^{(2)}\lambda_R\big(s_2^{(2)}\big)$. Using (26), we obtain the equilibrium *aggregate* off-exchange liquidity supply

$$\int_0^M x_{m2}^{(2)}\mathrm{d}m = \alpha^{(2)}\Big(\zeta - \frac{\gamma}{2}\Big)\omega_R. \tag{29}$$

The best-execution requirement says that if $s_2^{(2)} < (>) s_1^{(2)}$, then all (no) R-orders are siphoned off-exchange, i.e., $\alpha^{(2)} = 1$ $(= 0)$. If $s_2^{(2)} = s_1^{(2)}$, then $\alpha^{(2)}$ can take any value in $[0, 1]$. Recall the definition of $\Delta^{(2)}$ from (17). On the one hand, if $\tau^{(2)} > 0$, then using the spread expressions (26) and (27) (the $\tau^{(2)} > 0$ case), we have

$$\tau^{(2)} > 0 \Longrightarrow \begin{cases} s_2^{(2)} < s_1^{(2)} \iff w_I^{(2)}\gamma\omega_I\sigma_I^2 < 2w_I^{(2)}\Big(\zeta\omega_I\sigma_I^2 - \phi_I D_I^{(1)}\bar{z}^{(1)}M\Big) \iff \Delta^{(2)} < 0 \\ s_2^{(2)} > s_1^{(2)} \iff w_I^{(2)}\gamma\omega_I\sigma_I^2 > 2w_I^{(2)}\Big(\zeta\omega_I\sigma_I^2 - \phi_I D_I^{(1)}\bar{z}^{(1)}M\Big) \iff \Delta^{(2)} > 0 \end{cases} \tag{30}$$

(Note that $w_I^{(2)}$ is strictly positive and, hence, can be cancelled out without affecting the inequalities above.) On the other hand, if $\tau^{(2)} \leq 0$, then $D_I^{(1)}\bar{z}^{(1)} \leq \frac{1}{\gamma w_I^{(2)}\phi_I}\big(\zeta - \frac{\gamma}{2}\big) < 0$. Using the spread expressions (26) and (27) (the $\tau^{(2)} \leq 0$ case), this inequality implies $s_2^{(2)} < s_1^{(2)}$. Recalling (17), the same inequality also implies $\Delta^{(2)} < 0$. It follows that we have also the following (with the second case holding vacuously)

$$\tau^{(2)} \leq 0 \implies \begin{cases} s_2^{(2)} < s_1^{(2)} \iff \Delta^{(2)} < 0 \\ s_2^{(2)} > s_1^{(2)} \iff \Delta^{(2)} > 0 \end{cases} \tag{31}$$

Combining (30) and (31) with the best-execution requirement, the relationship between $\Delta^{(2)}$ and $\alpha^{(2)}$ is precisely as stated in the lemma.

Finally, assume market clearing from $t = 1$. Then $\bar{z}^{(1)} = -\frac{1}{M}\lambda_I\big(s_1^{(1)}\big)D_I^{(1)}$. Plugging this into (17), the expression for $\Delta^{(2)}$ becomes

$$\Delta^{(2)} = -\frac{\phi_I\lambda_I\big(s_1^{(1)}\big)\sigma_I^2}{\zeta - \frac{\gamma}{2}} - \sigma_I^2\omega_I,$$

which is negative. Following the previous paragraph, best-execution therefore requires $\alpha^{(2)} = 1$, and, hence, $w_I^{(2)} = 1$. □

Proposition 9

Proof. The first step is to characterize the order flow characteristics in each marketplace j. Similar to $t = 2$, defining $w_I^{(1)} = \frac{\omega_I}{\omega_I + (1-\alpha^{(1)})\omega_R}$ as the on-exchange weight of I-orders, we have

$$\boldsymbol{\mu}^{(1)} = \mathbb{E}\begin{bmatrix} D_1^{(1)} \\ D_2^{(1)} \end{bmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \Sigma^{(1)} = \operatorname{var}\begin{bmatrix} D_1^{(1)} \\ D_2^{(1)} \end{bmatrix} = \begin{pmatrix} \big(w_I^{(1)}\big)^2\sigma_I^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

The second step is to derive the objective function $\pi_m^{(1)} = \boldsymbol{x}_m^{(1)\top}\boldsymbol{s}_1^{(1)} + \mathbb{E}\big[\pi_m^{(2)}(\cdot)\big]$. In particular, we are only interested in how it is affected by the supply $\boldsymbol{x}_m^{(1)}$. To do so, we first evaluate $\pi_m^{(2)}(\cdot)$ using the $t = 2$ solution derived above. In particular, recall that under $t = 1$ market clearing, $\Delta^{(2)} < 0$ and so $w_I^{(2)} = 1$. Then plug into (16) the half spread $s_2^{(2)}$ as given by (26) and the posterior moments (24) to get

$$\pi_m^{(2)} = \left(s_1^{(2)} - \frac{\gamma}{2}\right)x_{m1}^{(2)} - \frac{\gamma}{2}\left(\left(z_m^{(1)}\right)^2 - 2\phi_I D_I^{(1)} z_m^{(1)} x_{m1}^{(2)} + \left(x_{m1}^{(2)}\right)^2\right).$$

To substitute in $s_1^{(2)}$ and $x_{m1}^{(2)}$, we need to discuss two cases.[28]

Case 1: Conjecture $\zeta - \frac{\gamma}{2} \le \frac{\gamma\phi_I\sigma_I^2}{M}\lambda_I\big(s_1^{(1)}\big)$. Given market clearing at $t = 1$, it follows that $\tau^{(2)} \le 0$:

$$\tau^{(2)} = \zeta - \frac{\gamma}{2} + \gamma w_I^{(2)}\phi_I D_I^{(1)}\bar{z}^{(1)} = \zeta - \frac{\gamma}{2} - \frac{\gamma\phi_I\sigma_I^2}{M}\lambda_I\big(s_1^{(1)}\big) \le 0,$$

where the first equality is the definition of $\tau^{(2)}$, and the second equality uses that, as discussed in the proof of Lemma 3, $t = 1$ market clearing implies (i) $\bar{z}^{(1)} = -\frac{1}{M}\lambda_I\big(s_1^{(1)}\big)D_I^{(1)}$ and (ii) $w_I^{(2)} = 1$. Substituting the $\tau^{(2)} \le 0$ cases of (27) and (28) into $\pi_m^{(2)}$ and then also, by market clearing, substituting $\bar{z}^{(1)} = -\frac{1}{M}\lambda_I\big(s_1^{(1)}\big)D_I^{(1)}$ into $\pi_m^{(2)}$, we get

$$\pi_m^{(2)} = -\frac{\gamma}{2}(1-\phi_I)^2\big(z_m^{(1)}\big)^2 + \frac{\gamma}{M}\lambda_I\big(s_1^{(1)}\big)\phi_I^2 D_I^{(1)} z_m^{(1)} + \big[\text{terms unaffected by } \boldsymbol{x}_m^{(1)}\big].$$

From the market maker's point of view in $t = 1$, the above involves two random variables: $z_m^{(1)}$ and $D_I^{(1)}$. Recall that $z_m^{(1)} = z_{m1}^{(1)} + z_{m2}^{(1)}$ and, hence, by Lemma 1, $\mathbb{E}\big[\big(z_m^{(1)}\big)^2\big] = \boldsymbol{x}_m^{(1)\top}\mathbf{1} + \boldsymbol{x}_m^{(1)\top}\Big(\Sigma^{(1)} + \boldsymbol{\mu}^{(1)}\boldsymbol{\mu}^{(1)\top}\Big)\boldsymbol{x}_m^{(1)} = x_{m1}^{(1)} + x_{m2}^{(1)} + \big(w_I^{(1)}\sigma_I\big)^2\big(x_{m1}^{(1)}\big)^2$. Note that $z_m^{(1)}$ and $D_I^{(1)}$ are correlated: $\mathbb{E}\big[D_I^{(1)}z_m^{(1)}\big] = \mathbb{E}\big[D_I^{(1)}z_{m1}^{(1)}\big] = -w_I^{(1)}x_{m1}^{(1)}\sigma_I^2$. Therefore, taking the unconditional expectation of $\pi_m^{(2)}$ and adding the $t = 1$ spread revenues $\boldsymbol{x}_m^{(1)\top}\boldsymbol{s}^{(1)}$, we obtain

$$\begin{aligned}\pi_m^{(1)} =& \left(s_1^{(1)} - \frac{\gamma}{2}(1-\phi_I^2)\right)x_{m1}^{(1)} + \left(s_2^{(1)} - \frac{\gamma}{2}(1-\phi_I^2)\right)x_{m2}^{(1)} - \frac{\gamma}{2}(1-\phi_I^2)\big(w_I^{(1)}x_{m1}^{(1)}\sigma_I\big)^2 \\ &- \frac{\gamma\sigma_I^2\phi_I^2}{M}\lambda_I\big(s_1^{(1)}\big)w_I^{(1)}x_{m1}^{(1)} + \big[\text{terms unaffected by } \boldsymbol{x}_m^{(1)}\big].\end{aligned}$$

As in the case of $t = 2$, the first-order conditions with respect to $x_{m1}^{(1)}$ and $x_{m2}^{(1)}$ determine

$$x_{m1}^{(1)} = \frac{\big(s_1^{(1)} - \frac{\gamma}{2}(1-\phi_I^2)\big)M - \lambda_I\big(s_1^{(1)}\big)w_I^{(1)}\gamma\sigma_I^2\phi_I^2}{(1-\phi_I^2)\big(w_I^{(1)}\sigma_I\big)^2 M\gamma}; \quad \text{and} \tag{32}$$

$$s_2^{(1)} = \frac{\gamma}{2}(1-\phi_I^2). \tag{33}$$

On-exchange market clearing requires $\int_0^M x_{m1}^{(1)}\mathrm{d}m = \lambda_I\big(s_1^{(2)}\big) + (1-\alpha^{(1)})\lambda_R\big(s_1^{(2)}\big)$, i.e.,

$$\frac{\big(s_1^{(1)} - \frac{\gamma}{2}(1-\phi_I^2)\big)M - \lambda_I\big(s_1^{(1)}\big)w_I^{(1)}\gamma\sigma_I^2\phi_I^2}{(1-\phi_I^2)\gamma\big(w_I^{(1)}\sigma_I\big)^2} = \max\big\{0, \zeta - s_1^{(1)}\big\}\frac{\omega_I}{w_I^{(1)}}.$$

[28] The off-exchange liquidity supply $x_{m2}^{(2)}$ is not needed: Recall from the proof of Lemma 3 that the equilibrium off-exchange half spread $s_2^{(2)}$ as given in (26) ensures that the off-exchange supply $x_{m2}^{(2)}$ is irrelevant for $\pi_m^{(2)}$.

Assuming $\zeta - s_1^{(1)} \le 0$, then $\lambda_I\big(s_1^{(1)}\big) = 0$, and the above equation gives $s_1^{(1)} = \frac{\gamma}{2}(1-\phi_I^2) \le \frac{\gamma}{2} < \zeta$, which contradicts the assumption. We conclude that $\zeta > s_1^{(1)}$, in which case $\lambda_I\big(s_1^{(1)}\big) = \big(\zeta - s_1^{(1)}\big)\omega_I$ and the above market-clearing condition yields

$$s_1^{(1)} = \frac{\gamma}{2}\frac{(1-\phi_I^2)M + 2w_I^{(1)}\omega_I\sigma_I^2\zeta}{M + w_I^{(1)}\gamma\omega_I\sigma_I^2}. \tag{34}$$

Direct computation gives

$$s_1^{(1)} - s_2^{(1)} = \frac{w_I^{(1)}\gamma\omega_I\sigma_I^2}{M + w_I^{(1)}\gamma\omega_I\sigma_I^2}\Big(\zeta - \frac{\gamma}{2}(1-\phi_I^2)\Big) > 0.$$

Therefore, best-execution requires all R-orders to be siphoned off-exchange in period 1, yielding $\alpha_1^{(1)} = 1$ and $w_I^{(1)} = 1$. Finally, we need to verify the initial conjecture. Using (34) (with $w_I^{(1)} = 1$),

$$\zeta - \frac{\gamma}{2} \le \frac{\gamma\phi_I\sigma_I^2}{M}\lambda_I\big(s_1^{(1)}\big) \iff \zeta \le \frac{\gamma}{2}\frac{M + \big(1-\phi_I+\phi_I^3\big)\gamma\omega_I\sigma_I^2}{M + (1-\phi_I)\gamma\omega_I\sigma_I^2},$$

that is, this case applies if and only if the last inequality holds.

Case 2: Conjecture $\zeta - \frac{\gamma}{2} > \frac{\gamma\phi_I\sigma_I^2}{M}\lambda_I\big(s_1^{(1)}\big)$. Given market clearing at $t = 1$, it follows that $\tau^{(2)} > 0$:

$$\tau^{(2)} = \zeta - \frac{\gamma}{2} + \gamma w_I^{(2)}\phi_I D_I^{(1)}\bar{z}^{(1)} = \zeta - \frac{\gamma}{2} - \frac{\gamma\phi_I\sigma_I^2}{M}\lambda_I\big(s_1^{(1)}\big) > 0,$$

where the first equality is the definition of $\tau^{(2)}$, and the second equality uses that, as discussed in the proof of Lemma 3, $t = 1$ market clearing implies (i) $\bar{z}^{(1)} = -\frac{1}{M}\lambda_I\big(s_1^{(1)}\big)D_I^{(1)}$ and (ii) $w_I^{(2)} = 1$. Substituting the $\tau^{(2)} > 0$ cases of (27) and (28) into $\pi_m^{(2)}$ and then also, by market clearing, substituting $\bar{z}^{(1)} = -\frac{1}{M}\lambda_I\big(s_1^{(1)}\big)D_I^{(1)}$ into $\pi_m^{(2)}$, we get

$$\pi_m^{(2)} = -\frac{\gamma}{2}(1-\phi_I^2)\big(z_m^{(1)}\big)^2 + \frac{\big(\zeta - \frac{\gamma}{2}\big)\omega_I + \phi_I\lambda_I\big(s_1^{(1)}\big)}{M + \gamma\omega_I\sigma_I^2}\gamma\phi_I D_I^{(1)} z_m^{(1)} + \big[\text{terms unaffected by } \boldsymbol{x}_m^{(1)}\big].$$

As in the previous case, taking the unconditional expectation of $\pi_m^{(2)}$ and adding the $t = 1$ spread revenues $\boldsymbol{x}_m^{(1)\top}\boldsymbol{s}^{(1)}$, we obtain

$$\begin{aligned}\pi_m^{(1)} =& \Big(s_1^{(1)} - \frac{\gamma}{2}(1-\phi_I^2)\Big)x_{m1}^{(1)} + \Big(s_2^{(1)} - \frac{\gamma}{2}(1-\phi_I^2)\Big)x_{m2}^{(1)} - \frac{\gamma}{2}(1-\phi_I^2)\big(w_I^{(1)}x_{m1}^{(1)}\sigma_I\big)^2 \\ & - \frac{\gamma\sigma_I^2\phi_I}{M + \gamma\sigma_I^2\omega_I}\Big(\Big(\zeta - \frac{\gamma}{2}\Big)\omega_I + \phi_I\lambda_I\big(s_1^{(1)}\big)\Big)w_I^{(1)}x_{m1}^{(1)} + \big[\text{terms unaffected by } \boldsymbol{x}_m^{(1)}\big].\end{aligned}$$

Then, as before, the first-order conditions pin down

$$x_{m1}^{(1)} = \frac{\left(s_1^{(1)} - \frac{\gamma}{2}(1-\phi_I^2)\right)\left(M + \gamma\sigma_I^2\omega_I\right) - \lambda_I\left(s_1^{(1)}\right)w_I^{(1)}\gamma\sigma_I^2\phi_I^2 - \left(\zeta - \frac{\gamma}{2}\right)w_I^{(1)}\phi_I\gamma\sigma_I^2\omega_I}{(1-\phi_I^2)\left(w_I^{(1)}\sigma_I\right)^2(M+\gamma\sigma_I^2\omega_I)\gamma} \quad \text{and ;} \tag{35}$$

$$s_2^{(1)} = \frac{\gamma}{2}\left(1 - \phi_I^2\right). \tag{36}$$

On-exchange market clearing requires $\int_0^M x_{m1}^{(1)}\mathrm{d}m = \lambda_I\left(s_1^{(2)}\right) + (1-\alpha^{(1)})\lambda_R\left(s_1^{(2)}\right)$, i.e.,

$$M \cdot \frac{\left(s_1^{(1)} - \frac{\gamma}{2}(1-\phi_I^2)\right)\left(M + \gamma\sigma_I^2\omega_I\right) - \lambda_I\left(s_1^{(1)}\right)w_I^{(1)}\gamma\sigma_I^2\phi_I^2 - \left(\zeta - \frac{\gamma}{2}\right)w_I^{(1)}\phi_I\gamma\sigma_I^2\omega_I}{(1-\phi_I^2)\left(w_I^{(1)}\sigma_I\right)^2(M+\gamma\sigma_I^2\omega_I)\gamma} = \max\left\{0, \zeta - s_1^{(1)}\right\}\frac{\omega_I}{w_I^{(1)}}.$$

Assuming $\zeta - s_1^{(1)} \leq 0$, then $\lambda_I\left(s_1^{(1)}\right) = 0$, and the above equation then gives

$$s_1^{(1)} - \zeta = \frac{\frac{\gamma}{2}(1-\phi_I^2)\cdot M + \left(\frac{\gamma}{2}(1-\phi_I^2) + \left(\zeta - \frac{\gamma}{2}\right)\phi_I w_I^{(1)}\right)\gamma\sigma_I^2\omega_I}{M + \gamma\sigma_I^2\omega_I} - \zeta,$$

which decreases in ζ and, hence, when $\zeta \downarrow \frac{\gamma}{2}$, reaches its maximum of $-\frac{\gamma}{2}\phi_I^2 < 0$, thus rejecting the assumption. We conclude that $\zeta > s_1^{(1)}$, in which case $\lambda_I\left(s_1^{(1)}\right) = \left(\zeta - s_1^{(1)}\right)\omega_I$ and the above market-clearing condition yields

$$s_1^{(1)} = \frac{\gamma}{2}\frac{(1-\phi_I^2)(M+\gamma\omega_I\sigma_I^2)\left(M + 2\zeta w_I^{(1)}\omega_I\sigma_I^2\right) + \left(2(\phi_I + \phi_I^2)\zeta - \phi_I\gamma\right)M w_I^{(1)}\omega_I\sigma_I^2}{(M+\gamma\omega_I\sigma_I^2)\left(M + \gamma w_I^{(1)}\omega_I\sigma_I^2\right) - w_I^{(1)}\phi_I^2\gamma^2\omega_I^2\sigma_I^4}. \tag{37}$$

Directly comparing the two half spreads in this case yields

$$s_1^{(1)} - s_2^{(2)} = \frac{\gamma}{2}\frac{\left[2(1+\phi_I)\zeta - \gamma\right]\phi_I M w_I^{(1)}\omega_I\sigma_I^2 + (1-\phi_I^2)w_I^{(1)}\phi_I^2\gamma^2\omega_I^2\sigma_I^4}{M^2 + (1 + w_I^{(1)})M\gamma\omega_I\sigma_I^2 + (1-\phi_I^2)w_I^{(1)}\gamma^2\omega_I^2\sigma_I^4} > 0,$$

where the last inequality is guaranteed by $\zeta > \frac{\gamma}{2}$. Therefore, best-execution requires all *R*-orders to be routed off-exchange in period 1, yielding $\alpha^{(1)} = 1$ and $w_I^{(1)} = 1$. Finally, we need to verify the initial conjecture. Using (37) (with $w_I^{(1)} = 1$),

$$\zeta - \frac{\gamma}{2} > \frac{\gamma\phi_I\sigma_I^2}{M}\lambda_I\left(s_1^{(1)}\right) \iff \zeta > \frac{\gamma}{2}\frac{M + \left(1 - \phi_I + \phi_I^3\right)\gamma\omega_I\sigma_I^2}{M + (1-\phi_I)\gamma\omega_I\sigma_I^2},$$

Summary: As seen above, in either case, $\alpha^{(1)} = 1$ (and, hence, $w_I^{(1)} = 1$). Define

$$\tau^{(1)} := \zeta - \frac{\gamma}{2}\frac{M + \left(1 - \phi_I + \phi_I^3\right)\gamma\omega_I\sigma_I^2}{M + (1-\phi_I)\gamma\omega_I\sigma_I^2}.$$

Then, combining (34) and (37), we obtain the equilibrium on-exchange half spread as

$$s_1^{(1)} = \begin{cases} \frac{\gamma}{2}\frac{(1-\phi_I^2)(M+\gamma\omega_I\sigma_I^2)\left(M+2\zeta\omega_I\sigma_I^2\right)+\left(2(\phi_I+\phi_I^2)\zeta-\phi_I\gamma\right)M\omega_I\sigma_I^2}{\left(M+\gamma\omega_I\sigma_I^2\right)^2-\phi_I^2\gamma^2\omega_I^2\sigma_I^4}, & \text{if } \tau^{(1)} > 0; \\ \frac{\gamma}{2}\frac{(1-\phi_I^2)M+2\omega_I\sigma_I^2\zeta}{M+\gamma\omega_I\sigma_I^2}, & \text{if } \tau^{(1)} \le 0. \end{cases} \tag{38}$$

Combining (32) and (35), then plugging in (38), we obtain an individual market maker *m*'s on-exchange equilibrium liquidity supply as

$$x_{m1}^{(1)} = \begin{cases} \frac{\left(\zeta-\frac{\gamma}{2}(1-\phi_I^2)\right)M\omega_I+\left((1-\phi_I)\zeta-\frac{\gamma}{2}(1-\phi_I-\phi_I^2)\right)\gamma\omega_I^2\sigma_I^2}{\left(M+\gamma\omega_I\sigma_I^2\right)^2-\phi_I^2\gamma^2\omega_I^2\sigma_I^4}, & \text{if } \tau^{(1)} > 0; \\ \frac{\left(\zeta-\frac{\gamma}{2}(1-\phi_I^2)\right)\omega_I}{M+\gamma\omega_I\sigma_I^2}, & \text{if } \tau^{(1)} \le 0. \end{cases} \tag{39}$$

Combining (33) and (36), we obtain the equilibrium off-exchange half spread as

$$s_2^{(1)} = \frac{\gamma}{2}\left(1-\phi_I^2\right). \tag{40}$$

Using (40), by market clearing, we obtain the equilibrium *aggregate* off-exchange liquidity supply

$$\int_0^M x_{m2}^{(1)}\mathrm{d}m = \left(\zeta - \left(1-\phi_I^2\right)\frac{\gamma}{2}\right)\omega_R. \tag{41}$$

□

C The effect of order flow correlation

The order flow correlation parameter ρ is a key driver of market makers' incentives. Indeed, as ρ decreases, *R*-orders offset *I*-orders more often, thus lowering the inventory risk of a given portfolio of order flows. We study this correlation channel in this appendix. To focus on this channel, we eliminate effects that may stem from differences between σ_I and σ_R by assuming, for this appendix only, that

$$\sigma_I = \sigma_R = \sigma. \tag{42}$$

To be consistent with the real world, we focus on the parameter ranges where there is siphoning of *R*-orders and where there is positive on-exchange volume; that is, both $\Delta < 0$ and (12) hold (*cf.* the assumption on page 23). With (42), assuming $\Delta < 0$ simplifies to assuming $\omega_R < \omega_I$, which we consider realistic: as discussed in Section 3.3, retail trading activity tends to be far smaller than that of institutions, despite its growth in recent years. Moreover, (42) also ensures condition (12). Hence, the focal parameter ρ has its full support on $(-1, 1)$ and we can exogenously vary it to examine how bid-ask spreads are affected.

Figure 6 illustrates the main findings. One pattern is that the price improvement $s_1 - s_2$ is decreasing in ρ. This is rather intuitive: at the extreme of $\rho = 1$, R-orders are interchangeable with I-orders, but as ρ decreases, R-orders become progressively more effective in hedging I-orders, and accordingly receive more price improvement. At first glance, it also appears that all spreads are increasing in ρ. Such a pattern might also seem intuitive: as ρ increases, inventories resulting from more correlated orders offset each other less often, adding to the expected inventory costs of the market makers, who compensate by charging larger spreads. Upon a more careful examination, however, the on-exchange spread s_1 (without the ban) can actually be hump-shaped in ρ, as illustrated in Figure 6(b). Formally, we have the following proposition.

Proposition 10 (Order flow correlation and spreads). Suppose condition (42) holds. The half spreads s_2 and s_b unambiguously increase monotonically in correlation ρ. If

$$M < (\omega_I - \omega_R)\gamma\sigma^2, \tag{43}$$

then s_1 is hump-shaped in ρ; or else, s_1 also increases monotonically in ρ. The price improvement $s_1 - s_2$ decreases monotonically in ρ.

Proof. Under $\Delta < 0$ and under (12), the equilibrium features both on-exchange and off-exchange volume. Hence, following the proofs of Proposition 2 and Corollary 1, $s_j = \frac{\gamma}{2} + \left(\zeta - \frac{\gamma}{2}\right)\beta_j$ for $j \in \{1, 2, b\}$, where β_1, β_2, and β_b are given by (19), (20), and (21), respectively. Therefore, to sign $\frac{\mathrm{d}s_j}{\mathrm{d}\rho}$, it equivalent to examine $\frac{\mathrm{d}\beta_j}{\mathrm{d}\rho}$.

Consider first s_2. Careful evaluation shows that $\operatorname{sign}\left[\frac{\mathrm{d}\beta_2}{\mathrm{d}\rho}\right] = \operatorname{sign}\left[M^2 + \left(\omega_I + (1-2\rho)\omega_R\right)\sigma^2\gamma M + (1-\rho)^2\sigma^4\omega_I\omega_R\gamma^2\right]$. Note that, under (42), the maintained assumption of $\Delta < 0$ becomes $\omega_I > \omega_R$. Hence, $\omega_I + (1-2\rho)\omega_R > 2(1-\rho)\omega_R > 0$. Therefore, $\frac{\mathrm{d}\beta_2}{\mathrm{d}\rho} > 0$.

Consider next s_b. Direct computation gives that $\frac{\mathrm{d}\beta_b}{\mathrm{d}\rho} = \frac{2M\gamma\sigma^2\omega_I\omega_R(\omega_I+\omega_R)^2}{\left((\omega+\omega_R)^2M+(\omega_I^2+2\rho\omega_I\omega_R+\omega_R^2)\sigma^2\gamma\right)^2} > 0$.

Finally, consider s_1. Careful evaluation shows that

$$\operatorname{sign}\left[\frac{\mathrm{d}\beta_1}{\mathrm{d}\rho}\right] = \operatorname{sign}\left[\rho^2 - 2\frac{M+\gamma\sigma^2\omega_R}{\gamma\sigma^2\omega_R}\rho + \frac{(M+\gamma\sigma^2\omega_I)(M+\gamma\sigma^2\omega_R)}{\gamma^2\sigma^4\omega_I\omega_R}\right],$$

Under (42), neither $\Delta < 0$ nor (12) imply any constraints on ρ. Hence, this quadratic expression in ρ has the full domain $\rho \in (-1, 1)$. It also has at most one root in $(0, 1)$ and is strictly positive if $\rho \to -1$. Therefore, the sign of the quadratic expression in the limit as $\rho \to 1$ determines the shape of β_1 in $\rho \in (-1, 1)$: If it is negative, then β_1 is initially increasing from $\rho \to -1$, peaks at the quadratic expression's unique root in $(0, 1)$, and then decreases toward $\rho \to 1$. If it is weakly positive, then β_1 is monotonically increasing throughout $(-1, 1)$. Indeed, evaluating the sign of the quadratic expression at $\rho = 1$ yields (43). □

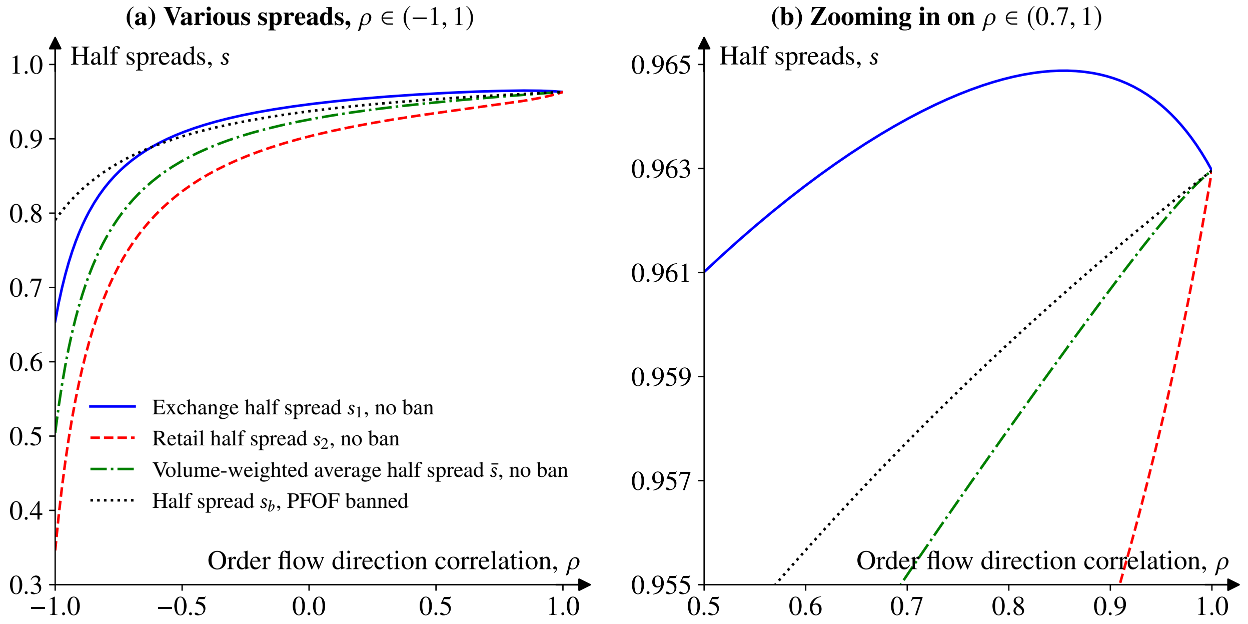


Figure 6: Bid-ask spreads: the order flow correlation. This figure shows how various bid-ask (half) spreads change as the order flow correlation ρ increases. Panel (a) plots the four different spreads across the full support of $\rho \in (-1, 1)$. Panel (b) zooms in on the support of $\rho \in (0.5, 1)$. The other parameters are set at $M = 3$, $\gamma = \zeta = 1$, $\omega_I = 100$, $\omega_R = 50$, $\sigma_I = \sigma_R = 0.5$, and $\mu_I = \mu_R = 0$.

Below we provide an intuitive discussion to shed light into this potential hump shape of s_1. From the market makers' first-order conditions and the market-clearing conditions, it can be shown that ρ affects s_1 via two channels:[29]

$$\frac{\mathrm{d}s_1}{\mathrm{d}\rho} = \frac{\gamma\sigma^2}{M + \gamma\sigma^2}\Big(\underbrace{\lambda_R(s_2)}_{\substack{\text{Direct effect}\\ (>0)}} + \underbrace{\rho\frac{\mathrm{d}\lambda_R}{\mathrm{d}s_2}\frac{\mathrm{d}s_2}{\mathrm{d}\rho}}_{\substack{\text{Indirect effect}\\ (<0,\text{ iff }\rho>0)}}\Big). \tag{44}$$

The positive direct effect corresponds to the hedging intuition alluded to earlier: an increase in ρ reduces the extent to which R-orders can be used to hedge I-orders, hence increasing the marginal inventory cost of an I-order. The market makers then charge a higher s_1 to compensate for the higher marginal cost. Moreover, this direct effect is proportional to $\lambda_R(s_2)$—it is stronger if market makers handle more R-orders.

Of course, the number of R-orders that a market maker handles is itself influenced by ρ, creating the above indirect effect. In particular, when ρ increases, s_2 increases (due to the same hedging intuition mentioned above), reducing R-investor volume, which, in the case of $\rho > 0$, lowers the marginal inventory cost of an I-order and, hence, also s_1.

When might this indirect effect be strong enough to overturn the direct effect and generate a hump shape as seen in Figure 6(b)? Proposition 10 gives the exact condition (43). For example, the condition indicates that no hump shape arises if M is too large. To see why, note that equation (44) indicates that the positive direct effect must dominate if $\frac{\mathrm{d}s_2}{\mathrm{d}\rho}$ is small. This is true if M is sufficiently large: in the limit as $M \to \infty$, each market maker expects to receive at most one order, in which case ρ has no effect on market makers' inventory costs, so that $\frac{\mathrm{d}s_2}{\mathrm{d}\rho} \to 0$. The condition (43) also indicates that no hump shape arises if $\gamma\sigma^2$ is too small. This is because, in the limit as $\gamma\sigma^2 \to 0$, inventory costs disappear (regardless of ρ). Hence, $\frac{\mathrm{d}s_2}{\mathrm{d}\rho} \to 0$ in this limit also, again implying a small indirect effect.[30]

[29] To derive (44), recall that a market maker's expected payoff can be written as (8). In particular, in the equilibrium of interest, all R-orders are siphoned off exchange and, hence, $\sigma_1 = \sigma_I$, $\sigma_2 = \sigma_R$, and $r = \rho$. Also, thanks to (42), we can write $\sigma := \sigma_I = \sigma_R$. The market maker's first-order condition with respect to her on-exchange liquidity supply x_{m1} is

$$\frac{\partial \pi_m}{\partial x_{m1}} = \left(s_1 - \frac{\gamma}{2}\right) - \gamma\sigma^2 \cdot (x_{m1} + \rho x_{m2}) = 0.$$

Note that, in equilibrium, $Mx_{m1} = \lambda_I(s_1) = (\zeta - s_1)\omega_I$ and $Mx_{m2} = \lambda_R(s_2)$ by market clearing. Therefore, the above first-order condition yields the following relation between the on-exchange spread s_1 and the off-exchange spread s_2:

$$s_1 = \frac{\gamma}{2} + \frac{\gamma\sigma^2}{M + \gamma\sigma^2}\left(\left(\zeta - \frac{\gamma}{2}\right) + \rho\lambda_R(s_2)\right),$$

from which (44) follows.

[30] Proposition 10 states that, unlike the on-exchange spread s_1 (for I-orders), the off-exchange spread s_2 (for R-orders) unambiguously increases monotonically in ρ. One might wonder what asymmetry between the two types of orders drives this difference. Indeed, following the same derivation above, one can write s_2 as a function of s_1 (similar to Footnote 29) and similarly identify both a positive direct effect and an indirect effect, whose sign depends on ρ.

References

Adams, Samuel and Connor Kasten. 2021. "Retail Order Execution Quality under Zero Commissions." Working paper.

Amihud, Yakov and Haim Mendelson. 1980. "Dealership Markets: Market Making with Inventory." *Journal of Financial Economics* 8 (1):31–53.

Babus, Ana and Cecilia Parlatore. 2022. "Strategic Fragmented Markets." *Journal of Financial Economics* 145 (3):876–908.

Baldauf, Markus and Joshua Mollner. 2020. "Trading in Fragmented Markets." *Journal of Financial and Quantitative Analysis* 56 (1):93–121.

Barardehi, Yashar H., Dan Bernhardt, Zhi Da, and Mitch Warachka. 2022. "Internalized Retail Order Imbalances and Institutional Liquidity Demand." Working paper.

Barber, Brad M., Xing Huang, Terrance Odean, and Christopher Schwarz. 2022. "Attention-Induced Trading and Returns: Evidence from Robinhood Users." *The Journal of Finance* 77 (6):3141–3190.

Barbon, Andrea, Marco Di Maggio, Francesco Franzoni, and Augustin Landier. 2019. "Brokers and Order Flow Leakage: Evidence from Fire Sales." *The Journal of Finance* 74 (6):2707–2749.

Barron's. 2021. "SEC Chairman Says Banning Payment for Order Flow Is 'On the Table'." https://bit.ly/3VfvCS6.

Battalio, Robert, Jason Greene, and Robert Jennings. 1997. "Do Competing Specialists and Preferencing Dealers Affect Market Quality?" *The Review of Financial Studies* 10 (4):969–993.

Battalio, Robert and Craig W. Holden. 2001. "A Simple Model of Payment for Order Flow, Internalization, and Total Trading Cost." *Journal of Financial Markets* 4:33–71.

Battalio, Robert H. 1997. "Third Market Broker-Dealers: Cost Competitors or Cream Skimmers?" *The Journal of Finance* 52 (1):341–352.

Ben-Rephael, Azi, Shmuel Kandel, and Avi Wohl. 2011. "The Price Pressure of Aggregate Mutual Fund Flows." *Journal of Financial and Quantitative Analysis* 46 (2):585–603.

———. 2012. "Measuring Investor Sentiment with Mutual Fund Flows." *Journal of Financial Economcis* 104 (4):363–382.

Bernhardt, Dan, Vladimir Dvoracek, Eric Hughson, and Ingrid M. Werner. 2004. "Why Do Larger Orders Receive Discounts on the London Stock Exchange?" *The Review of Financial Studies* 18 (4):1343–1368.

Bloomberg. 2022. "Retail Trading Army Is Seeing Power Wane in Stock Market It Once Ruled." https://bloom.bg/3O4OmBs.

Boehmer, Ekkehart, Charles M. Jones, Xiaoyan Zhang, and Xinran Zhang. 2021. "Tracking Retail Investor Activity." *The Journal of Finance* 76 (5):2249–2305.

However, it can be shown that for s_2, the indirect effect is never large enough to overturn the direct effect, as long as $\omega_R < \omega_I$ holds (even if (42) does not hold).

Brogaard, Jonathan and Corey Garriott. 2019. "High-Frequency Trading Competition." *Journal of Financial and Quantitative Analysis* 54 (4):1469–1497.

Bruche, Max and John C.F. Kuong. 2021. "Dealer Funding and Market Liquidity." Working paper.

Cespa, Giovanni and Xavier Vives. 2022. "Exchange Competition, Entry, and Welfare." *The Review of Financial Studies* 35 (5):2570–2624.

Chao, Yong, Chen Yao, and Mao Ye. 2019. "Why Discrete Price Fragments U.S. Stock Exchanges and Disperses Their Fee Structures." *The Review of Financial Studies* 32 (3):1068–1101.

Chen, Daniel and Darrell Duffie. 2021. "Market Fragmentation." *American Economic Review* 111 (7):2247–74.

Chordia, Tarun and Avanidhar Subrahmanyam. 1995. "Market Making, the Tick Size, and Payment-for-Order Flow: Theory and Evidence." *Journal of Business* 68 (5):543–575.

Chowdhry, Bhagwan and Vikram Nanda. 1991. "Multimarket Trading and Market Liquidity." *Review of Financial Studies* 4 (3):483–511.

Comerton-Forde, Carole, Katya Malinova, and Andreas Park. 2018. "Regulating Dark Trading: Order Flow Segmentation and Market Quality." *Journal of Financial Economics* 130 (2):347–366.

Corts, Kenneth S. 1998. "Third-Degree Price Discrimination in Oligopoly: All-Out Competition and Strategic Commitment." *The RAND Journal of Economics* 29 (2):306–323.

Coval, Joshua and Erik Stafford. 2007. "Asset Fire Sales (and Purchases) in Equity Markets." *Journal of Financial Economics* 86 (2):479–512.

Daures-Lescourret, Laurence and Sophie Moinas. 2022. "Fragmentation and Strategic Market-Making." *Journal of Financial and Quantitative Analysis* Forthcoming:1–26.

Degryse, Hans, Frank de Jong, and Vincent van Kervel. 2015. "The Impact of Dark Trading and Visible Fragmentation on Market Quality." *Review of Finance* 19 (4):1587–1622.

Desgranges, Gabriel and Theirry Foucault. 2005. "Reputation-Based Pricing and Price Improvements." *Journal of Economics and Business* 57 (6):493–527.

Duffie, Darrell, Lei Qiao, and Yeneng Sun. 2020. "Continuous Time Random Matching." Working paper.

Easley, David, Nicholas M. Kiefer, and Maureen O'Hara. 1996. "Cream-Skimming or Profit-Sharing? The Curious Role of Purchased Order Flow." *The Journal of Finance* 51 (3):811–833.

Eaton, Gregory W., T. Clifton Green, Brian S. Roseman, and Yanbin Wu. 2022. "Retail Trader Sophistication and Stock Market Quality: Evidence from Brokerage Outages." *Journal of Financial Economics* 146 (2):502–528.

Elsas, Ralf, Lutz Johanning, and Erik Theissen. 2022. "Payment for Order Flow and Market Quality: A Field Experiment." Working paper.

Ernst, Thomast and Chester Spatt. 2022. "Payment for Order Flow and Asset Choice." Working paper.

ESMA. 2021. "ESMA warns Firms and Investors about Risks Arising from Payment for Order

Flow.” https://bit.ly/3XihWrA.

Financial Times. 2021. “Citadel Securities Founder ‘Quite Fine’ with Ending Payment for Order Flow.” https://on.ft.com/3OxLj3y.

Foucault, Thierry and Albert J. Menkveld. 2008. “Competition for Order Flow and Smart Order Routing Systems.” *The Journal of Finance* 63 (1):119–158.

Garman, Mark B. 1976. “Market Microstructure.” *Journal of Financial Economics* 3:257–275.

Garriott, Corey and Adrian Walton. 2018. “Retail Order Flow Segmentation.” *The Journal of Trading* 13 (3):13–23.

Gensler, Gary. 2022. “SEC’s Meme Stock Response Coming Next Week, Gensler Says.” https://bloom.bg/3UKA7EK.

Glode, Vincent and Christian Opp. 2016. “Asymmetric Information and Intermediation Chains.” *American Economic Review* 106 (9):2699–2721.

Glossner, Simon, Pedro Matos, Stefano Ramelli, and Alexander F. Wagner. 2022. “Do Institutional Investors Stabilize Equity Markets in Crisis Periods? Evidence from COVID-19.” Working paper.

Glosten, Lawrence R. and Paul R. Milgrom. 1985. “Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Agents.” *Journal of Financial Economics* 42 (1):71–100.

Greenwood, Robin. 2005. “Short- and Long-Term Demand Curves for Stocks: Theory and Evidence on the Dynamics of Arbitrage.” *Journal of Financial Economics* 75:607–649.

Griffin, Ken. 2021. “Testimony of Kenneth C. Griffin, Founder and CEO of Citadel and Founder and Principal Shareholder of Citadel Securities, before the Committee on Financial Services, United States House of Representatives.” Report.

Hagerty, Kathleen and Robert L. McDonald. 1996. “Brokerage, Market Fragmentation, and Securities Market Regulation.” In *The Industrial Organization and Regulation of the Securities Industry*. University of Chicago Press, 35–62.

Harris, Lawrence and Eitan Gurel. 1986. “Price and Volume Effects Associated with Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures.” *The Journal of Finance* 41 (4):815–829.

Hatheway, Frank, Amy Kwan, and Hui Zheng. 2017. “An Empirical Analysis of Market Segmentation on U.S. Equity Markets.” *Journal of Financial and Quantitative Analysis* 52 (6):2399–2427.

Hendershott, Terrence and Albert J. Menkveld. 2014. “Price Pressures.” *Journal of Financial Economics* 114 (3):405–423.

Ho, Thomas and Hans R. Stoll. 1981. “Optimal Dealer Pricing under Transaction Cost and Return Uncertainty.” *Journal of Financial Economics* 9 (1):47–73.

———. 1983. “The Dynamics of Dealer Markets Under Competition.” *The Journal of Finance* 38:1053–1074.

Holmes, Thomas J. 1989. “The Effects of Third-Degree Price Discrimination in Oligopoly.” *The American Economic Review* 79 (1):244–250.

Hu, Edwin and Dermot Murphy. 2022. "Competition for Retail Order Flow and Market Quality." Working paper.

Jain, Pankaj K., Suchismita, Mishra, Shawn O'Donoghue, and Le Zhao. 2021. "Trading Volume Shares and Market Quality: Pre- and Post-Zero Commissions." Working paper.

Jones, Charle M., Donghui Shi, Xiaoyan Zhang, and Xinran Zhang. 2022. "Understanding Retail Investors: Evidence from China." Working paper.

Kandel, Eugene and Leslie M. Marx. 1999. "Payments for Order Flow on Nasdaq." *The Journal of Finance* 54 (1):35–66.

Kyle, Albert S. 1985. "Continuous Auctions and Insider Trading." *Econometrica* 53 (6):1315–1336.

Kyle, Albert S. and Anna Obizhaeva. 2016. "Market Microstructure Invariance: Empirical Hypotheses." *Econometrica* 84 (4):1345–1404.

Mackintosh, Phil. 2022. "Retail Activity Remains Strong in U.S. Markets." https://bit.ly/3YQXk9k.

Pagano, Marco. 1989. "Trading Volume and Asset Liquidity." *The Quarterly Journal of Economics* 104 (2):255–274.

Pagnotta, Emiliano and Thomas Philippon. 2018. "Competing on Speed." *Econometrica* 86 (3):1067–1115.

Parlour, Christine and Uday Rajan. 2003. "Payment for Order Flow." *Journal of Financial Economics* 68:379–411.

Reuters. 2023. "EU lawmakers set up clash with member states over share trade ban." https://reut.rs/3n3TTil.

Schwarz, Christopher, Brad M. Barber, Xing Huang, Philippe Jorion, and Terrance Odean. 2022. "The 'Actual Retail Price' of Equity Trades." Working paper.

SEC. 2021. "Staff Report on Equity and Options Market Structure Conditions in Early 2021." Report, SEC.

———. 2022. "Order Competition Rule (proposed, release No. 34-96495)." https://www.sec.gov/rules/proposed/2022/34-96495.pdf.

Stole, Lars A. 2007. "Chapter 34: Price Discrimination and Competition." In *Handbook of Industrial Organization*, vol. 3, edited by M. Armstrong and R. Porter. Elsevier, 2221–2299.

Stoll, Hans R. 1978. "The Supply of Dealer Services in Securities Markets." *The Journal of Finance* 33 (4):1133–1151.

van Kervel, Vincent. 2015. "Competition for Order Flow with Fast and Slow Traders." *Review of Financial Studies* 28 (7):2094–2127.

Yang, Liyan and Haoxiang Zhu. 2020. "Back-Running: Seeking and Hiding Fundamental Information in Order Flows." *The Review of Financial Studies* 33 (4):1484–1533.

Ye, Mao. 2011. "A Glimpse into the Dark: Price Formation, Transaction Costs and Market Share of the Crossing Network." Working paper.

Zhu, Haoxiang. 2014. "Do Dark Pools Harm Price Discovery?" *Review of Financial Studies* 27 (3):747–789.