Anticompetitive Price Referencing*

Vincent van Kervel† Bart Zhou Yueshen‡

This version: July 18, 2024

* This paper benefits tremendously from helpful discussions with Paul Barton (discussant), Sabrina Buti (discussant), Carole Comerton-Forde, Giovanni Cespa, Jérôme Dugast (discussant), Tom Ernst, Sergei Glebkin, Naveen Gondhi, Shiyang Huang, Pete Kyle (discussant), Joshua Mollner, Emiliano Pagnotta, Gideon Saar (discussant), Chester Spatt, Jian Sun, Jonathan Wallen (discussant), Milena Wittwer, and Yajun Wang (discussant); and the seminar and conference participants at INSEAD, Finance UC conference 2024, Plato MI3 innovator London, Rotman School of Management, Microstructure Exchange, Lee Kong Chian School of Business, 6th Sydney Market Microstructure and Digital Finance Meeting, Rotterdam Liquidity Conference 2024, SFS Cavalcade 2024, 6th Future of Financial Information Conference, Shanghai Advanced Institute of Finance, and Renmin Business School. There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.

† Vincent van Kervel (vincentvankervel@gmail.com) is affiliated with the School of Business and Economics, Universidad de los Andes, Chile and gratefully acknowledges the financial support of the Fondecyt Regular (project 1231072).

‡ Bart Zhou Yueshen (b@yueshen.me) is affiliated with Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road, 178899 Singapore.

Anticompetitive Price Referencing

Abstract

Off-exchange trades are often executed by referencing on-exchange prices. In equilibrium, such price referencing softens market makers' on-exchange competition and makes liquidity expensive for investors. Additionally, by equalizing on- and off-exchange prices, price referencing guarantees "best-execution" and makes investors indifferent where to trade. Market makers effectively obtain a license to fragment orders off exchange, raising their profits but reinforcing market-wide illiquidity. This inefficiency remains tenacious even if more market makers enter and if they are forced to compete off exchange, as in the SEC's proposed order-by-order auction. The model yields important implications for regulating off-exchange trading.

Keywords: order protection rule, market fragmentation, payment for order flow, order-by-order auctions

(There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.)

1 Introduction

Close to half of trading volume in U.S. equity markets currently takes place off exchange: dark pools, broker-dealer internalization, over-the-counter trading, and wholesalers' direct execution of (retail) orders from payment for order flow (PFOF) arrangements.[1] Such off-exchange trading is subject to Regulation NMS Rule 611, the Order Protection Rule (OPR), which requires the off-exchange trade to execute at or inside the prevailing on-exchange bid and ask quotations. Under such regulation, off-exchange trades need to refer to exchanges for pricing, and this "price referencing" practice has become ubiquitous: For example, broker-dealers reference the prevailing national best bid and offer (NBBO) when they internalize client orders or retail orders received via PFOF. The same happens when dark pools match buy and sell orders at prices inside the NBBO, and in size discovery mechanisms such as "workup" and "matching sessions." In Europe, dark pools are required under MiFID II to reference widely published prices. As yet another example, the SEC's Order Competition Rule 615 proposal requires off-exchange retail orders to be auctioned, order by order, with a price compliant with the OPR (Footnote 250 of SEC, 2022b).

We develop a theoretical model to study the implications of such price referencing. The core idea is that, while price referencing is about how trades are executed off exchange, it affects market makers' on-exchange incentive to supply liquidity, which in turn determines the "reference price" for off-exchange trades. Specifically, our analysis yields the following two main insights. When off-exchange trades refer to on-exchange prices:

(i) Market makers compete less on-exchange, lowering market liquidity and profiting from investors' heightened trading cost.

(ii) Market makers can divert order flow off exchange and endogenously fragment the market. This is because price referencing equalizes on- and off-exchange prices, making investors indifferent

[1] In the first quarter of 2023, 54% of the U.S. equity volume is executed on-exchange, 10.7% in ATS/dark pools, and the remaining 35% are in the form of broker-dealer internalization, PFOF arrangements, and over-the-counter trading (TD Cowen, 2023).

about where to trade and fulfilling brokers' best-execution obligation.[2]

These two effects reinforce each other: Because of (i), market makers have incentives to fragment order flow off exchange, and because of (ii), they can do so easily, for example, via PFOF arrangements with brokers. We will show that these reinforcing effects remain tenacious even as the market-making sector grows.

To the best of our knowledge, we are the first to theoretically study price referencing and find that it jointly affects market liquidity (higher trading costs as in (i)) and market fragmentation (order routing as in (ii)). This endogeneity has important implications for empirical examinations of off-exchange trading, like wholesalers' execution of retail orders, single-dealer platforms, and various dark pools. In terms of policy implications, both effects run counter to the intention behind regulations like the OPR, which is to encourage liquidity supply competition across fragmented marketplaces by virtually integrating them via price referencing (see p. 461 of O'Hara and Ye, 2011).

We next discuss our model in detail and build intuition for the above insights. Section 2 considers the simplest form of regulatory price referencing: off-exchange prices equal on-exchange prices. Section 2.1 sets up the model. Liquidity demanders have hedging and speculation motives and choose optimal market orders that are exogenously fragmented on and off exchange (fragmentation is endogenous in Section 3). On-exchange trading is modeled as a standard uniform price auction (Kyle, 1989), where a finite number of market makers post supply schedules to clear the market order demand. They also simultaneously absorb off-exchange market orders at the same on-exchange price—price referencing. This setup likens the current practice of handling retail orders: brokers direct orders off exchange to market makers for execution at exchange-referenced prices.

Section 2.2 characterizes the equilibrium and highlights our first main insight: because of price referencing, market makers reduce their on-exchange supply and soften competition.[3] To see why,

[2] While the notion of best- execution is multi-dimensional (FINRA, 2014), in practice, regulations like Reg NMS only enforce the price dimension, as highlighted by Li, Ye, and Zheng, 2023, p.319.

[3] The term "price referencing" might seem reminiscent of practices like "price matching" and "price guarantees," which facilitate oligopolistic sellers' collusive behavior (see, e.g., Salop, 1986; Shapiro, 1989). However, in our setting there is no collusion: a market maker reduces on-exchange supply to raise her *own* off-exchange profits.

note that their on-exchange price impact, due to imperfect competition, also moves the referenced off-exchange price in the same direction. The more a market maker trades off exchange, the stronger is this *referenced* price impact, which might turn her on-exchange supply even negative (i.e., she would demand liquidity). Notably, this liquidity reduction is reinforced by other market makers: Supplying less liquidity is as if the market maker is (partially) leaving the exchange, and so the others operate in a less competitive environment and respond by also reducing their supplies. This illiquidity feedback spills over to off-exchange trading through price referencing.

Our model further predicts that illiquidity driven by price referencing exacerbates mispricing, raises investors' trading costs, and thus dampens their willingness to trade and the overall trading volume. Importantly, the mispricing generated by price referencing is *transitory* in nature. This differs from the predictions by "cream-skimming" theories (Easley, Kiefer, and O'Hara, 1996; Battalio and Holden, 2001), which would leave on-exchange orders more informed (toxic), thus wielding larger *permanent* price effects.

Our second main insight is that price referencing allows market makers to endogenously fragment investors' order flow off exchange. This is because, with the same execution price on and off exchange, investors become indifferent in terms of routing, and their brokers always fulfill the best-execution obligation, irrespective of where the trades take place. This endogenous fragmentation, driven by price referencing, contrasts the canonical liquidity-begets-liquidity view (Pagano, 1989) that "trading naturally gravitates towards a single marketplace because each trader benefits from the presence of others" (p. 254 of Foucault, Pagano, and Roell, 2013). Our model, therefore, contributes a novel fragmentation mechanism to the literature, which, as we review below, typically either assumes exogenous fragmentation or endogenizes it via investor heterogeneity (neither is required in our model).

We formalize this second main insight in Section 3. Specifically, we augment the baseline model with a pre-trading stage, during which market makers choose their off-exchange exposures at a cost, as motivated by real-world PFOF to brokers.[4] By endogenizing fragmentation, perhaps surprisingly, we show that the market becomes more *il*liquid if the market making sector grows larger, for example, over the long run when more of them enter. This is because more market makers in aggregate divert more orders off exchange, thus reducing their on-exchange competition and worsening overall illiquidity and trading efficiency.

Price referencing in our model has so far been stipulated by regulations like the OPR. To better understand its role, we study what happens, under endogenous fragmentation, when we remove this regulatory requirement. We find that in our stylized static model, price referencing can no longer be sustained, consolidated trading is restored, market illiquidity is alleviated, and a larger market making sector no longer hurts efficiency. We caveat, however, that our result does not suggest that regulations like the OPR should be removed. Instead, we argue that even in absence of regulation, market makers may still have the incentive and the means to commit to price referencing, for example, in more general settings with repeated interactions or reputation effects. That is, regulations like the OPR are sufficient to implement price referencing, but not necessary as market makers may also voluntarily adopt such practice.

Our analysis so far has not considered market makers' potential off-exchange competition, which is relevant in many platforms like dark limit order books and periodic (off-exchange) auctions. Notably, off-exchange competition also underlies the SEC's recent proposal of order-by-order (OBO) auction (SEC, 2022b). We therefore extend the model to allow for multiple market makers' off-exchange competition in Section 4. The analysis generalizes the mechanism from Section 2 and reveals a novel liquidity shift: Under regulations like the OPR, competing market makers reduce their on-exchange liquidity supply and move it off exchange. Notably, just like in the main model, such liquidity shift hurts the overall market quality when there is more off-exchange trading. We further study in Appendix B

[4] There is a fast-growing recent literature empirically examining *how* market makers acquire retail orders and *how* the economic rents are divided among market makers, brokers, and retail traders (e.g., via PFOF and price improvements). See Dyhrberg, Shkilko, and Werner (2023), Ernst et al. (2024), Huang et al. (2023), and Battalio and Jennings (2023), among many others. We complement this literature by contributing to the understanding of *why* market makers want to acquire retail orders off exchange in the first place.

an extension where market makers endogenously divert orders off exchange and demonstrate the robustness of the findings from Section 3.

The result from Section 4 supports the SEC's OBO auction proposal, albeit via a different mechanism. The SEC motivates their proposal from the observation that off-exchange retail orders are not as toxic as on-exchange (institutional) orders, and, hence, will enjoy lower trading costs once additional off-exchange competition is introduced. Instead, our mechanism does not rely on order heterogeneity. Notably, our model predicts that SEC's OBO auction would reduce not only off-exchange but also on-exchange trading costs.

Related literature and contribution

Our model first contributes to the theory literature studying market fragmentation by adding a new perspective of cross-venue price referencing driven by, for example, regulations like the OPR. Exogenous market fragmentation is a standing assumption in the literature, as in, among others, Chowdhry and Nanda (1991), Easley, Kiefer, and O'Hara (1996), Foucault and Menkveld (2008), van Kervel (2015), and Chen and Duffie (2021). To explain how fragmentation arises, the literature typically resorts to various forms of heterogeneity among investors, as in, for example, Battalio and Holden (2001), Babus and Parlatore (2021), Baldauf, Mollner, and Yueshen (2024). Absent such heterogeneity, trading naturally gravitates to a single dominant venue as liquidity begets liquidity (Pagano, 1989). Our model shows that price referencing enables market makers to fragment *homogeneous* investors' trading, and by doing so they can avoid the fierce on-exchange competition.[5]

Second, as the handling of retail orders serves as a featured application, our model contributes

[5] The topic of market fragmentation is broad and has been examined from many different angles: trading concentration (Pagano, 1989, Chowdhry and Nanda, 1991); market structure (Glosten, 1994, Hendershott and Mendelson, 2000, Parlour and Seppi, 2003); speed heterogeneity (Foucault and Menkveld, 2008, van Kervel, 2015); venue operators' competition (Colliard and Foucault, 2012; Pagnotta and Philippon, 2018; Chao, Yao, and Ye, 2019; Baldauf and Mollner, 2021; Cespa and Vives, 2022); price discovery across dark and lit markets (Ye, 2011; Zhu, 2014; Ye and Zhu, 2020); size discovery protocols (Duffie and Zhu, 2017; Duffie, Dworczak, and Zhu, 2017; Antill and Duffie, 2021); urgency and execution costs (Menkveld, Yueshen, and Zhu, 2017); price impact (Chen and Duffie, 2021); investor disagreement (Babus and Parlatore, 2021); and market makers' inventory costs (Daures-Lescourret and Moinas, 2022).

to the theories studying PFOF. The literature has identified various reasons why practices like PFOF arise. One is information. In Easley, Kiefer, and O'Hara (1996) and Battalio and Holden (2001), retail orders are less informed, hence less toxic to market makers, who naturally pay brokers to acquire them. Glode and Opp (2016) argue that PFOF can sustain intermediation chains that reduce information asymmetry. Yang and Zhu (2020) argue that retail flow via PFOF is informative about future institutional flows. These works build on the heterogeneous investors facing intermediaries, who hence need to screen them, while in our model investors are *intentionally* assumed to be homogeneous. Two is market makers' inventory frictions. In Baldauf, Mollner, and Yueshen (2024), retail orders inflict lower inventory cost on market makers, who then would like to acquire such orders via PFOF. Our model instead assumes away market makers' inventory costs. Three is competition. Parlour and Rajan (2003) show that because investors' limit orders compete against market makers' limit orders, only spreads wider than the competitive level can justify market makers' PFOF costs. The current paper, which is also about competition, does not study the role of such competing limit orders.

Third, our work also relates to the recent literature on order routing decisions in fragmented markets. Li, Ye, and Zheng (2023) examine exchanges' routing under Rule 610 and 611 (i.e., OPR) of Reg NMS and find that, because the rules only define best price in terms of a stock's gross price (before exchange fees or rebates), 62% exchange routings lead to worse net prices. Such fee considerations also underlie the findings from Battalio, Corwin, and Jennings (2016) and from Anand et al. (2021) that brokers' routing may not be in the best interest for their retail nor institutional clients, respectively. Degryse, Markovic, and Wuyts (2023) study the different impacts of the OPR and make-take fees on the liquidity demand and supply sides across two limit order markets. Unlike these papers, whose focus largely lies on the various forms of exogenous fees and costs, we turn instead to market makers' price referencing behavior.

The price referencing mechanism is reminiscent of price manipulation in cash-settled derivatives, where "manipulators" trade strategically to move the underlying price and raise their payoffs from the derivative contracts; see, e.g., Kumar and Seppi (1992); Dutt and Harris (2005); Griffin and

Shams (2018); Zhang (2022). While manipulation too is a form of strategic trading, we offer a novel perspective of how market fragmentation reinforces—and thus is endogenous to—the price distortion.

2 A model of price referencing

This section presents a model where market makers provide liquidity both on and off exchange. In particular, they execute off-exchange trades by referencing the prevailing on-exchange price, following regulations like the OPR in the U.S. and those governing dark pools under MiFID II in the EU. Section 2.1 sets up the model by fixing an exogenous amount of off-exchange trading (which we endogenize later in Section 3). Section 2.2 then characterizes the equilibrium, and Section 2.3–2.5 discuss various predictions and implications.

2.1 Model setup

Assets. There is one risky asset and one risk-free numéraire. After trading, each unit of the risky asset will pay v units of the numéraire, where v is random ex ante.

Trading venues. The asset is traded both on and off exchange. The on-exchange trading price is denoted by p (to be endogenously determined). The off-exchange price $\hat{p}$ refers to the exchange and is identical to p, i.e., $\hat{p} = p$ (hence also endogenous). All variables denoted with a "ˆ" overhead refer to the off-exchange venue and their counterparts without the "ˆ" to the on-exchange venue.

Liquidity demand. There is a continuum of (small) investors of measure μ (> 0), indexed by $i \in [0, \mu]$. They are risk-neutral, but if one holds q_i units of the risky asset after trading, she incurs a quadratic inventory cost of $\frac{\rho}{2}q_i^2$, where ρ (> 0) measures the severity of the holding cost. All investors receive a common endowment shock of u units of the risky asset and also observe the asset payoff v. The realizations $\{u, v\}$ are their private information. To hedge u and to profit from v, they demand liquidity to trade the risky asset.

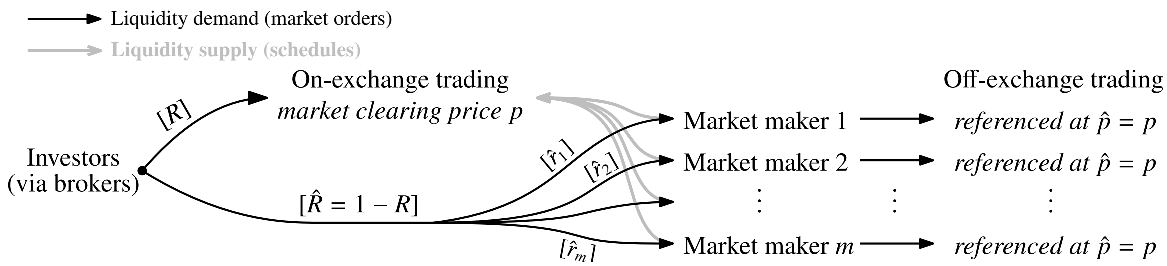

Figure 1: Order flow illustration. This figure illustrates how investors' order flows are handled under the setup of Section 2.1. The order flows are shown in black arrowed lines. Market makers' supplies are shown in gray arrowed lines. The on-exchange trading price p is determined via market clearing, while the off-exchange price $\hat{p}$ simply refers to p.

Liquidity supply. There are m (large) market makers, indexed by $j \in \{1, ..., m\}$, where m is an integer. They are risk-neutral, have no inventory costs, and do not observe the realizations of u or v. They supply liquidity, both on and off exchange, as follows.

Trading. There is one round of trading, in which

- each market maker j posts her supply schedule $x_j(p)$ on exchange; and
- each investor i submits a market order $z_i(v, u)$.

Each market order is independently routed to the exchange with probability R and off exchange to market maker j with probability $\hat{r}_j$, so that the total off-exchange probability is $\hat{R} := \sum_{j=1}^{m} \hat{r}_j = 1 - R$. The exogenous probabilities $\{\hat{r}_1, ..., \hat{r}_m\}$ are the respective market makers' "off-exchange exposures" (which we endogenize later in Section 3). See Figure 1 for illustration.

Market clearing. Market makers only choose their on-exchange supplies $\left\{x_j(p)\right\}_{j\in\{1,...,m\}}$, which determine the trading price p according to

$$\sum_{j=1}^{m} x_j(p) = R \int_0^{\mu} z_i(v, u)\mathrm{d}i. \tag{1}$$

They do not choose their off-exchange liquidity supply $\{\hat{x}_j\}$. Instead, all market orders routed off exchange to market maker j are executed at the on-exchange price p—price referencing (due, e.g., to regulations like the OPR). Then the off-exchange market clearing condition for market maker j is

$$\hat{x}_j = \hat{r}_j \int_0^{\mu} z_i(v,u)\mathrm{d}i =: \hat{z}_j, \tag{2}$$

where for notation simplicity we write $\hat{z}_j$ as the aggregate order routed to the market maker j.

Equilibrium. There are three sets of endogenous objects: (i) market makers' on-exchange liquidity supply schedules $\{x_j(p)\}_{j\in\{1,\ldots,m\}}$; (ii) investors' market orders $\{z_i(v,u)\}_{i\in[0,\mu]}$; and (iii) the trading price p. An equilibrium is such that, taking as given everyone else's strategy, each investor i chooses her market order z_i to maximize

$$\mathbb{E}\Big[(z_i+u)v - pz_i - \frac{\rho}{2}(z_i+u)^2 \,|\, v,u\Big]; \tag{3}$$

each market maker j chooses her on-exchange supply x_j to maximize

$$\mathbb{E}\big[(p-v)(x_j(p)+\hat{z}_j) \,|\, p\big]; \tag{4}$$

and, finally, the markets clear as in (1) and (2). In particular, each investor i is *small*, in that her market order z_i does *not* affect the price p, while each market maker j is *large*, in that her on-exchange supply x_j affects p via market clearing.

Parameters. The random variables v and u are independently normally distributed with respective means $\bar{v}$ and 0 and variances τ_v^{-1} and τ_u^{-1}. Sufficiently many market makers are needed:

$$m > 1 + \frac{1}{1-\phi}\frac{1}{R}, \text{ with } \phi := \frac{\tau_v^{-1}}{\tau_v^{-1}+\tau_u^{-1}\rho^2} \in (0,1), \tag{a}$$

which requires $R > 0$, i.e., on-exchange trading is non-zero, for otherwise the off-exchange trading has no price reference. (As will be shown shortly, the parameter ϕ is the signal to noise ratio of investor demand.) The condition also ensures that $m \geq 3$ (for m has to take integer values), which is a common restriction that guarantees sufficient competition seen, e.g., in Kyle (1989). Indeed, later in Section 3, after endogenizing market makers' off-exchange exposures $\{\hat{r}_j\}_{j\in\{1,\ldots,m\}}$ and hence also R, we will see that Condition (a) reduces to $m \geq 3$ (Condition (b)).

Remarks:

Remark 1 (Interpretations of $\hat{p} = p$). We offer three interpretations for the price referencing of $\hat{p} = p$. First, $\hat{p} = p$ can proxy how market makers (or wholesalers) set prices for the retail orders they purchase via PFOF from brokers. Second, $\hat{p} = p$ also proxies (midpoint) dark pools, which execute trades at prices referenced to on-exchange prices, effectively delegating price discovery to (lit) exchanges. Third, broker-dealers and single-dealer platforms who absorb client flows, e.g., in over-the-counter trading, also refer to on-exchange prices.

Remark 2 (Price improvement). In practice, market makers who operate as wholesalers interact with brokerage platforms to determine off-exchange order routing and price improvement (offered by wholesalers to traders); see, e.g., Schwarz et al. (2023) for empirical quantification. We abstract away from such interaction and simply set $\hat{p} = p$. Later, we use the example on page 14 to show that price improvement does not qualitatively affect our main results.

Remark 3 (Liquidity supply and demand). We model liquidity supply via large strategic market makers who have price impacts, as in Kyle (1989). We intentionally consider risk-neutral market makers (without inventory costs), so as to mute their incentives to share risks among themselves or to trade across marketplaces (on exchange vs. off exchange) to diversify inventory costs (Baldauf, Mollner, and Yueshen, 2024). The liquidity demand side is modeled after Vayanos and Wang (2012), but for simplicity we (i) assume quadratic inventory costs instead of constant absolute risk aversion (CARA) utility; and (ii) restrict the investors to using market orders. We focus on price-taking, small investors to fit into our key application of retail trading.[6]

Remark 4 (Homogeneous investors). The liquidity-demanding investors on and off exchange are intentionally assumed to be homogeneous. This way, we shut down the cream-skimming channel (Easley, Kiefer, and O'Hara, 1996; Battalio and Holden, 2001), through which market makers can directly offer liquidity to less informed retail orders via PFOF or to liquidity-driven institutional orders

[6] All results in this section hold if we i) assume CARA utility for the investors, ii) allow them to use price schedules, and/or iii) consider large investors with price impacts.

in dark pools. In Section 2.3 (page 18), we contrast our model predictions with those from the cream-skimming literature. Appendix A considers heterogeneous orders to demonstrate the robustness of our main insights.

Remark 5 (Market makers' off-exchange exposures). Following Remark 1, in the case of PFOF, a market maker j's off-exchange exposure $\hat{r}_j$ can be interpreted as the consequence of her (unmodeled) persistent arrangement with retail brokers (see, e.g., Huang et al., 2023 for evidence). In the case of dark pools, $\hat{r}_j$ can reflect the market maker's ex-ante choice of dark pool participation, e.g., the frequency and the quantity she posts orders there.

Remark 6 (Supply schedules). We let market makers use supply schedules $x_j(p)$ in their on-exchange competition. As in Kyle (1989), such a supply schedule game is a tractable device to model the oligopolistic market makers' imperfect competition.[7] We follow Rostek and Yoon (2021), Wittwer (2021), and Chen and Duffie (2021), and allow the on-exchange supply to only depend on p and is thus "disconnected" from $\hat{p}$. This is realistic, as the supply schedule is an approximation for a sequence of limit orders, which at time of submission cannot depend on $\hat{p}$. We do note that in the market maker optimization, by introducing price referencing $p = \hat{p}$ (absent in aforementioned papers), the on-exchange price perfectly reveals the off-exchange price.

2.2 Equilibrium characterization

We conjecture, and later verify, the following linear equilibrium trading strategies:

$$\begin{aligned} z_i &= \alpha_i + \beta_i \cdot (v - \bar{v} - \gamma_i u) \quad \text{for an investor } i; \\ x_j(p) &= a_j + b_j \cdot (p - \bar{v}) \quad \text{for a market maker } j. \end{aligned} \tag{5}$$

The coefficients $\{\alpha_i, \beta_i, \gamma_i\}$ and $\{a_j, b_j\}$ are to be determined. To simplify notation, for the investors we define the averages $\alpha = \frac{1}{\mu}\int_0^\mu \alpha_i \mathrm{d}i$, $\beta = \frac{1}{\mu}\int_0^\mu \beta_i \mathrm{d}i$, and $\gamma = \frac{1}{\mu}\int_0^\mu \gamma_i \mathrm{d}i$; and for the market makers we define the aggregates $A = \sum_{j=1}^m a_j$, $B = \sum_{j=1}^m b_j$, $A_{-j} = \sum_{j' \neq j} a_{j'}$, and $B_{-j} = \sum_{j' \neq j} b_{j'}$. The coefficients $\{b_j\}$,

[7] The significant market maker profits are suggestive of imperfect competition, see, e.g., Financial Times (2024).

hence also B, are key determinants of market liquidity: they measure how aggressively each market maker j supplies liquidity for a given price premium of $(p - \bar{v})$.

2.2.1 Investors' market order choices

Consider an arbitrary investor i. From her point of view, under the above conjecture and notation, the on-exchange market clearing condition (1) becomes

$$A + B \cdot (p - \bar{v}) = R\mu \cdot (\alpha + \beta \cdot (v - \bar{v} - \gamma u)) \iff p = \bar{v} - \frac{A}{B} + \frac{R\mu\alpha}{B} + \frac{R\mu\beta}{B}(v - \bar{v} - \gamma u). \tag{6}$$

Note that given their identical information $\{v, u\}$, the investors effectively know p. Maximizing the quadratic optimization problem (3) yields the optimal market order of $z_i = \frac{1}{\rho}(v - p) - u$. After substituting in p from (6), we see that z_i is indeed linear in v and u as conjectured in (5). The following lemma pins down the coefficients.

Lemma 1 (Optimal market order, given market makers' supplies). Fix market makers' on-exchange supplies as conjectured in (5). In equilibrium, every liquidity-demanding investor i indeed submits her market order z_i according to the linear form conjectured in (5), with coefficients

$$\alpha_i = \frac{A}{\rho B + R\mu}, \quad \beta_i = \frac{B}{\rho B + R\mu}, \quad \text{and} \quad \gamma_i = \rho. \tag{7}$$

Unsurprisingly, since the investors are homogeneous, they all use the same linear strategy. Hence, the averages $\alpha = \alpha_i$, $\beta = \beta_i$, and $\gamma = \gamma_i$.

Recall that the investors demand liquidity for two reasons, hedging and speculation, which jointly determine their trading motive. We summarize this motive in a single statistic

$$t := v - \bar{v} - \rho u, \tag{8}$$

which according to (6) can be inferred by the market maker through the market clearing price p. Then, the market order becomes $z_i = \alpha + \beta t$, where β measures her aggressiveness in demanding liquidity. Indeed, Lemma 1 confirms that β is higher precisely when there is more on-exchange liquidity supply (larger B).

2.2.2 A market makers' on-exchange supply

Consider a market maker j. She takes as given all others' conjectured strategies as given in (5). In particular, she knows that the investors trade on the combined motive t in the form of $z_i = \alpha + \beta t$, and, hence her off-exchange orders amount to $\hat{z}_j = \hat{r}_j \mu \cdot (\alpha + \beta t)$. Through the market clearing condition (1), her on-exchange supply x_j affects the price according to:

$$(9) \qquad x_j + A_{-j} + B_{-j} \cdot (p - \bar{v}) = R\mu \cdot (\alpha + \beta t) \iff p = \bar{v} + \frac{1}{B_{-j}}\big(R\mu \cdot (\alpha + \beta t) - A_{-j} - x_j\big).$$

Therefore, conditioning on p via her supply schedule, she equivalently observes t as defined in (8) and infers, by Bayes' theorem, that

$$(10) \qquad \mathbb{E}[v \mid p] = \mathbb{E}[v \mid t] = \bar{v} + \phi t,$$

where ϕ is defined in (a) and reflects the informativeness of t about v. Note also that

$$\mathbb{E}[\hat{z}_j \mid p] = \mathbb{E}[\hat{z}_j \mid t] = \hat{r}_j \mu \cdot (\alpha + \beta t) = \hat{z}_j.$$

That is, by conditioning on the on-exchange price p, the market maker perfectly infers her off-exchange orders $\hat{z}_j$, which can be considered as a function of t. Following (4), the market maker then solves

$$\max_{x_j} (p - \bar{v} - \phi t)\big(x_j + \hat{z}_j(t)\big),$$

knowing that her supply x_j impacts the price p via (9).

How does off-exchange exposure affect on-exchange supply? Notably, the market maker's off-exchange exposure $\hat{z}_j$ is executed at the same price p as on-exchange trades. Her first-order condition with respect to x_j yields

$$(11) \qquad -\frac{1}{B_{-j}}\big(x_j + \hat{z}_j(t)\big) + (p - \bar{v} - \phi t) = 0 \implies x_j = B_{-j} \cdot (p - (\bar{v} + \phi t)) - \hat{z}_j(t).$$

Note how the off-exchange exposure $\hat{z}_j(t)$ negatively affects the on-exchange supply x_j. To see why this is beneficial for the market maker, suppose the investors are buying (hence $\hat{z}_j(t) > 0$). Then by reducing x_j, the market maker pushes up the on-exchange price p and thus raises the proceeds from her off-exchange sale. Crucially, as we will see from the generalization below, the above negative effect only arises with price referencing.

Generalization: the distortionary effect of price referencing. We now allow the off-exchange price $\hat{p}$ to refer to the on-exchange price p via a general function $\hat{p}(p)$, which can reflect, for example, off-exchange price improvement (the example below) or off-exchange competition among various market makers (Section 4). Then the market maker's problem becomes

$$\max_{x_j}(p - \bar{v} - \phi t)x_j + (\hat{p}(p) - \bar{v} - \phi t)\hat{z}_j(t),$$

and the first-order condition with respect to x_j yields

$$\frac{\mathrm{d}p}{\mathrm{d}x_j}x_j + p - (\bar{v} + \phi t) + \frac{\mathrm{d}\hat{p}}{\mathrm{d}p}\frac{\mathrm{d}p}{\mathrm{d}x_j}\hat{z}_j(t) = 0 \implies x_j = -\left(\frac{\mathrm{d}p}{\mathrm{d}x_j}\right)^{-1}(p - (\bar{v} + \phi t)) - \frac{\mathrm{d}\hat{p}}{\mathrm{d}p}\hat{z}_j(t). \tag{12}$$

Therefore, the on-exchange supply x_j is always distorted by price referencing, as long as (i) $\frac{\mathrm{d}p}{\mathrm{d}x_j} \neq 0$—there is price impact and $\left(\frac{\mathrm{d}p}{\mathrm{d}x_j}\right)^{-1}$ is well-defined; (ii) $\frac{\mathrm{d}\hat{p}}{\mathrm{d}p} \neq 0$—there is price referencing; and (iii) $\hat{z}_j(t) \neq 0$—the market maker j has non-zero off-exchange exposure.[8]

Example (Price improvement)**.** An example of the generalization is that in current equity markets, virtually all off-exchange retail trading receives price improvement. Suppose market makers offer a price improvement of δ dollars (see Remark 2), yielding $\hat{p}(p) = p - \delta \mathrm{sign}[\hat{z}_j]$. Then $\frac{\mathrm{d}\hat{p}}{\mathrm{d}p} = 1$. Therefore, the distortion effect of price referencing implied by (12) remains unaffected even in the presence of price improvement δ. While outside the scope of this paper, δ could be endogenously determined together with wholesalers' competition, PFOF, and execution speed (Huang et al., 2023;

[8] The distortion term $-\frac{\mathrm{d}\hat{p}}{\mathrm{d}p}\hat{z}_j$ also arises with delta hedging. For example, suppose that a trader holds $\hat{z}_j$ units of a derivative whose payoff is $f(p)$. Then the distortion becomes $-\frac{\mathrm{d}f(p)}{\mathrm{d}p}\hat{z}_j$, which is the trader's delta-hedging of her derivative position $\hat{z}_j$ and $\frac{\mathrm{d}f(p)}{\mathrm{d}p}$ is the delta-hedging ratio. This mechanism also arises in the literature on cash settlement of derivatives, which shows that above distortion is an incentive to "manipulate" the underlying price p to achieve higher profits (see, e.g., Kumar and Seppi, 1992; Dutt and Harris, 2005; Zhang, 2022).

Dyhrberg, Shkilko, and Werner, 2023; Ernst et al., 2024).

Solving the strategy coefficients. The market maker's first-order condition (11) verifies the linear supply conjectured in (5). The coefficients are given by the following lemma.

Lemma 2 (Optimal on-exchange supply schedule, given investors' market orders). Fix investors market order strategies as conjectured in (5) and suppose $\beta > 0$. Then a market maker j's optimal demand schedule $x_j(p)$ is indeed in the linear form as conjectured in (5), with coefficients

$$a_j = \left(\frac{1}{m-1} - \hat{r}_j\right)\frac{(m-1)R-1}{(m-1)R+1}\mu\alpha \text{ and } b_j = \left(\frac{1}{m-1} - \hat{r}_j\right)\frac{(m-1)R-1}{(m-1)R+1}\frac{\mu\beta}{\phi}. \tag{13}$$

2.2.3 Equilibrium

Lemmas 1 and 2 have verified the conjectured linear strategies (5). The implied coefficient equations (7) and (13) form a linear equation system that uniquely determines the equilibrium.

Proposition 1 (Equilibrium with exogenous off-exchange trading). Define an illiquidity index

$$\zeta := 1 + \frac{1}{(m-1)R-1}, \tag{14}$$

which is bounded by $\frac{m-1}{m-2} \leq \zeta < \frac{1}{\phi}$. Then the coefficients in (5) are given by

$$a_j = 0,\ b_j = \left(\frac{1}{m-1} - \hat{r}_j\right)\frac{1-\zeta\phi}{2\zeta-1}\frac{\mu}{\phi\rho},\ \alpha_i = 0,\ \beta_i = \frac{1}{\rho}(1-\zeta\phi), \text{ and } \gamma_i = \rho. \tag{15}$$

Further, fixing $\{v, u\}$, hence also t as defined in (8), the trading price p is given by

$$p = \bar{v} + \zeta\phi t. \tag{16}$$

2.3 Equilibrium properties and predictions

Below we discuss some properties of the equilibrium.

Illiquidity due to market makers' imperfect competition: ζ**.** Compared to the informationally efficient price $\mathbb{E}[v|\,t\,] = \bar{v} + \phi t$, the equilibrium price (16) reacts excessively to the trading motive t due to the amplification by the illiquidity index $\zeta \geq \frac{m-1}{m-2} > 1$. Note that ζ closely relates to commonly used empirical illiquidity measures: The trades' permanent price impact is $\mathbb{E}[v|\,p\,] - \bar{v} = \phi t$, and they pay in total an effective spread of $p - \bar{v} = \zeta\phi t$. Therefore, ζ is the ratio of the effective spread over the permanent price impact.

Two parameters, R and m, drive the illiquidity ζ. Figure 2(a) plots ζ against the total off-exchange fraction $\hat{R} = 1 - R$ for three different levels of $m \in \{3, 5, 10\}$. First, consistent with the conventional wisdom, ζ decreases with m, i.e., liquidity is higher when more market makers compete. Second, novel in this paper, ζ increases in the level of off-exchange trading $\hat{R}$, approaching the upper bound of $\zeta = \frac{1}{\phi}$ for a sufficiently high $\hat{R}$. This is a key insight: As off-exchange prices reference on-exchange prices, market makers refrain from competing too fiercely on exchange, as such competition now additionally hurts their off-exchange profit.

Note also that illiquidity ζ does *not* depend on the absolute size μ of liquidity demanders. What matters is R, or the *relative* size of on- vs. off-exchange trading. All else being equal, a market with $\{R = 0.9, \mu = 1\}$ is still more liquid and efficient than $\{R = 0.8, \mu = 100\}$, even though in the latter the absolute amount of on-exchange demand $R\mu$, and also the overall trading volume, is much higher.

> **Prediction 1:** In stocks with higher *fractions* of price referencing off-exchange trading, on-exchange liquidity supply competition is lower. This can manifest in wider effective and realized spreads and in a higher ratio of effective spread over permanent price impact.

Existing evidence supports this prediction. Ernst and Spatt (2022) exploit variation in assignments of designated market makers (DMMs) in the U.S. options exchanges and show that market making firms who internalize more trades using their DMM-allocation also earn larger spreads.[9] Aramian and

[9] Options trading in the U.S. is on-exchange only. Nevertheless, options exchanges allow their appointed DMMs to "internalize" their acquired orders via DMM priority rules and price improvement mechanisms. See Ernst and Spatt (2022) and Bryzgalova, Pavlova, and Sikorkaya (2023a) for institutional details.

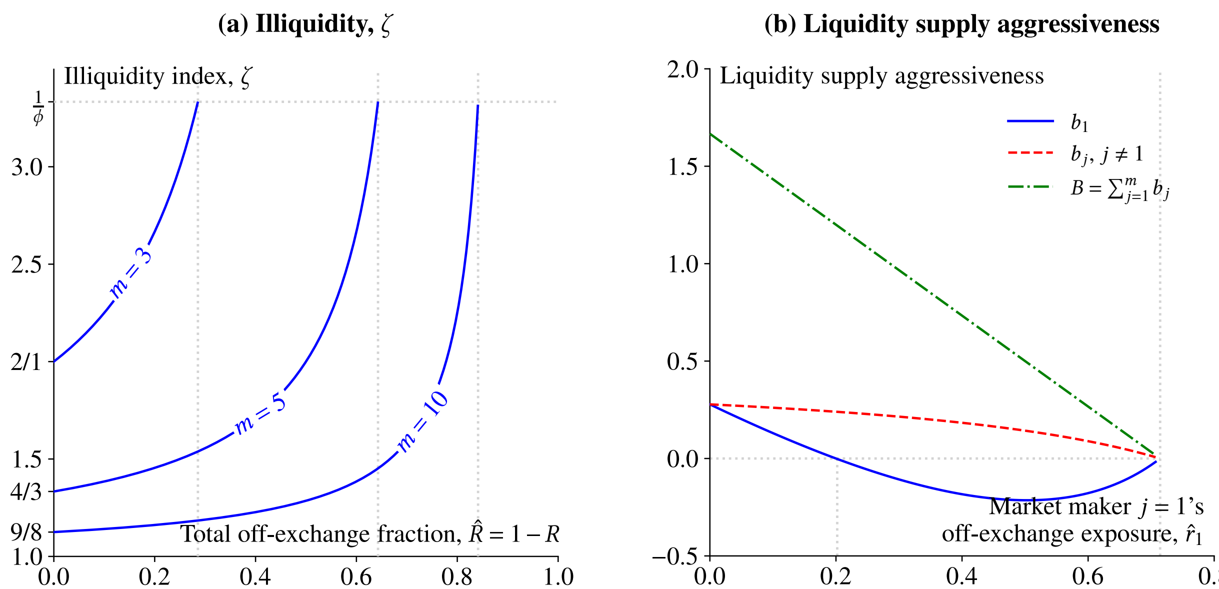

Figure 2: Off-exchange trading and market quality. This figure illustrates how off-exchange trading affects market quality in the trading equilibrium. Panel (a) shows how illiquidity ζ changes as more orders are traded off exchange (larger $\hat{R}$), for three different numbers of market makers, $m \in \{3, 5, 10\}$. Panel (b) shows how one market maker $j = 1$'s off-exchange exposure $\hat{r}_1$ affects various liquidity supply aggressiveness b_1, b_j ($j \neq 1$), and $B = \sum_{j=1}^{m} b_j$, assuming that all other market makers have no off-exchange exposures ($\hat{r}_j = 0, \forall j \neq 1$) and that $m = 6$. The other parameters are set at $\bar{v} = 0.0$, $\mu = \rho = 1.0$, $\tau_v = 0.1$, and $\phi = 0.3$.

Norden (2021) use Swedish data to show that on days where HFT execute more off-exchange volume at their single-dealer platforms, the spread on exchange increases.

Figure 2(a) also reveals that, intuitively, increasing the number of market makers reduces illiquidity. Therefore, the marginal effect of off-exchange trading on illiquidity becomes less severe in m: $\frac{\mathrm{d}^2\zeta}{\mathrm{d}\hat{R}\mathrm{d}m} < 0$. Put conversely,

> **Prediction 2:** The marginal impact of off-exchange trading on illiquidity is more negative for stocks with fewer market makers.

As larger stocks typically have more high-frequency liquidity providers (Biais and Foucault (2014), §II.3), this prediction is consistent with evidence that the impact of dark trading on liquidity is negative and more severe in small stocks (Degryse, de Jong, and van Kervel, 2015 and Foley and Putniņš, 2016).

Difference with cream-skimming models. The equilibrium price (16) can be decomposed into the efficient price $\mathbb{E}[v|p]$ plus a mispricing $(\zeta - 1)\phi t$, which is *transitory* in nature, because in the (unmodeled) long run, informed investors will trade against it. This transitory price component helps distinguish our mechanism from cream-skimming models, where wholesalers would like to attract less informed order flow off exchange, making the residual on-exchange order flow more toxic. Consequently, just like in Prediction 1, on-exchange trading also becomes less liquid (Easley, Kiefer, and O'Hara, 1996; Battalio and Holden, 2001; and Hu and Murphy, 2022). The key difference is that the illiquidity predicted by these theories arises from the worsening adverse selection, i.e., larger *permanent* price impacts, while our prediction on the illiquidity ζ builds on market makers' imperfect competition, i.e., larger *transitory* price impacts.

Market makers' liquidity supply aggressiveness $\{b_j\}$. We see from (15) that an increase in a market maker j' off-exchange exposure $\hat{r}_j$ reduces her liquidity supply aggressiveness b_j via two channels. First, directly, it lowers the first term $\frac{1}{m-1} - \hat{r}_j$. Intuitively, she reduces exchange supply due to the exacerbating distortion from her off-exchange exposure $\hat{z}_j = \hat{r}_j \int_0^\mu z_i(v, u)\mathrm{d}i$, as seen in (11). Notably, if her $\hat{r}_j$ is sufficiently large, exceeding $\frac{1}{m-1}$, the market maker no longer supplies liquidity, but, demands it instead from the other market makers. Indeed, with $m = 6$ market makers in Figure 2(b), b_1 of market maker $j = 1$ decreases with her $\hat{r}_1$, and it drops below zero after $\hat{r}_1 > \frac{1}{m-1} = 0.2$.[10]

> **Prediction 3:** Market makers with larger off-exchange exposures are more likely to demand liquidity on-exchange. This can manifest in their submitting more market(able) orders or posting more aggressive limit orders in the same direction as the off-exchange liquidity demand.

This result is consistent with the empirical evidence in Aramian and Norden (2021), who show that a one-standard deviation increase in HFT off-exchange volume at their single-dealer platform is associated with an 18% reduction of the on-exchange passive liquidity supply (i.e., the passive volume reduces from the sample mean of 50% to 41%).

[10] Of course, if a market maker j is consuming liquidity ($b_j < 0$), then the other market makers must be providing liquidity, i.e., $\sum_{l \neq j} b_l = B - b_j > 0$. We show in the proof of Proposition 1 that $B - b_j > 0$ always holds in equilibrium, which in fact ensures the market makers' second-order conditions.

Second, $\hat{r}_j$ reduces b_j indirectly via a higher illiquidity ζ. (More precisely, the second term in b_j, $\frac{1-\zeta\phi}{2\zeta-1}$, is monotone decreasing in ζ, which in turn rises with $\hat{r}_j$.) This is because in a more illiquid market, market makers face larger price impacts and therefore supply less liquidity. Importantly, this channel spills over to all other market makers: an increase in $\hat{r}_1$ reduces every market maker $j \neq 1$'s supply b_j via this second channel.

Prediction 4: An increase in one market makers' off-exchange exposure reduces not only her on-exchange liquidity supply but also that of all other market makers.

Figure 2(b) illustrates this "spill over" effect via b_j ($j \neq 1$, dashed line) and via $B = \sum_{j=1}^{m} b_j$ (dot-dashed line). As market maker 1's $\hat{r}_1$ continues to increase, both these curves monotonically decrease. Eventually, b_1, b_j ($j \neq 1$), and B all converge to zero, because the illiquidity ζ becomes too severe and the investors no longer trade ($\beta \to 0$).

2.4 Gains from trade

Following (3), an investor i's trading gain π_i^{I} can be defined as the difference of her certainty equivalents with vs. without trading: $\mathbb{E}\left[(z_i + u)v - pz_i - \frac{\rho}{2}(z_i + u)^2\right] = \mathbb{E}\left[\pi_i^{\mathrm{I}} + uv - \frac{\rho}{2}u^2\right]$, yielding

$$\pi_i^{\mathrm{I}} = \frac{(1-\zeta\phi)^2}{2\rho\tau_v\phi}. \tag{17}$$

A market maker j's expected profit can be directly computed as

$$\pi_j^{\mathrm{M}} := \mathbb{E}\left[(p - v)(x_j + \hat{x}_j)\right] = \left(1 - \hat{R}_{-j}\right)\frac{(\zeta-1)^2(1-\zeta\phi)}{2\zeta - 1}\frac{\mu}{\rho\tau_v}, \tag{18}$$

where $\hat{R}_{-j} := \sum_{j' \neq j} \hat{r}_{j'}$ is all other market makers' aggregate off-exchange exposure. Combining the above, we obtain:

Corollary 1 (Gains from trade and market illiquidity). Define the aggregate gains from trade w as the sum of all investors' gains from trade and all market makers' expected profits. Then

$$w = \int_0^{\mu} \pi_i^{\mathrm{I}} \mathrm{d}i + \sum_j \pi_j^{\mathrm{M}} = (1-\zeta\phi)(1 - \phi + (\zeta - 1)\phi)\frac{\mu}{2\rho\tau_v\phi}, \tag{19}$$

and it is monotonically increasing in m and in R.

Only investors' hedging trades are socially beneficial, transferring endowment shocks from investors with inventory costs to market makers who do not suffer from such costs. Investors' speculative trades on the private information v, on the other hand, lead only to zero-sum transfers. As ζ increases, investors hedge less: $\frac{\mathrm{d}z_i}{\mathrm{d}u} = \beta_i \gamma_i = 1 - \zeta\phi$ decreases. Therefore, overall gains from trade w increases in m and in R because ζ decreases in m and in R (Proposition 1).

2.5 Order routing

The investors' trading gain (17) reveals that every one of them is better off if more of them (require their brokers to) route orders to the exchange:

$$\frac{\mathrm{d}\pi_i^{\mathrm{I}}}{\mathrm{d}R} = \frac{\mathrm{d}\pi_i^{\mathrm{I}}}{\mathrm{d}\zeta}\frac{\mathrm{d}\zeta}{\mathrm{d}R} = \frac{1-\zeta\phi}{\rho\tau_v}\frac{m-1}{((m-1)R-1)^2} > 0.$$

However, individually they have no incentive to do so: each order will be executed at the same price $p = \hat{p}$, regardless of routing. That is, price referencing dissolves individual investors' (or their brokers') incentive to specify order routing.[11]

Investors' indifference in routing further facilitates market makers' fragmenting their orders off exchange: Regardless of how illiquid the market becomes, any investor remains indifferent between trading on or off exchange. We explore market makers' endogenous choices of their off-exchange exposures $\{\hat{r}_j\}$ in the next section. In particular, in Section 3.3, we show that price referencing is indeed the source of investors' indifference: once removed, they strictly prefer routing to the exchange.

[11] It would require significant coordination among dispersed (retail) investors to collectively direct their orders to the exchange. Further, such coordination is not always in the interest of investors (who might receive off-exchange price improvement) nor brokers (who would lose PFOF revenue).

3 Market makers' endogenous off-exchange exposures

Section 2 has analyzed the trading equilibrium, taking as given market makers' off-exchange exposures $\{\hat{r}_1, ..., \hat{r}_m\}$. We endogenize these exposures in this section by extending the model with a pre-trading period, in which each market maker j, taking as given all others' $\{\hat{r}_{j'}\}_{j' \neq j}$, chooses her $\hat{r}_j$ to maximize her expected profit π_j^{M} as given by (18), subject to a symmetric cost:

$$\max_{\hat{r}_j} \pi_j^{\mathrm{M}} - \kappa \hat{r}_j \mu, \tag{20}$$

where the parameter κ (> 0) measures how costly it is for one to obtain off-exchange exposure. Since $R = 1 - \sum_{j=1}^{m} \hat{r}_j$ is now endogenous, instead of Condition (a), we assume

$$m \geq 3 \tag{b}$$

to ensure sufficient competition among market makers.

Remark 7 (Costly acquisition of off-exchange orders). Market makers have certain discretion over the degree of off-exchange trading, which likely comes at a cost. As discussed in Remark 5, such cost can reflect, for example, the market maker's payment to brokers for retail orders or her ex-ante dark pool participation cost. We consider a linear cost $\kappa \hat{r}_j \mu$ only for tractability and simplicity. In unreported analysis, we show that our results remain robust if the cost is instead convex.

3.1 Equilibrium characterization

Given the homogeneous market makers, it is natural to focus on their symmetric strategies. The following proposition states the equilibrium.

Proposition 2 (Equilibrium with endogenous off-exchange exposures). Suppose the cost κ is no larger than a threshold $\bar{\kappa}(m)$ as defined by (C.11) in the proof. Then there exists a unique symmetric-strategy equilibrium, in which each market maker j acquires the same off-exchange exposure $\hat{r}_j = \hat{r} \in \left(0, \frac{1}{m}\right)$.

Below we discuss heuristically the construction of the symmetric-strategy equilibrium, deferring the details to the proof. Consider first the tradeoff market maker j faces when choosing her optimal $\hat{r}_j$, fixing all others' aggregate $\hat{R}_{-j} = \sum_{j' \neq j} \hat{r}_j$. This tradeoff is illustrated in Figure 3(a). Her marginal trading revenue $\frac{\partial \pi_j^{\mathrm{M}}}{\partial \hat{r}_j}$ (dashed line) is hump-shaped in $\hat{r}_j$, reflecting two countervailing forces: On the one hand, as seen in Proposition 1, more off-exchange trading softens on-exchange competition, so that the market maker can sustain a higher trading costs for—and can profit from—investors. On the other hand, the resulting illiquidity ζ reduces investors' demand β and overall trading volume, thus lowering the market maker's revenue accordingly. Her marginal cost of acquiring off-exchange exposure is κ, fixed at $\kappa = 1$ in this numerical illustration (solid flat line). Her optimal level of $\hat{r}_j$ is obtained when the marginal revenue equates the marginal cost, as shown by the circle. (The other intersection fails the second-order condition: increasing $\hat{r}_j$ raises marginal profit.)

In Panel (b), we repeat the above by identifying the best response $\hat{r}_j$ for various levels of $\hat{R}_{-j}$ (on the horizontal axis) and obtain the best response curve, the (blue) bold-solid line. For example, the circle at $\hat{R}_{-j} = 0.2$ matches the one in Panel (a). Also shown in Panel (b) are the contours of the market maker's net profit $\pi_j^{\mathrm{M}} - \kappa \hat{r}_j \mu$. Indeed, it can be seen that the best response curve traces the maximum net profit for every level of $\hat{R}_{-j}$. To obtain the symmetric-strategy equilibrium, we look for the intersection of the best response curve and the dashed line of $\hat{r}_j = \hat{R}_{-j}/(m-1)$, as implied by symmetry. The solid square indicates such an equilibrium, which is unique given the monotonically decreasing best-response curve.[12]

3.2 Equilibrium properties

We next study the market quality implied by the above equilibrium. In particular, we consider the effects of the acquisition cost κ and the number of market makers m.

[12] We analytically prove that the best-response curve is weakly decreasing in the proof of Proposition 2. The main economic force can be seen from the expression (18) for π_j^{M}: As $\hat{R}_{-j}$ increases, the smaller $(1 - \hat{R}_{-j})$ directly scales down π_j^{M}, hence also the marginal trading revenue proportionally. Given the constant marginal cost κ, the market maker therefore responds by reducing her $\hat{r}_j$.

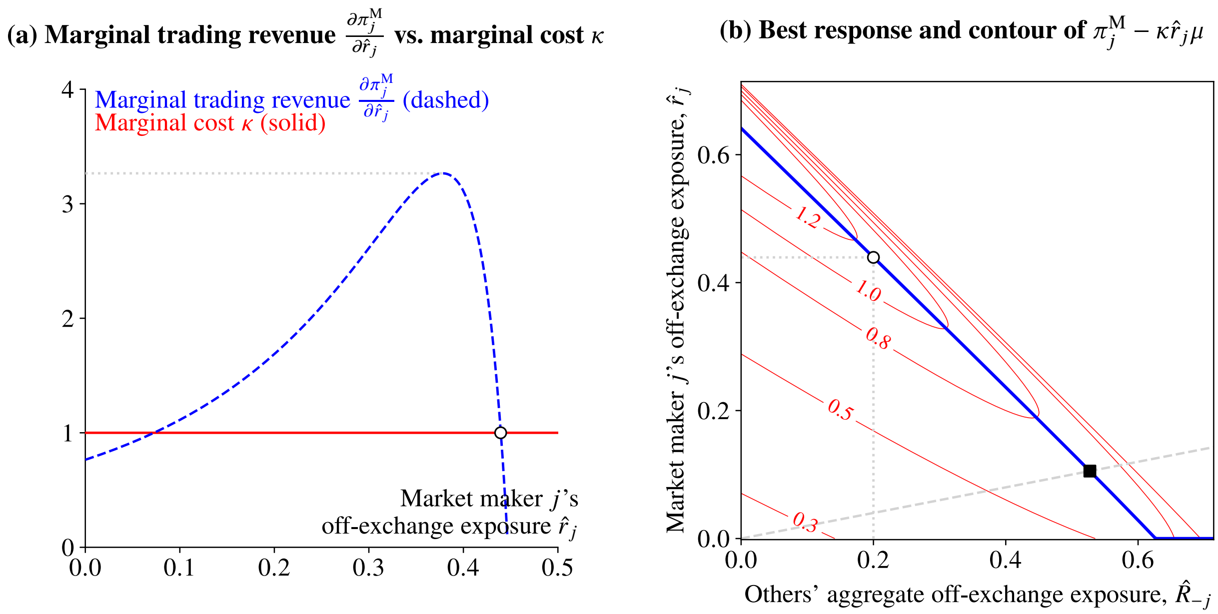

Figure 3: A market maker's tradeoff in choosing her off-exchange exposure. Consider a market maker j. Panel (a) fixes the others' aggregate exposure $\hat{R}_{-j} = \sum_{j' \neq j} \hat{r}_{j'} = 0.2$ and plots the market maker j's tradeoff between her marginal trading revenue (dashed) and her marginal cost $\kappa = 1$ (solid). The circle indicates the market maker's optimal choice. Panel (b) illustrates in a contour graph how the market maker's net profit $\pi_j^{\mathrm{M}} - \kappa \hat{r}_j \mu$ responds to her own off-exchange exposure $\hat{r}_j$ and to other' aggregate $\hat{R}_{-j}$. The blue bold-solid line indicates the market maker's $\hat{r}_j$ as a best response to $\hat{R}_{-j}$, with the circle mirroring the circle in Panel (a), where $\hat{R}_{-j} = 0.2$. The dashed line indicates the symmetry line of $\hat{r}_j = \hat{R}_{-j}/(m-1)$, and the black square indicates the symmetric equilibrium. The other parameters are set at $\bar{v} = 0.0$, $\mu = \rho = 1.0$, $\tau_v = 0.1$, $\phi = 0.3$, and $m = 6$.

Corollary 2 (Effects of off-exchange exposure cost). In the symmetric-strategy equilibrium, an increase in the off-exchange exposure cost κ raises on-exchange trading R, lowers illiquidity ζ, and raises the aggregate gains from trade w. That is, $\frac{\mathrm{d}R}{\mathrm{d}\kappa} \geq 0$, $\frac{\mathrm{d}\zeta}{\mathrm{d}\kappa} \leq 0$, and $\frac{\mathrm{d}w}{\mathrm{d}\kappa} \geq 0$.

The effects are illustrated in Figure 4(a). Intuitively, a higher cost directly reduces market makers' incentive to acquire off-exchange exposures, thus leaving more orders on exchange, i.e., a higher $R = 1 - \sum_{j=1}^{m} \hat{r}_j$, as shown in the rising solid line. Following (14) and (19), in turn, illiquidity ζ lowers, as seen in the decreasing dashed line.

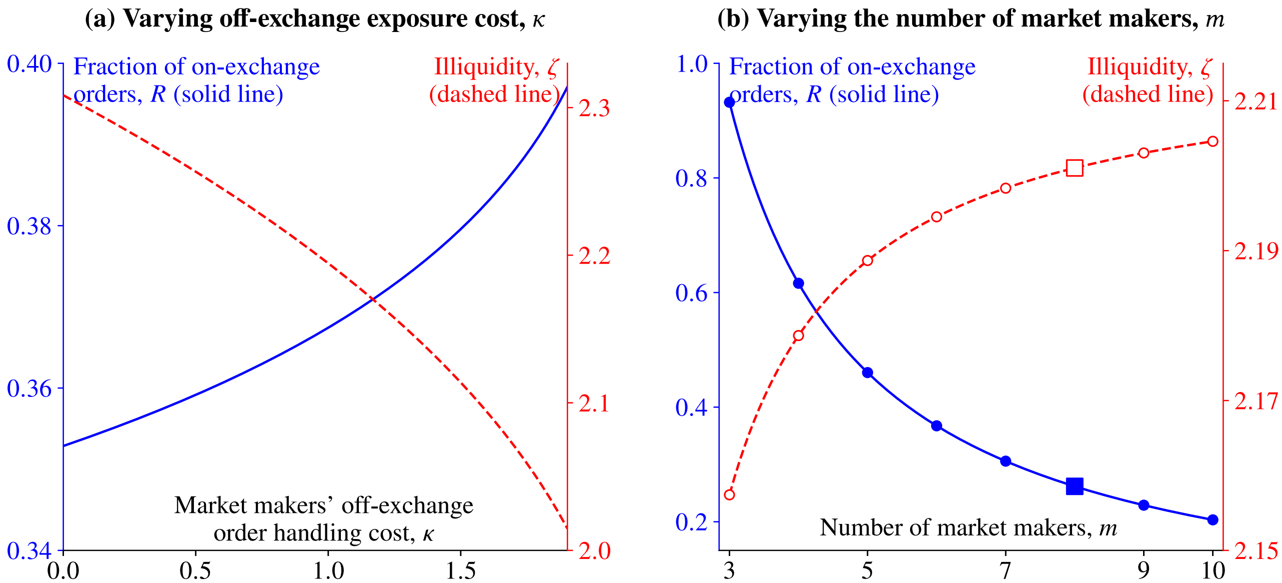

Figure 4: Market quality under market makers' endogenous off-exchange exposures. This figure shows how the fraction of on-exchange orders R (solid line, left axis) and illiquidity ζ (dashed line, right axis) vary with market makers' off-exchange exposure cost κ in Panel (a) and with the number of market makers m in Panel (b). The squares in Panel (b) correspond to the free-entry number of market makers in the long-run (Corollary 4). In Panel (a), $m = 6$; and in Panel (b), $\kappa = 1.0$ and the entry cost is $\kappa_e = 0.4$. The other parameters are set at $\bar{v} = 0.0$, $\mu = \rho = 1.0$, $\tau_v = 0.1$, and $\phi = 0.3$.

Corollary 3 (Effects of more market makers). In the symmetric-strategy equilibrium, an increase in the number of market makers m reduces on-exchange trading R, raises illiquidity ζ, and lowers the aggregate gains from trade w. That is, assuming differentiability in m, $\frac{dR}{dm} \leq 0$, $\frac{d\zeta}{dm} \geq 0$, and $\frac{dw}{dm} \leq 0$.

Figure 4(b) illustrates these results. As the number of market makers increases from $m = 3$ to $m = 10$, rather intuitively, they acquire more orders off exchange, and the on-exchange trading $R = 1 - \sum_{j=1}^{m} \hat{r}_j$ (solid line) decreases. Less intuitive is the rising ζ (dashed line): more market makers *hurt* liquidity. To see why, recall from (14) that $\zeta = 1 + \frac{1}{(m-1)R-1}$, which decreases in both m and R. Hence, an increase in m generates two effects:

$$\frac{d\zeta}{dm} = \frac{\partial \zeta}{\partial m} + \frac{\partial \zeta}{\partial R}\frac{dR}{dm}.$$

First, there is the direct effect, $\frac{\partial \zeta}{\partial m} < 0$: more market makers compete more and provide more liquidity.

Second, there is a novel indirect effect, via the reduced on-exchange trading (lower R): $\frac{\partial \zeta}{\partial R}\frac{\mathrm{d}R}{\mathrm{d}m} > 0$.[13] The proof shows that the indirect negative effect dominates in equilibrium, thus giving rise to the seemingly surprising finding that having more market makers worsens liquidity. To emphasize, this second effect arises precisely because of price referencing, under which market makers want to gain off-exchange exposure to soften supply competition.

The number of market makers m and the off-exchange exposure cost κ affect the aggregate gains from trade w only via illiquidity ζ; see (19). In particular, higher ζ worsens w. Therefore, the overall trading efficiency deteriorates when acquisition cost decreases ($\frac{\mathrm{d}w}{\mathrm{d}\kappa} > 0$) and when more market makers enter ($\frac{\mathrm{d}w}{\mathrm{d}m} < 0$).

Finally, we note that the above negative effect of more market makers carries through in the long run, when market makers can endogenously enter at a cost:

> **Corollary 4 (Entry of market makers in the long run.).** Suppose that there are $m_\circ$ (≥ 3) incumbent market makers initially, with $\kappa < \bar{\kappa}(m_\circ)$ (defined in (C.11)), and that the entry cost is $\kappa_{\mathrm{e}} > 0$. Then there exists an equilibrium with m^* market makers, where $m^* \geq m_\circ$.

Following Corollary 3, as more market makers enter, fewer orders remain on exchange (lower R), market illiquidity exacerbates (higher ζ), and the total gains from trade reduce (lower w). Notably, a high level of illiquidity persists, suggesting long-run market inefficiency. As an example, the long-run equilibrium levels of R and ζ are shown in the squares in Figure 4(b) (assuming $m_\circ = 3$). Our analysis therefore highlights the importance of the proper regulation of market makers' price referencing and their acquisition of off-exchange order exposures (e.g., from PFOF arrangements with brokers).

[13] The countervailing direct and indirect effects can be seen via the numerical illustration in Figure 2(a). For example, when all orders are on-exchange ($\hat{R} = 0$) and $m = 3$, the illiquidity is $\zeta = \frac{3-1}{3-2} = 2$. When m increases to $m = 5$, illiquidity drops to $\zeta = \frac{5-1}{5-2} = 4/3$—this is the direct effect. However, this liquidity gain is completely undone if these two additional market makers also raise off-exchange trading to $\hat{R} = 0.5$—this is the indirect effect.

3.3 The role of the price referencing regulation

In this subsection, we investigate the role of the price referencing regulations like the OPR. Specifically, we remove the requirement that off-exchange prices must equate on-exchange prices and examine what happens in our model laboratory. Note that, irrespective of the regulation, market makers can always refer to on-exchange prices when executing off-exchange orders[14]—but do they still have incentive to? How is market fragmentation affected in equilibrium? Do investors benefit? To address these questions, we make two changes to the previous model setup:

- It is no longer required that $\hat{p} = p$. Instead, each market maker j can now choose the price $\hat{p}_{ij}$ when executing the i-th off-exchange order.
- Accordingly, investors are no longer necessarily indifferent between trading on and off exchange—each of them (or their brokers) now may want to direct her order to the venue with the best (expected) price. Hence, we let each investor i choose, at the time of submitting her order, the probability $\hat{r}_{ij}$ to route her order to any market maker j (or to the exchange with probability $R_i := 1 - \sum_j \hat{r}_{ij}$).

The equilibrium, therefore, is composed of the following objects: (i) each market maker j's on-exchange liquidity supply schedule $x_j(p)$; (ii) each market maker j's off-exchange pricing rule $\hat{p}_{ij}$ for each order i routed to her; (iii) each investor i's market order size $z_i(v, u)$; (iv) each investor i's order routing probability $\hat{r}_{ij}$ to market maker j; and (v) the on-exchange market clearing price p. As before, we restrict our attention to linear on-exchange supply $x_j(p)$ and linear demand $z_i(v, u)$. Below we first state the equilibrium and then discuss the intuition and implications.

Proposition 3 (Equilibrium without regulatory price referencing). Absent regulatory requirements, there exists a class of equilibria, in which

[14] Price referencing is possible as long as there is (timely) dissemination of exchange prices. For example, between 1870–1915, so-called bucket shops effectively facilitated off-exchange trading at on-exchange prices thanks to the widespread adoption of low-cost telegraph called "ticker" (Hochfelder, 2006). In modern times, exchange prices have been made available via "securities information processors," or SIPs, since the enactment of Section 11A in the 1975 amendments to the Securities Exchange Act.

- for every order i routed to her, market maker j charges a worse off-exchange price $\hat{p}_{ij}$ than the on-exchange price p: $(\hat{p}_{ij} - p)z_i > 0$;
- all orders are routed on exchange, i.e., $\hat{r}_{ij} = 0$ and $R_i = 1$ for all $i \in [0, \mu]$;
- the on-exchange supply $x_j(p)$, the demand $z_i(v, u)$, and the on-exchange market clearing price p are as given in Proposition 1 with illiquidity $\zeta = 1 + \frac{1}{m-2}$ and $\hat{r}_j = 0$ for all $j \in \{1, ..., m\}$.

Intuition. Absent price referencing, when processing a market(able) order z_i, market maker j will have an incentive to charge an infinite price $\hat{p}_{ij}$ to maximize the expected trading profit. This means that investors would incur a prohibitive off-exchange trading cost and will thus only route to the exchange, resulting in $\hat{r}_{ij} = 0$ for all i and all j. As all orders are on-exchange, the exact pricing rule $\hat{p}_{ij}$ becomes irrelevant. Equilibria multiplicity arises, as $\hat{p}_{ij}$ only needs to be worse than p, i.e., $(\hat{p}_{ij} - p)z_i > 0$, to be incentive-compatible with investors' on-exchange routing. But the resulting market quality is always the same: trading consolidates on exchange, and the equilibrium supply, demand, and on-exchange price follow the special case of $R = 1$ in Proposition 1.

Alternative commitment devices. It follows that in the current static model, absent regulations like the OPR, no market maker can commit to offering an equal or better off-exchange price. However, in a repeated game or in a setting with reputation costs, such commitment could be possible; see Bryzgalova, Pavlova, and Sikorkaya (2023b) for a recent example. This means that the regulations suffice, but may not be necessary, for the market makers to implement price referencing. For example, Virtu (2021) states that "[w]holesalers provided over \$3.6B" and "Virtu alone provided over \$3B in Real Price Improvement to retail investors in 2020," a claim that puts Virtu's reputation at stake and could serve as a commitment device to implement price referencing and facilitate off-exchange trading.

In the remainder of this section, we discuss further implications of price referencing, without specifying whether it is driven by regulation or other means.

Implication: without price referencing, the effect of market maker competition is restored. Previously, under regulatory price referencing, Corollary 3 shows that when a new market maker

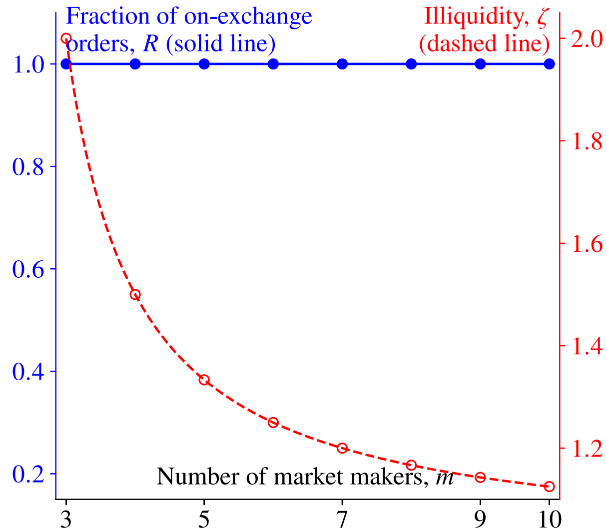

Figure 5: Market illiquidity when there is no price referencing. This figure shows the effects of more market makers (m on the horizontal axis) on the on-exchange market share R (the inverse of market fragmentation, solid line on the left axis), and on illiquidity (ζ, the dashed line, on the right axis), when there is no price referencing. The parameters are set to be the same as in Figure 4(b).

enters (increasing m), market illiquidity exacerbates (larger ζ). Corollary 4 further shows that this negative effect persists in the long run. This is because the new entrant market maker is able to acquire more orders off exchange, lowering R and the on-exchange competition. Proposition 3 shows that without price referencing, we revert to the standard result that illiquidity $\zeta = 1 + 1/(m-2)$ is *decreasing* in m through market makers' competition. Figure 5 illustrates this pattern and contrasts Figure 4(b). Following Corollary 1, the aggregate gains from trade w then also increases with m.

Implication: Who benefits from price referencing? We can compute the investors' and the market makers' expected gains from trade using expressions from (17) and (18) (by setting $\zeta = 1 + \frac{1}{m-2}$ and $\hat{R}_{-j} = 0$). Comparing them to the equilibrium with regulatory price referencing (Proposition 2), we find:

Corollary 5 (Gains from trade and price referencing). The investors always strictly prefer *no* price referencing. The market makers strictly prefer price referencing if and only if the market makers' cost κ of acquiring off-exchange exposure is sufficiently low. The total gains from trade w is always higher without price referencing.

Intuitively, without price referencing (Proposition 3), trading is consolidated, illiquidity reaches the lowest level of $\zeta = 1 + \frac{1}{m-2}$, and the investors incur the lowest trading costs. While this also means that market makers' trading proceeds are the lowest, the absence of price referencing also waives them from the cost $\kappa\mu\hat{r}_j$ to acquire off-exchange orders. Therefore, they prefer price referencing if and only if κ is sufficiently low. Finally, Corollary 1 has shown that the total gains from trade w is monotone increasing in R, and it follows that the highest $R = 1$ absent of price referencing maximizes w.

4 Off-exchange competition

In Sections 2 and 3, there is no competition for off-exchange orders as the price $\hat{p}$ is fixed to the on-exchange price p. A natural question is whether our results extend to other forms of off-exchange trading where $\hat{p}$ refers to p but is also determined via competition. Examples include dark limit order books, periodic auctions, and the recently proposed order-by-order (OBO) auctions under the Order Competition Rule 615 by SEC (2022b). In this extension we explore the robustness of the main results and compare the two settings, with vs. without off-exchange competition.

4.1 Model setup

We now assume a single off-exchange trading venue that operates identically to the exchange. We maintain the regulatory constraint that the off-exchange price must be equal or better than the on-exchange price.

Liquidity supply. As before, there are m market makers, who are all risk-neutral, have no inventory costs, and do not observe investors' private u or v. We let n (out of m) of them be "cross market makers," who can supply liquidity both on and off exchange, while the rest $\ell := m - n$ are "local," who can supply liquidity only on exchange. We allow $\ell \geq 0$ but require sufficient cross market makers:

(c) $$n > 1 + \frac{1}{1-\phi},$$

where ϕ is the same as defined in (a). (In Appendix B.1, for completeness, we also characterize the equilibrium with additional $\hat{\ell}$ off-exchange local market makers.)

Trading. There is one round of trading, in which

- each j of the ℓ on-exchange local market makers posts her supply schedule $y_j(p)$;
- each j of the n cross market makers posts both her supply schedules $x_j(p)$ and $\hat{x}_j(\hat{p})$; and
- each investor i submits a market order $z_i(v, u)$.

As before, each market order is independently routed on exchange with probability R and off exchange with probability $\hat{R} = 1 - R$. These probabilities are exogenous in this section and later endogenized in Appendix B.2. Figure 6 illustrates the game. Compared to Figure 1, the key differences are thus i) the local and cross market makers, and ii) the cross market makers compete in the off-exchange auction, rather than internalizing them. The trading prices, $\{p, \hat{p}\}$, are determined via market clearing:

(21) $$\sum_{j=1}^{n} x_j(p) + \sum_{j=1}^{\ell} y_j(p) = Z \text{ and } \sum_{j=1}^{n} \hat{x}_j(\hat{p}) = \hat{Z},$$

where for notation simplicity we write $Z := R\int_0^\mu z_i(v,u)\mathrm{d}i$ and $\hat{Z} := \hat{R}\int_0^\mu z_i(v,u)\mathrm{d}i$ as the aggregate order flows on and off exchange, respectively.

The price constraint. The off-exchange trading price $\hat{p}$ must weakly improve upon the on-exchange price p and is modeled as an inequality constraint of

(22) $$(\hat{p} - p)\hat{Z} \leq 0,$$

which the cross market makers must honor when choosing their supplies (see Remark 10 below).

Equilibrium. There are three sets of endogenous objects: (i) market makers' liquidity supply schedules $\{y_j(p)\}_{j\in\{1,\ldots,\ell\}}$ and $\{x_j(p), \hat{x}_j(\hat{p})\}_{j\in\{1,\ldots,n\}}$; (ii) investors' market orders $\{z_i(v,u)\}_{i\in[0,\mu]}$; and (iii) the trading prices $\{p, \hat{p}\}$. An equilibrium is such that, taking as given everyone else's strategy,

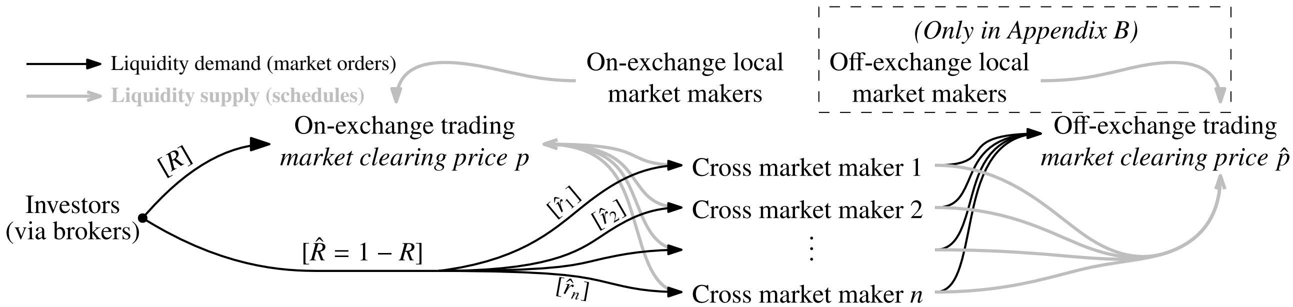

Figure 6: Order flow illustration, with off-exchange competition. This figure illustrates how investors' market orders are handled under the setup of Section 4.1. Investors' order flows are shown in black arrowed lines. Various types of market makers' supplies are shown in gray arrowed lines. For simplicity, in the main model we assume away off-exchange local market makers and consider the general case later in the Appendix B. Both the on-exchange trading price p and the off-exchange price $\hat{p}$ are determined via market clearing.

the agents solve their respective problems:

$$\max_{z_i} \mathbb{E}\Big[(z_i + u)v - \Big(Rp + \hat{R}\hat{p}\Big)z_i - \frac{\rho}{2}(z_i + u)^2 | \, v, u\Big] \text{ by an investor } i;$$

$$\max_{y_j(p)} \mathbb{E}\big[(p - v)y_j(p) | \, p\big] \text{ by an on-exchange local market maker } j; \tag{23}$$

$$\max_{x_j(p), \hat{x}_j(\hat{p})} \mathbb{E}\big[(p - v)x_j(p) | \, p\big] + \mathbb{E}\big[(\hat{p} - v)\hat{x}_j(\hat{p}) | \, \hat{p}\big] \text{ s.t. (22) by a cross market maker } j;$$

while the markets clear according to (21). Note that we continue to assume that the on-exchange supply schedules do not depend on off-exchange trading (Remark 6). Accordingly, the cross-market makers optimization in (23) contains two separate conditional expectations.

Remarks:

Remark 8 (Off-exchange trading). The model setup speaks to the SEC (2022b) proposal of "Open Competition Rule 615," which requires a mandatory auction for all "segmented orders," a term "designed to encompass those orders of individual investors [...]" (p. 70). The proposal allows "open competition trading centers"—including exchanges—to operate such auctions. However, the auctions are always conducted *separately* from conventional on-exchange trading, and, hence, we refer to them as "off-exchange." More generally, the model also applies to other forms of off-exchange trading where market makers compete to supply liquidity, like dark limit order books or periodic auctions. In all these settings, the endogenous clearing prices are subject to the price constraint (22).[15]

Remark 9 (Market maker types). The n cross market makers naturally include conventional high-frequency trading companies who can provide liquidity both on exchange and in the auction (off exchange). The SEC's proposal encourages broader participation in the auction, especially by "institutional investors" (p. 14, SEC, 2022b). However, it is possible that not all of them are able or willing to do so, and they form the ℓ on-exchange local market makers.[16]

Remark 10 (The price constraint). According to SEC (2022b), the proposed auction is "prohibited by Rule 611 from executing the segmented order" if it "did not generate a price that was at or within the best-priced protected quotations" (Footnote 250, p. 120). We model this constraint as inequality (22) and apply it to the cross market makers, notably the wholesalers (high-frequency market makers). This is because the wholesalers are in a unique position to influence the auction prices: The proposal discusses in an example that "[b]efore executing internally, however, the wholesaler would be required to submit the segmented order to a qualified auction with a specified limit price," which "would inform auction responders on how to price their orders ..." (p. 71). That is, a cross market maker not only triggers the auction but also influences its pricing, which must be subject to the OPR.

[15] A caveat is that the OPR only imposes obligations on "trading centers" to prevent trade-throughs. We do not model how such trading centers design policies like rerouting or outright rejection to prevent trade-throughs. Instead, we analyze how market makers might strategically adjust their supplies to prevent trade-throughs. This distinction between trading centers and market makers does not arise in Sections 2 and 3, where market makers or wholesalers effectively serve as off-exchange trading centers.

[16] It is a general concern whether institutional investors will participate regularly in the auctions (i.e., supply liquidity off exchange, in the context of our model). For example, Bryzgalova, Pavlova, and Sikorkaya (2023a) document the lack of participation in auctions organized by options designated market makers. In addition, Commissioner Hester M. Peirce caveats that "although allowing a broader set of market participants to interact with retail order flow is a goal of the proposal, institutional investors may not expend much effort to participate regularly in auctions" (Peirce, 2022). Further, in a comment letter to the SEC, Council of Institutional Investors writes, "for many institutional investors, the risk of potentially revealing their identities and trade interest to even a single dealer by participating in the proposed new auction mechanisms could materially outweigh any potential benefits of receiving the executions."

4.2 Equilibrium

Our equilibrium analysis yields a key novel insight of "liquidity shift," which we sketch below. The other details are deferred to the proof of Proposition 4. As in Section 2, we look for linear-strategy equilibria, where each investor submits the same optimal market order $z_i \beta t$, with β being their trading (endogenous) aggressiveness and t their trading motive as defined in (8). Our main insight arises from a cross market maker's optimal liquidity supply: On each market, she can condition her supply on the price of that market which reveals t to update $\mathbb{E}[v|\, t\,]$. Thus, for any given t, she chooses $\{x_j, \hat{x}_j, \omega_j\}$ to maximize the Lagrangian

$$\mathcal{L}(x_j, \hat{x}_j, \omega_j; t) = \big(p(x_j) - \mathbb{E}[v|\, t\,]\big)x_j + \big(\hat{p}(\hat{x}_j) - \mathbb{E}[v|\, t\,]\big)\hat{x}_j - \omega_j \cdot \big(\hat{p}(\hat{x}_j) - p(x_j)\big)\hat{Z},$$

where $\omega_j \geq 0$ is the Lagrangian coefficient that corresponds to the price constraint (22). Notably, her trades move prices $p(x_j)$ and $\hat{p}(\hat{x}_j)$ via market clearing (21).

The first-order condition with respect to the on-exchange supply x_j is

$$\text{(24)} \qquad \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}x_j} = \frac{\mathrm{d}p}{\mathrm{d}x_j}x_j + (p - \mathbb{E}[v|\, t\,]) + \omega_j \frac{\mathrm{d}p}{\mathrm{d}x_j}\hat{Z} = 0 \implies x_j = \left(\frac{\mathrm{d}p}{\mathrm{d}x_j}\right)^{-1}(\mathbb{E}[v|\, t\,] - p) - \omega_j \hat{Z}.$$

The resulting expression is reminiscent of (11) and (12) from Section 2. The difference is that the Lagrangian coefficient $\omega_j \geq 0$ is now in place of the price referencing derivative $\frac{\mathrm{d}\hat{p}}{\mathrm{d}p}$. In other words, ω_j reflects how binding the regulatory constraint (22) is and, hence, also how "tightly" the off-exchange $\hat{p}$ refers to the on-exchange p. In particular, when $\omega_j > 0$, the constraint binds in such a way that the market maker *reduces* her on-exchange liquidity supply. Doing so, she worsens the on-exchange p to satisfy the constraint that the off-exchange $\hat{p}$ must be weakly better. That is, consistent with the findings from Section 2, price referencing still softens on-exchange liquidity supply competition. The only difference is that the strength of this effect switches from the exogenous $\frac{\mathrm{d}\hat{p}}{\mathrm{d}p}$ to the endogenous ω_j.

This reduction in on-exchange supply is indeed a shift towards the off-exchange supply. The first-order condition of the Lagrangian with respect to $\hat{x}_j$ is

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\hat{x}_j} = \frac{\mathrm{d}\hat{p}}{\mathrm{d}\hat{x}_j}\hat{x}_j + (\hat{p} - \mathbb{E}[v|\,t\,]) - \omega_j \frac{\mathrm{d}\hat{p}}{\mathrm{d}\hat{x}_j}\hat{Z} = 0 \implies \hat{x}_j = \left(\frac{\mathrm{d}\hat{p}}{\mathrm{d}\hat{x}_j}\right)^{-1} (\mathbb{E}[v|\,t\,] - \hat{p}) + \omega_j \hat{Z}. \tag{25}$$

Note that the same term $\omega_j \hat{Z}$ appears, but now with the opposite sign to that in (24). This is the liquidity shift that closes the price gap that otherwise would violate the OPR inequality (22). The magnitude of ω_j measures the strength of the shift.

Equilibrium. The following proposition characterizes the equilibrium.

Proposition 4 (Equilibrium with off-exchange competition). There exists a linear-strategy equilibrium in the form of

$$\begin{aligned} &z_i = \beta_i \cdot (v - \bar{v} - \gamma_i u) \ \text{ for an investor } i; \\ &x_j(p) = b_j \cdot (p - \bar{v}) \text{ and } \hat{x}_j(\hat{p}) = \hat{b}_j \cdot (\hat{p} - \bar{v}) \ \text{ for a cross market maker } j; \\ &y_j(p) = c_j \cdot (p - \bar{v}) \ \text{ for an on-exchange local market maker } j, \end{aligned} \tag{26}$$

with coefficients given by

$$b_j = \hat{b}_j = \frac{(\zeta - 1)R - \omega_j \zeta \hat{R}}{(2\zeta - 1)\zeta} \frac{\mu\beta_i}{\phi}, \ c_j = \frac{(\zeta - 1)R}{(2\zeta - 1)\zeta} \frac{\mu\beta_i}{\phi}, \ \beta_i = \frac{1}{\rho}(1 - \zeta\phi), \text{ and } \gamma_i = \rho,$$

where the illiquidity index ζ is given by

$$\zeta = 1 + \frac{1}{n + R\ell - 2}; \tag{27}$$

and the cross market makers' Lagrangian coefficients $\{\omega_j\}_{j\in\{1,\dots,n\}}$ satisfy

$$\sum_{j=1}^{n} \omega_j = \frac{\ell}{n + R\ell - 1} \ \text{ and } \ 0 \le \omega_j < 1, \ \forall j \in \{1, \dots, n\}. \tag{28}$$

Further, fixing $\{v, u\}$, hence also t as defined in (8), the trading prices $\{p, \hat{p}\}$ are given by $\hat{p} = p = \bar{v} + \zeta\phi t$.

Recall that the Lagrangian coefficient ω_j measures the strength of the liquidity shift by market maker j. Equation (28) shows that such a shift exists if and only if there are local market makers, i.e., $\ell > 0$; and further, the aggregate shift $\sum_j \omega_j$ is monotone increasing in ℓ. Intuitively, when $\ell \geq 1$, there are more market makers on exchange than off exchange, yielding a better on-exchange price. But the regulatory constraint (22) does not allow it: the off-exchange $\hat{p}$ must be better than (or equal to) the on-exchange p. Therefore, the cross market makers must shift liquidity off exchange to improve $\hat{p}$, worsen p, and close the gap. This gap would be wider for larger ℓ, and thus requires a larger supply shift to close it.

We note that only the sum of all the Lagrangian coefficients is determined, as given by (28). In this sense, we have multiple equilibria, as the values of the individual ω_j cannot be uniquely determined, although the resulting market outcome, as characterized by ζ in (27), is unique.

Market quality. The equilibrium illiquidity (27) entails the following results.

Corollary 6 (Illiquidity with off-exchange competition). The illiquidity ζ exacerbates

- if there is more off-exchange trading (i.e., ζ increases with $\hat{R}$); or
- if more of the m market makers are constrained to be "local" (i.e., fixing m, ζ increases with ℓ).

The corollary shows that, just like in the baseline, illiquidity exacerbates with off-exchange fragmentation $\hat{R}$; cf. the ζ in (14). That is, even when there is off-exchange competition, Prediction 1 remains robust. The underlying intuition is the same: being exposed to more off-exchange orders makes every market maker less willing to compete on exchange.

Further, the corollary shows that illiquidity would be worse if more market makers are constrained to only supply liquidity on exchange. To see why, note that with fewer cross market makers competing off exchange ($n = m - \ell$), they each get a larger exposure of the off-exchange orders, which significantly reduce their incentive to compete on exchange. Contrarily, if there are many cross market makers, each getting a very thin slice of the off-exchange orders, then there is very little impact on their incentive to compete on exchange.

This result caveats that for the SEC's proposal of order-by-order auction to be efficient, indeed all the m market makers should be encouraged to participate in the (off-exchange) auction (which can be challenging for the reasons discussed in Footnote 16). The model of Ernst, Spatt, and Sun

(2022) also points to the importance of sufficient participation in the order-by-order auction but from an information channel, orthogonal to our paper.

Further extensions. Appendix B contains additional extensions. In Appendix B.1, we first generalize the equilibrium characterization by considering both ℓ on-exchange local and, additionally, $\hat{\ell}$ off-exchange local market makers (as shown in the dashed box in Figure 6). We then study in Appendix B.2 cross market makers' endogenous choices of their off-exchange exposures, as in Section 3, and show that the same key results are obtained: (i) absent the price referencing requirement (22), all orders are routed on exchange (cf. Proposition 3); and (ii) liquidity and trading efficiency (total gains from trade) are always higher without price referencing (cf. Corollary 5). Finally, Appendix B.3 further extends the policy analysis of Section 4.3 below.

4.3 Policy implications: The effectiveness of OBO auction

We now use Propositions 1 and 4 to compare the equilibrium outcomes of Sections 2 and 4.

Corollary 7 (With vs. without off-exchange competition). Given the same level of off-exchange fragmentation $\hat{R}$, off-exchange competition alleviates illiquidity ζ. Accordingly, gains from trade w are also higher. That is, for the same $\hat{R}$, ζ is lower and w is higher under Proposition 4 than under Proposition 1.

The intuition is as follows: Recall that market makers reduce on-exchange liquidity supply because they profit also from the referenced off-exchange prices. When competition shrinks such off-exchange profits, this incentive to reduce on-exchange liquidity supply weakens, and illiquidity is alleviated.

The corollary speaks to the SEC's OBO auction proposal, which requires each individual retail order be auctioned against all participating agents, just like in this extension. On the other hand, under the current practice, retail orders are routed off exchange to individual market makers for execution, subject to no off-exchange competition, just like in the baseline model in Section 2. According to Corollary 7, therefore, the proposed OBO auctions may indeed improve retail orders' execution quality.

While our policy implication echoes the SEC's proposal, it is worth noting that the underlying mechanism differs from the starting point of the SEC (2022a): They argue that retail flow is less toxic than on-exchange flow, and, therefore, should receive improved prices once more competition is introduced (SEC, 2022a). An important distinction is that our channel predicts that the proposed OBO auction would also help on-exchange liquidity via price referencing ($\hat{p} = p$). That is, the on-exchange institutional investors would also benefit, whereas the SEC's toxicity-driven argument is only about retail orders.

We note that the comparison in Corollary 7 assumes constant fragmentation $\hat{R}$. This is a plausible assumption as, under the current practice, almost all retail orders are routed off exchange to market makers, suggesting that $\hat{R}$ has reached its maximum possible level. Therefore, if the OBO auction proposal is implemented, $\hat{R}$ is unlikely to further increase. It is possible that $\hat{R}$ might decrease under the OBO auction proposal. For example, if the competition significantly reduces market makers' off-exchange profits, they may no longer have the incentive to spend PFOF on brokers, who then instead route more retail orders on exchange. If this happens, however, the market will become less fragmented, and market quality will still improve (Prediction 1 and Corollary 6), consistent with Corollary 7 albeit for a different reason. Appendix B.3 compares the case of fixed $\hat{R}$ and the case of endogenous $\hat{R}$ in more detail.

The broader policy insight of Corollary 7 is that off-exchange competition should be encouraged—to the extent that fragmentation $\hat{R}$ is unaffected—to improve market quality. Beyond retail trading, the lack of competition arises in midpoint crossing networks, single-dealer platforms, off-exchange broker-dealer internalization of institutional orders, and over-the-counter trading. Our model suggests that better market quality could be achieved by migrating trades to, for example, periodic auctions and dark limit order books, where there is competition in off-exchange trading.

5 Conclusion

Under Reg NMS and the OPR, off-exchange trades need to refer to the on-exchange quotations (the NBBO) to ensure that they do not execute at worse prices. While the regulation is about off-exchange order execution per se, we argue that such price referencing affects market makers' incentives to provide liquidity on exchange and thus also the on-exchange price that is used as the reference. In particular, we show that in response to off-exchange trading, market makers soften their on-exchange liquidity supply competition, which makes liquidity more expensive and raises trading costs for investors.

Market makers benefit from such illiquidity and from diverting orders off exchange. The latter is also facilitated by price referencing, because it guarantees best execution for all orders by equalizing on- and off-exchange prices. With no concern of their best-execution obligations, brokers can route orders to market makers as they please, for example via PFOF arrangements. Absent regulation, we argue that market makers may still have incentives and the means to voluntarily implement price referencing, for example by committing to an off-exchange pricing rule.

Surprisingly, the inefficiency induced by price referencing may exacerbate with more market makers. The reason is that they acquire (through brokers) more orders off exchange, exacerbating the softened on-exchange competition and the resulting illiquidity. Encouraging market makers to compete for off-exchange orders does not resolve the inefficiency either. In this case, market makers shift part of their on-exchange supply off-exchange to prevent price violations. The incentives to raise off-exchange trading persist.

These findings extend to several forms of off-exchange trading, including wholesalers' in-house execution of retail orders (from PFOF arrangements), broker-dealer internalization, single-dealer platforms—as long as the off-exchange prices are a function of on-exchange quotations.

Appendix

A Heterogeneous on- and off-exchange orders

Our main analysis has made the simplifying assumption of homogeneous investors, who submit identical orders on- and off-exchange. In reality, however, these orders are heterogeneous and likely have different characteristics. For example, currently most of retail orders in the U.S. are routed off exchange, leaving mostly institution orders on exchange. In this extension, we reexamine our main mechanism in view of heterogeneous on- and off-exchange orders.

Specifically, we take the on-exchange aggregate order Z and the market maker j's off-exchange orders $\hat{z}_j$ as given and revisit her optimization as outlined in Section 2.2.2. We consider the general price referencing rule of $\hat{p}(p)$. As before, by conditioning on the price p, the market maker equivalently observes the on-exchange aggregate order Z, thanks to the market clearing condition $x_j + \sum_{j' \neq j} x_{j'}(p) = Z$. She therefore chooses her on-exchange supply x_j to maximize

$$\mathbb{E}\big[(p - v)x_j + (\hat{p}(p) - v)\hat{z}_j \,|\, p\big] = (p - \mathbb{E}[v|\,Z])x_j + (\hat{p}(p) - \mathbb{E}[v|\,Z])\mathbb{E}\big[\hat{z}_j|\,Z\big],$$

wary of her price impact $\frac{\mathrm{d}p}{\mathrm{d}x_j}$. Analogous to (12), her first-order condition yields

$$\frac{\mathrm{d}p}{\mathrm{d}x_j}x_j + (p - \mathbb{E}[v|\,Z]) + \frac{\mathrm{d}\hat{p}}{\mathrm{d}p}\frac{\mathrm{d}p}{\mathrm{d}x_j}\mathbb{E}\big[\hat{z}_j|\,Z\big] = 0 \implies x_j = -\left(\frac{\mathrm{d}p}{\mathrm{d}x_j}\right)^{-1}(p - \mathbb{E}[v|\,Z]) - \frac{\mathrm{d}\hat{p}}{\mathrm{d}p}\mathbb{E}\big[\hat{z}_j|\,Z\big].$$

As can be seen, the off-exchange exposure $\hat{z}_j$ continues to distort the on-exchange supply x_j, but now only via the conditional expectation $-\mathbb{E}\big[\hat{z}_j|\,Z\big]$.

To compare, our previous analysis in “Generalization” on page 14 finds three conditions for price referencing to distort liquidity supply: (i) $\frac{\mathrm{d}p}{\mathrm{d}x_j} \neq$, (ii) $\frac{\mathrm{d}\hat{p}}{\mathrm{d}p} \neq 0$, and (iii) $\hat{z}_j \neq 0$. This extension therefore generalizes (iii) to $\mathbb{E}\big[\hat{z}_j|\,Z\big] \neq 0$. That is, despite the demand heterogeneity, the model's main insight remains robust, as long as the market maker's off-exchange orders $\hat{z}_j$ can be predicted by the on-exchange orders Z. Notably, it is natural that the price referencing rules satisfy $\frac{\mathrm{d}\hat{p}}{\mathrm{d}p} > 0$. Then as long as $\mathrm{sign}\big[\mathbb{E}\big[\hat{z}_j|\,Z\big]\big] = \mathrm{sign}[Z]$, the distortion always *hurts* the on-exchange liquidity supply.

The requirement of $\text{sign}\big[\mathbb{E}\big[\hat{z}_j | Z\big]\big] = \text{sign}[Z]$ might seem strong at first sight. However, it should be noted that in real-world trading, the buy and the sell sides are effectively separate: When market makers execute off-exchange buy (sell) orders, they refer to the prevailing on-exchange ask (bid) price, which is only applicable to on-exchange buy (sell) orders. That is, the requirement becomes, respectively for buy and sell orders, $\text{sign}\big[\mathbb{E}\big[\hat{z}_j \big| Z, \hat{z}_j > 0\big]\big] > 0$ and $\text{sign}\big[\mathbb{E}\big[\hat{z}_j \big| Z, \hat{z}_j < 0\big]\big] < 0$, which always hold.

Empirically, suppose we use a linear model to approximate thee relation between Z and $\hat{z}_j$:

$$\mathbb{E}\big[\hat{z}_j | Z\big] \approx \alpha + \beta Z,$$

where α and β are some (regression) coefficients. This approximation then brings the following empirical prediction:

> **Prediction 5:** A market maker supplies less liquidity on exchange if her off-exchange orders $\hat{z}_j$ are more sensitive (larger β) to the on-exchange orders Z.

This is a novel prediction arising from the extension above and, more importantly, from the price referencing mechanism that incentivizes market makers to acquire off-exchange orders.

B Further extensions of off-exchange competition

In this appendix, we study the following further extensions of Section 4. First, while Section 4 only allows ℓ on-exchange local market makers, Appendix B.1 generalizes the equilibrium by adding $\hat{\ell}$ off-exchange local market makers. Second, Appendix B.2 studies how cross market makers' endogenous acquisition of off-exchange exposures affects trading efficiency. Finally, Appendix B.3 examines the effectiveness of the SEC's proposal of order-by-order auction, relative to the current market practice.

B.1 Allowing off-exchange local market makers

In Section 4, we assume that there are ℓ on-exchange local market makers, who can only supply liquidity on exchange. For symmetry, we generalize Proposition 4 by introducing also $\hat{\ell}$ (≥ 0) off-exchange local market makers, who can only supply liquidity in the off-exchange auction (see the dashed box in Figure 6). Specifically, echoing the description in "Trading" on p. 30, each j of these market makers posts her supply schedule $\hat{y}_j(\hat{p})$ to solve $\max_{\hat{y}_j(\hat{p})} \mathbb{E}\big[(\hat{p} - v)\hat{y}_j(\hat{p}) | \hat{p}\big]$. As a result, the off-exchange market clearing condition becomes $\sum_{j=1}^{n} \hat{x}_j(\hat{p}) + \sum_{j=1}^{\hat{\ell}} \hat{y}_j(\hat{p}) = \hat{Z}$; cf. (21). All other model elements remain the same. Conjecturing that the off-exchange local market makers also play the linear strategy of the form

$$\hat{y}_j(\hat{p}) = \hat{c}_j \cdot (\hat{p} - \bar{v}), \tag{B.1}$$

we then obtain the following equilibrium:

Proposition 5 (Generalized equilibrium with off-exchange competition). There always exists a linear-strategy equilibrium in the form of (26) and (B.1), with coefficients given by

$$b_j = \frac{(\zeta - 1)R - \omega_j \zeta \hat{R}}{(2\zeta - 1)\zeta} \frac{\mu\beta_i}{\phi}, \ \hat{b}_j = \frac{(\hat{\zeta} - 1)\hat{R} - \omega_j \hat{\zeta} \hat{R}}{(2\hat{\zeta} - 1)\hat{\zeta}} \frac{\mu\beta_i}{\phi},$$
$$c_j = \frac{(\zeta - 1)R}{(2\zeta - 1)\zeta} \frac{\mu\beta_i}{\phi}, \ \hat{c}_j = \frac{(\hat{\zeta} - 1)\hat{R}}{(2\hat{\zeta} - 1)\hat{\zeta}} \frac{\mu\beta_i}{\phi}, \ \beta_i = \frac{1}{\rho}\Big(1 - (\zeta R + \hat{\zeta}\hat{R})\phi\Big), \text{ and } \gamma_i = \rho,$$

where the on-exchange illiquidity index ζ, the off-exchange $\hat{\zeta}$, and the cross market makers' Lagrangian coefficients $\{\omega_j\}_{j\in\{1,\ldots,n\}}$ are given by the following, depending on whether the regulatory constraint (22) is imposed and on whether it binds:

- Suppose that (22) is imposed. Then it binds if and only if $\hat{\ell} < \ell$, in which case $\{\omega_j\}_{j\in\{1,\ldots,n\}}$ jointly satisfy (hence not uniquely determined)

$$\sum_{j=1}^{n} \omega_j = \frac{\ell - \hat{\ell}}{n + R\ell + \hat{R}\hat{\ell} - 1} \text{ and } 0 \leq \omega_j < 1, \ \forall j \in \{1, \ldots, n\}; \tag{B.2}$$

and the illiquidity indices are the same:

$$\zeta = \hat{\zeta} = h\big(n + R\ell + \hat{R}\hat{\ell}\big), \text{ where } h(x) := 1 + \frac{1}{x-2} \text{ for } x > 2. \tag{B.3}$$

- Suppose $\hat{\ell} \geq \ell$ or that (22) is *not* imposed. Then, in equilibrium, $\omega_j = 0$, $\forall j \in \{1, ..., n\}$; and $\zeta = h(n + \ell)$ and $\hat{\zeta} = h(n + \hat{\ell})$, where $h(\cdot)$ is given in (27) above.

Further, fixing $\{v, u\}$, hence also t as defined in (8), the trading prices $\{p, \hat{p}\}$ are given by $p = \bar{v} + \zeta\phi t$ and $\hat{p} = \bar{v} + \hat{\zeta}\phi t$.

Note that when $\hat{\ell} = 0$, the previous Proposition 4 is obtained as a special case. Indeed, the intuition remains exactly the same: If $\ell > \hat{\ell}$, then there are more market makers supplying liquidity on exchange than off exchange ($n+\ell > n+\hat{\ell}$), yielding a better on-exchange price. But the regulatory constraint (22), if imposed, does not allow it: the off-exchange $\hat{p}$ must be better than (or equal to) the on-exchange p. Therefore, the cross market makers must shift liquidity off exchange, improving $\hat{p}$ while worsening p, to close the gap. On the contrary, if $\hat{\ell} \geq \ell$ or if no such regulation is imposed, the two venues clear separately and no such shift occurs.

Figure B.1 illustrates the illiquidity indices $\{\zeta, \hat{\zeta}\}$ by varying $\hat{\ell} \in [0, 10]$ on the horizontal axis while fixing $\ell = 4$. When $\ell < \hat{\ell}$, illiquidity is lower off exchange than on exchange (the dashed $\hat{\zeta}$ below the dot-dashed ζ), price is better off exchange, and the equilibrium is unconstrained. When $\hat{\ell}$ drops below ℓ, illiquidity ζ and $\hat{\zeta}$ without the regulatory constraint (22) are shown in the two diverging gray lines; while with the regulation, both $\hat{\zeta}$ and ζ collapse into an average of the two gray lines.

The illiquidity function $h(\cdot)$ given in (B.3) highlights that what matters is the average number of market makers facing every order: when the regulatory constraint (22) binds in Proposition 5, an order expects $n + \ell$ market makers if on-exchange (with probability R) and $n + \hat{\ell}$ if off-exchange (with probability $\hat{R} = 1 - R$). This insight, in fact, also underlies Proposition 1, where (14) can be equivalently written as $\zeta = h(Rm + (1 - R) \cdot 1)$: an order expects m market makers if it is on-exchange and a monopolist if off-exchange.

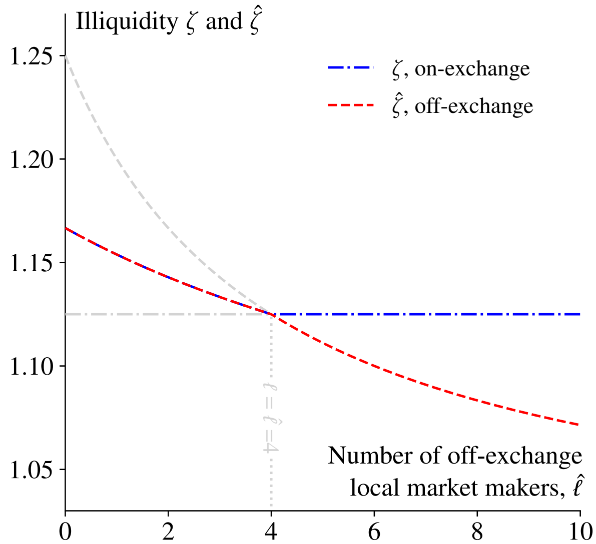

Figure B.1: Illiquidity with off-exchange competition. This figure plots the equilibrium illiquidity indices ζ (on-exchange, dot-dashed) and $\hat{\zeta}$ (off-exchange, dashed) against varying $\hat{\ell} \in [0, 10]$, the number of off-exchange local market makers. The number of on-exchange local market makers is fixed at $\ell = 4$. The vertical dotted line indicates when $\ell = \hat{\ell}$. The other parameters are set at $\bar{v} = 0.0$, $\mu = \rho = 1.0$, $\tau_v = 0.1$, $\phi = 0.3$, $n = 6$, and $R = \hat{R} = 0.5$.

B.2 Endogenous routing and trading efficiency

This part of the appendix endogenizes order routing $\{R, \hat{R}\}$ and compares two cases, with vs. without the price constraint (22). As such a constraint can be driven by regulations like the OPR, for notation clarity, we shall subscript equilibrium objects with the constraint by "OPR" and those without by "no-OPR," while keeping in mind that there are alternative mechanisms that can drive price referencing (see Section 3.3).

The price referencing constraint (22) **exists.** As seen from Proposition 4, both on- and off-exchange prices become the same: $p = \hat{p}$. In other words, consistent with the finding from Section 3.3, the price referencing constraint (22) dissolves (individual) investors' incentive to direct their orders and helps brokers always fulfill their best-execution obligation. Investors' indifference then facilitates market makers' fragmenting their orders' off exchange. For concreteness, consider the following setup similar to Section 3: we add a pre-trading period in which each cross market maker j can choose her off-exchange exposure $\hat{r}_j$, so that $\hat{R} = \sum_{j=1}^{n} \hat{r}_j$. That is, taking all others $\hat{r}_{j'}$ as given ($\forall j' \neq j$), a cross market maker j chooses her $\hat{r}_j$ to maximize her expected trading profit minus the cost:

$$\max_{\hat{r}_j} \mathbb{E}\big[(p-v)x_j(p) + (\hat{p}-v)\hat{x}_j(\hat{p})\big] - \kappa \hat{r}_j \mu, \tag{B.4}$$

where the expected trading profit can be derived following the trading equilibrium characterized by Proposition 4. Similar to Proposition 2, it can be shown that a unique symmetric-strategy Nash equilibrium exists. For the following discussion, however, it suffices to note that in such an equilibrium, assuming a sufficiently low cost κ, there is always non-zero off-exchange trading, i.e., $\hat{R}_{\text{OPR}} > 0$.

Without the price referencing constraint (22). Proposition 5 characterizes the no-constraint equilibrium. Notably, the on- and off-exchange trades are then cleared *separately*. There is no longer liquidity shift from on exchange to off exchange. Below we show that the same result from Proposition 3 remains robust that $R_{\text{no-OPR}} = 1$ and $\hat{R}_{\text{no-OPR}} = 0$ in equilibrium, and we do so by assuming the opposite, i.e., $\hat{R} \in (0, 1)$. According to Proposition 5, under the maintained assumption of $\hat{\ell} = 0 < \ell$ (Remark 9), the off-exchange illiquidity is worse than on-exchange, i.e., $\hat{\zeta} > \zeta$. (This can be seen from Figure B.1, where the gray dot-dashed line for ζ is below the gray dashed line for $\hat{\zeta}$.) Consequently, the off-exchange per-share trading cost is higher: $z_i \cdot (\hat{p} - p) = (\hat{\zeta} - \zeta)\phi\beta t^2 > 0$, meaning that investors have strict preference to direct their orders to the exchange, leading to $R_{\text{no-OPR}} = 1$ and $\hat{R}_{\text{no-OPR}} = 0$ in equilibrium.

Comparing trading efficiency. Consistent with the result from Section 3.3, regulations like the OPR worsen trading efficiency:

Corollary 8 (Trading efficiency with vs. without OPR). Both the total gains from trade and the investors' gains from trade are higher if the OPR is lifted: $w_{\text{no-OPR}} \geq w_{\text{OPR}}$ and $\pi^{\text{I}}_{\text{no-OPR}} \geq \pi^{\text{I}}_{\text{OPR}}$.

To see why, recall from Section 4.2 that the binding price constraint (22) effectively shifts cross market makers' supply off exchange. That is, the equilibrium illiquidity can be thought of as some form of the average between the more competitive on-exchange outcome and the less competitive off-exchange outcome. However, by lifting the price constraint (22), the better on-exchange liquidity is attainable to all orders (because $R_{\text{no-OPR}} = 1$), thus improving also trading efficiency.

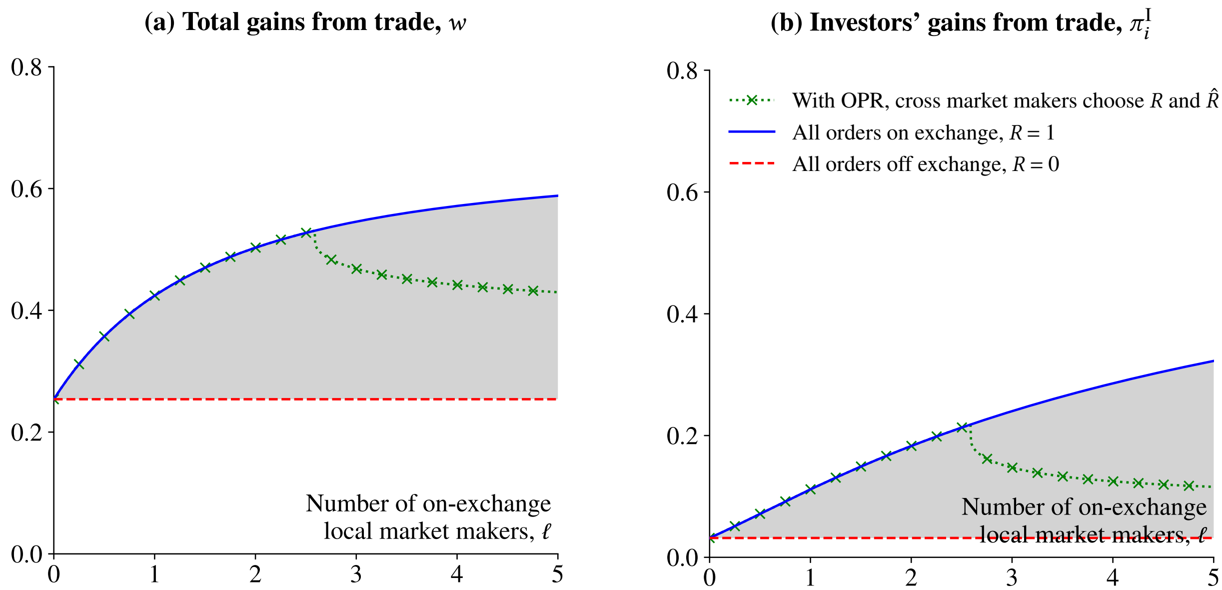

Figure B.2: Trading efficiency, with vs. without the price referencing constraint (22). This figure illustrates the total gains from trade w in Panel (a) and the investors' gains from trade π^{I} in Panel (b) by varying the number of on-exchange local market makers. Without the constraint (22), all orders routed to the exchange, so that $R = 1$, as shown by the blue solid line. With the constraint, cross market makers can endogenously choose $\hat{R}$ and R, and the shaded area indicates all possible outcomes. In particular, the cross-dotted line shows the equilibrium levels of w and π_i^{I}, when each cross market maker can acquire off-exchange exposure at a flat cost of $\kappa = 0.02$ per unit, i.e., when they all solve the optimal off-exchange exposures according to (B.4). The dashed line shows the lower bound of $R = 0$, i.e., all orders off exchange. The other parameters are set at $\bar{v} = 0.0$, $\mu = \rho = 1.0$, $\tau_v = 0.1$, $\phi = 0.7$, and $n = 5$, and $\hat{\ell} = 0$.

Figure B.2 illustrates the implications of the price referencing constraint (22) for the total gains from trade w in Panel (a) and for the investors' gains from trade π_i^{I} in (b). In both panels, we fix the number of cross market makers constant at $n = 5$, varying the number of on-exchange local market makers, $\ell \in [0, 5]$. (There are no off-exchange local market makers, i.e., $\hat{\ell} = 0$.) Without the price referencing constraint, all orders will be routed on exchange for best-execution, so that $R_{\text{no-OPR}} = 1$, as shown in the solid line. It increases with ℓ, because with in total $n + \ell$ market makers competing to provide liquidity on exchange, trading price becomes more efficient as ℓ increases.

With the constraint (22), prices will bind with $p = \hat{p}$ and best-execution is always guaranteed irrespective of routing. This allows cross market makers to arrange routing with brokers to their benefit. The equilibrium R_{OPR}, however, is always bounded between 0 and 1, as shown in the shaded area. The lowest possible gains from trade are achieved when $R_{\text{OPR}} = 0$, as shown in the dashed line, which is flat in ℓ. This is because when $R_{\text{OPR}} = 0$, all orders are off exchange, and so only the n cross market makers compete to supply liquidity to them. As an example, the cross-dotted line illustrates the equilibrium w and π_i^{I} when each cross market maker solves the problem (B.4).

Summary. Lifting the price referencing constraint (22) improves both the overall trading efficiency and the investors' gains from trade. To emphasize, the key intuition underlying this result is that: (i) cross market makers soften their on-exchange competition to satisfy the price referencing requirement, so that $\hat{p} = p$ binds; and that (ii) the binding $\hat{p} = p$ dissolves investors' (or their brokers') incentive to route orders, allowing cross market makers to arrange routing to their benefit but at the cost of the investors' and of the overall efficiency. Both (i) and (ii) are robust features of price referencing, as seen also earlier in Sections 2 and 3.

B.3 Effectiveness of the SEC's proposal of Rule 615 with endogenous off-exchange trading

However, as we have seen in Section 2.4, the n cross market makers' incentives to acquire off-exchange exposure will change. Previously in Section 4.3, we consider the case where the SEC's proposed OBO auction would not affect the amount of off-exchange trading $\hat{R}$. We examine in this appendix the general case of endogenous responses of market makers' $\{\hat{r}_j\}$, hence also $\hat{R}$ and R, under the OBO auction proposal. Specifically, we allow the n cross market makers to endogenously choose their off-exchange exposures $\{\hat{r}_j\}$. Proposition 2 has shown that without OBO auction, there exists a unique symmetric-strategy equilibrium. We can show that the same holds with the OBO auction (but for brevity the characterization is omitted). Figure B.3 then numerically examines the effectiveness of the OBO auction by varying the number of on-exchange local market makers l.

Panel (a) plots the equilibrium fraction R of on-exchange orders. It can be seen that either with

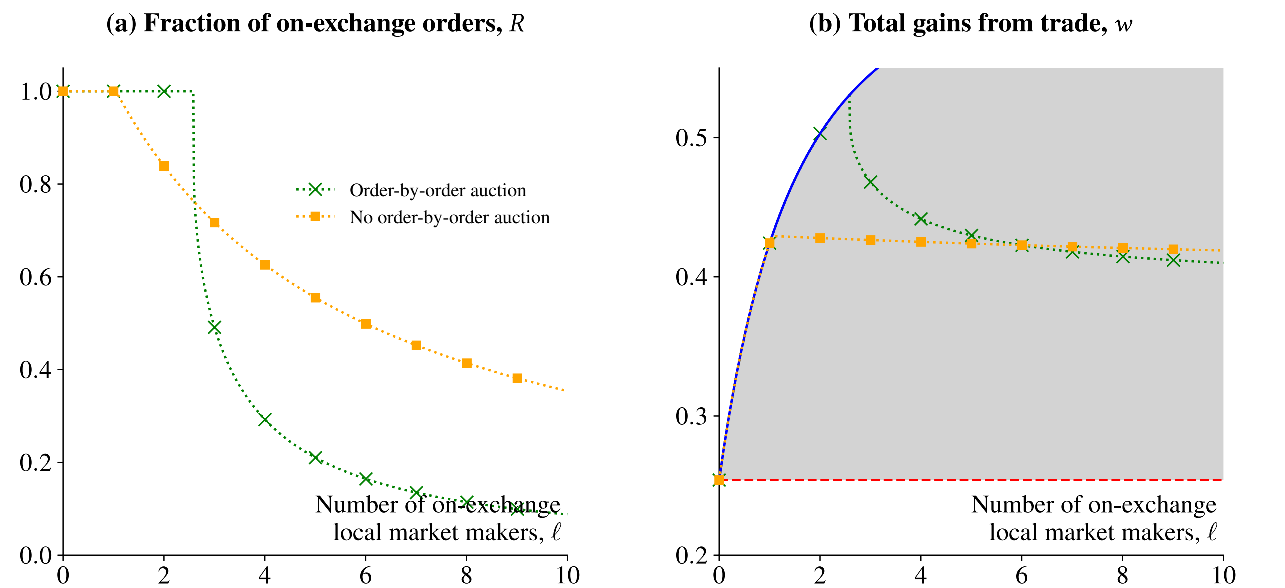

Figure B.3: Effectiveness of the SEC's proposal of Rule 615. This figure illustrates the effectiveness of the SEC's order-by-order auction proposal (proposed Rule 615). Panel (a) plots the endogenous fraction R of on-exchange orders, while Panel (b) plots the total gains from trade, with the number of on-exchange local market makers l varying on the horizontal axis. In particular, the cross-dotted, the blue-solid, and the red-dashed lines in Panel (b) are the same as in Figure B.2(a), with the shaded area representing all possible outcomes of $R \in [0, 1]$. The other parameters are set at $\kappa = 0.02$, $\bar{v} = 0.0$, $\mu = \rho = 1.0$, $\tau_v = 0.1$, $\phi = 0.7$, $n = 5$, and $\hat{\ell} = 0$.

or without OBO auction, the n cross market makers start to acquire orders off exchange only when l is sufficiently large. Intuitively, this is because more on-exchange local market makers raise the on-exchange liquidity supply competition, thus making it relatively more profitable to fragment orders off exchange where there is less competition. However, we see two differences:

- When there is no OBO auction (the square-dotted line), cross market makers start to acquire off-exchange orders at around $\ell = 1$, sooner than when there is order-by-order auction (the cross-dotted line) at around $\ell = 3$. This is because the benefit of fragmenting orders off exchange is stronger without OBO auction, as the fragmenting market maker becomes the monopolist liquidity supplier for those fragmented orders.
- Yet, for relatively large ℓ, more orders are fragmented off exchange with OBO auction than without. This is because of the investors' endogenous response: All else being equal, diverting

order flow off exchange increases illiquidity and makes investors reduce their trading aggressiveness β, which, in turn, lowers market makers' profits and thus their willingness to do so. This channel is stronger in the setting without OBO auctions, because the impact of fragmentation on illiquidity is stronger (as shown in the previous subsection).

As seen in Panel (a), these two effects let the two lines of R cross. In particular, it is possible that (when ℓ is large) more orders are fragmented off exchange in the setting with OBO auction. The elevated level of off-exchange trading worsens illiquidity, a channel that dominates the increased competition effect of the OBO-auction. Panel (b) first replicates Figure B.2(a) and then adds the gains from trade without OBO (the square-dotted line). Notably, for sufficiently large ℓ, the cross-dotted line crosses below the square-dotted line, suggesting that the OBO auction can hurt welfare when market makers endogenously acquire off-exchange exposure.

C Proofs

Lemma 1

Proof. The discussion in (2.2.1) before the lemma has shown that the optimal market order is $z_i = \frac{1}{\rho}(v - p) - u$, where p, as given in (6), is a linear combination of u and v, both of which are known to the investor. Therefore, we have verified the linear conjecture of $z_i = \alpha_i + \beta_i \cdot (v - \bar{v} - \gamma_i u)$. Matching the coefficients, we require $\alpha_i = (A - R\mu\alpha)/(B\rho)$, $\beta_i = (B - R\mu\beta)/(B\rho)$, and $\beta_i\gamma_i = 1 - R\mu\beta\gamma/(B\rho)$. This linear equation system uniquely determines $\{\alpha_i, \beta_i, \gamma_i\}$ in terms of $\{\alpha, \beta, \gamma\}$. Importantly, this solution is the same across all i, and hence we must have $\alpha_i = \alpha$, $\beta_i = \beta$, and $\gamma_i = \gamma$, solving which yields (7). □

Lemma 2

Proof. The preceding discussion has derived a market maker j's first-order condition as in (11). Substitute in $\hat{z}_j = \hat{r}_j\mu \cdot (\alpha + \beta t)$ and then, by market clearing (1), $t = \frac{1}{R\mu\beta}(A + B \cdot (p - \bar{v})) - \frac{\alpha}{\beta}$ to get:[17]

$$x_j = \left(\left(\frac{\alpha\phi}{\beta} - \frac{A\phi}{R\mu\beta}\right)B_{-j} - \frac{A}{R}\hat{r}_j\right) + \left(\left(1 - \frac{B\phi}{R\mu\beta}\right)B_{-j} - \frac{B}{R}\hat{r}_j\right)(p - \bar{v}),$$

thus verifying the linear conjecture of $x_j = a_j + b_j \cdot (p - \bar{v})$ made in (5). To proceed, we aggregate the supply across all market makers to get

$$\sum_{j=1}^{m} x_j = \left(\left(\frac{\alpha\phi}{\beta} - \frac{A\phi}{R\mu\beta}\right)(m-1)B - \frac{A}{R}(1-R)\right) + \left(\left(1 - \frac{B\phi}{R\mu\beta}\right)(m-1)B - \frac{B}{R}(1-R)\right)(p - \bar{v}),$$

which must match the coefficients A and B in $\sum_{j=1}^{m} x_j = A + B \cdot (p - \bar{v})$. Solving for A and B (and ignoring the trivial root of zero), we get

$$A = \left(R - \frac{1}{m-1}\right)\mu\alpha \quad \text{and} \quad B = \left(R - \frac{1}{m-1}\right)\frac{\mu\beta}{\phi}. \tag{C.5}$$

Substitute these into the solution of x_j above, match coefficients of a_j and b_j, and we obtain the solution of a_j and b_j stated in (13). Finally, to ensure that the first-order condition (11) indeed identifies a maximum, we need to examine whether the second-order condition $-2B_{-j} \leq 0$ holds at the above solution: $-B_{-j} = -B + b_j = -R_{-j}\frac{(m-1)R-1}{(m-1)R+1}\frac{\beta\mu}{\phi}$, which is indeed negative because we assume $\beta > 0$ and because Condition (a) implies $(m-1)R > \frac{1}{1-\phi} > 1$. □

Proposition 1

Proof. Lemma 1 and (C.5) imply a linear equation system that uniquely solve for $\{\alpha, \beta, \gamma, A, B\}$. The solution of $\{\alpha, \beta, \gamma\}$ are as stated in the proposition (recall from Lemma 1 that all investors have the same $\alpha_i = \alpha$, $\beta_i = \beta$, and $\gamma_i = \gamma$). Note that indeed $\beta > 0$, thus ensuring the market makers' second-

[17] Recall that the on-exchange supply x_j is a function of the on-exchange p, but not of the off-exchange orders $\hat{z}_j$—the on- and off-exchange markets are "disconnected" (see Remark 6). This assumption ensures that, in this step, the variable t can only be written as a linear function of p via the market clearing condition (1). If instead the the supply can be contingent on both p and $\hat{z}_j$, then one can in addition substitute $t = \frac{1}{\hat{r}_j\mu}\hat{z}_j - \frac{\alpha}{\beta}$ into the first-order condition (11). This flexibility would make the equilibrium expression of $x_j(p, \hat{z}_j)$ indeterminate. That is, this assumption ensures the unique linear functional form of the supply.

order conditions (see the proof of Lemma 2). Substituting these into a_j and b_j as given by Lemma 2, we obtain the solution of $\{a_j, b_j\}$ as stated in the proposition. Finally, the equilibrium price p is found via the market clearing condition of $A + B \cdot (p - \bar{v}) = R\mu\beta t$. □

Corollary 1

Proof. The expression (19) directly follows π_i^{I} and π_j^{M} given by (17) and (18). Observe that the effects of m and R on w only go through ζ. In particular, w is a concave quadratic function of ζ, reaching its maximum at $\zeta = 1$. Therefore, w monotonically decreases in $\zeta \in \left(\frac{m-1}{m-2}, \frac{1}{\phi}\right)$. Then, following (14), w is indeed monotone increasing in m and in R. □

Proposition 2

Proof. We prove a more general version of the proposition: Only the first n of the m market makers are able to endogenously choose their off-exchange exposure $\hat{r}_j$, for $j \in \{1, ..., n\}$, while the other $m - n$ cannot and have $\hat{r}_j = 0$, for $j \in \{n + 1, ..., m\}$. The proposition then becomes a special case by setting $n = m$. This generalization helps the analysis in Appendix B.3. For clarity, we refer to the first n market makers as "cross" market makers, while the other $m - n$ as "local" market makers.

Consider a cross market maker j's problem (20). She takes as given all others' $\hat{r}_{j'}$ and choose her $\hat{r}_j$ to solve (20). In particular, the objective function only depends on $\hat{R}_{-j} = \sum_{j' \neq j} \hat{r}_{j'}$. Therefore, we want to find her best response of $\hat{r}_j$ to $\hat{R}_{-j}$.

Note that $\hat{r}_j$ can range from 0 up to $1 - \hat{R}_{-j} - \frac{1}{(m-1)(1-\phi)}$, where the upper bound is implied by the requirement of $\zeta = 1 + \frac{1}{(m-1)R-1} < \frac{1}{\phi}$ from Proposition 1, with $R = 1 - \hat{R} = 1 - \hat{r}_j - \hat{R}_{-j}$. We first assume the existence of the market maker's *interior* best response $\hat{r}_j^*$ to $\hat{R}_{-j}$, i.e., $0 < \hat{r}_j^* < 1 - \hat{R}_{-j} - \frac{1}{(m-1)(1-\phi)}$. Later we turn back to establish the existence.

Given the focus on interior solutions, we examine the market maker's first-order derivative:

$$\text{(C.6)} \qquad \frac{\mathrm{d}}{\mathrm{d}\hat{r}_j}\Big(\pi_j^{\mathrm{M}}(\hat{r}_j;\hat{R}_{-j}) - \kappa\hat{r}_j\mu\Big) = \frac{\mathrm{d}\pi_j^{\mathrm{M}}}{\mathrm{d}\zeta}\frac{\mathrm{d}\zeta}{\mathrm{d}R}\frac{\mathrm{d}R}{\mathrm{d}\hat{r}_j} - \kappa\mu = \frac{\mu}{\rho\tau_v}\Big(1-\hat{R}_{-j}\Big)(m-1)f(\zeta;\phi) - \kappa\mu,$$

where we define an auxiliary function

$$\text{(C.7)} \qquad f(\zeta;\phi) := \frac{(\zeta-1)^3}{(2\zeta-1)^2}\Big(-4\phi\zeta^2 + (2+3\phi)\zeta - \phi\Big).$$

Clearly, the second-order derivative is proportional to $f'(\zeta)\frac{\partial\zeta}{\partial\hat{r}_j}$, meaning that it has the same sign as $f'(\zeta)$, because $\frac{\partial\zeta}{\partial\hat{r}_j} > 0$.

We now characterize the shape of $f(\zeta)$ in $\zeta \in [1, 1/\phi]$: (1) It starts at $f(1) = 0$ and ends at $f(1/\phi) = -\frac{(1-\phi)^4}{(2-\phi)\phi^2} < 0$. (2) It is initially increasing: $\lim_{\zeta\downarrow 1}\frac{f'(\zeta)}{(\zeta-1)^2} = 6(1-\phi) > 0$. (3) It is quasi-concave, because at any stationary point, we have $f'(\zeta) = 0 \implies \phi = \frac{4\zeta^2-2\zeta+1}{12\zeta^3-16\zeta^2+8\zeta-1}$, plugging which into $f''(\zeta;\phi)$ yields a strictly negative value (as $\zeta > 1$). (4) Combining the above, we know that there exists a unique maximum of $f_{\max} > 0$ at some $\zeta_{\max} \in (1, 1/\phi)$, defined as $f_{\max} = f(\zeta_{\max}) \geq f(\zeta)$ for all $\zeta \in [1, 1/\phi]$. (5) Finally, there exists a unique $\zeta_0 \in (\zeta_{\max}, 1/\phi)$ such that $f(\zeta_0) = 0$ and $f'(\zeta) < 0$ for all $\zeta \in (\zeta_{\max}, \zeta_0]$.

We turn back to the market maker's first-order derivative (C.6), which is a linear transformation of $f(\cdot)$: f is scaled by $\frac{\mu}{\rho\tau_v}\Big(1-\hat{R}_{-j}\Big)$ and then shifted down by $\kappa\mu$. Given the above characterization of the shape of $f(\cdot)$, therefore, as long as

$$\text{(C.8)} \qquad \kappa \leq \frac{1}{\rho\tau_v}\Big(1-\hat{R}_{-j}\Big)(m-1)f_{\max},$$

the first-order condition—setting (C.6) to zero—has one or two roots in terms of ζ, and only the larger root satisfies both the first-order and the second-order conditions. (Recall that the second-order derivative is negative if and only if $f'(\zeta) < 0$, which holds at the larger root because $f(\cdot)$ is downward sloping there.) Denote this root by $\zeta^*(\hat{R}_{-j})$. Then the *interior* best response $\hat{r}_j^*$ is the unique solution implied by $\zeta^*(\hat{R}_{-j}) = 1 + \frac{1}{(m-1)(1-\hat{r}_j^*-\hat{R}_{-j})-1}$.

In particular, when (C.8) holds, under the first-order and the second-order conditions, by implicit function theorem, ζ^* is monotone decreasing in $\hat{R}_{-j}$. Note also that $\hat{r}_j^*$ increases with ζ^*. As such, the best response $\hat{r}_j^*$ decreases with $\hat{R}_{-j}$, until when (C.8) fails. (This best response is numerically illustrated by the thick blue line in Figure 3(b).) Therefore, if a symmetric-strategy equilibrium exists, it must be unique, because the downward sloping best response of $\hat{r}_j = \hat{r}_j^*(\hat{R}_{-j})$ always intersects once and only once with the symmetry-implied $\hat{r}_j = \frac{1}{n-1}\hat{R}_{-j}$ (only n cross market makers can choose $\hat{r}_j$).

Finally, we examine the existence of the *interior* symmetric-strategy equilibrium. This requires two conditions: (i) The first-order condition of (C.6) equal to zero must have solution when $\hat{R}_{-j} = (n-1)\hat{r}_j$ and also $\zeta = 1 + \frac{1}{(m-1)(1-n\hat{r}_j)-1}$. (ii) No cross market maker has incentive to deviate to either the lower corner of $\hat{r}_j = 0$ or the upper corner of $\hat{r}_j = 1 - \hat{R}_{-j} - \frac{1}{(m-1)(1-\phi)}$.

For (i), note that symmetry implies $\hat{r}_j = \frac{(m-2)\zeta-(m-1)}{(\zeta-1)(m-1)n}$, plugging which into the first-order condition (setting (C.6) to zero) yields

$$\text{(C.9)} \qquad \kappa = \frac{1}{\rho\tau_v}\frac{m-1}{n}g(\zeta), \quad \text{where } g(\zeta) := \frac{(2-\frac{m-n}{m-1})\zeta - 1}{\zeta - 1} f(\zeta).$$

This is the symmetry-implied first-order condition. For it to have solution, it is equivalent to require

$$\text{(C.10)} \qquad \kappa \le \frac{1}{\rho\tau_v}\frac{m-1}{n}\left(\max_{\zeta} g(\zeta)\right) =: \bar{\bar{\kappa}}(m,n).$$

Repeating the above steps of characterizing the shape of $f(\cdot)$, we can show that $g(\cdot)$ is also quasi-concave, initially increasing from $g(1) = 0$ and eventually dropping toward some $g(1/\phi) < 0$, and has a finite maximum value at the unique ζ that solves $g'(\zeta) = 0$. Thus, as long as (C.10) holds, the unique interior solution $\hat{r}^*$ exists.

For (ii), let us examine the shape of the first-order derivative (C.6) in terms of $\hat{r}_j$: Initially for small $\hat{r}_j$, it might be either positive (small κ) or negative (large κ), then for moderate $\hat{r}_j$ it is positive, and eventually for large $\hat{r}_j$ it becomes negative. (See Figure 3(a), subtracting the flat solid line—marginal cost—from the dashed curve—marginal trading benefit.) In other words, the market maker's objective $\pi_j^{\mathrm{M}}(\hat{r}_j; \hat{R}_{-j}) - \kappa\hat{r}_j\mu$ always has a single-peak in the interior (when the derivative equals zero), but this interior peak may or may not be the global maximum. The competing alternative is the lower corner of $\hat{r}_j = 0$ (the upper corner is never optimal). That is, to ensure the optimality of the interior symmetric solution above, we need to compare $\pi_j^{\mathrm{M}}\big(\hat{r}^*; \hat{R}^*_{-j}\big) - \kappa\hat{r}^*\mu$ and $\pi_j^{\mathrm{M}}\big(0; \hat{R}^*_{-j}\big)$, where we note that symmetry implies $\hat{R}^*_{-j} = (n-1)\hat{r}^*$. As analyzed above, when $\kappa \downarrow 0$, the objective becomes single-peaked, and therefore $\pi_j^{\mathrm{M}}\big(\hat{r}^*; \hat{R}^*_{-j}\big) - \kappa\hat{r}^*\mu > \pi_j^{\mathrm{M}}\big(0; \hat{R}^*_{-j}\big)$. When $\kappa \uparrow \bar{\bar{\kappa}}(m, n)$, the objective becomes monotone decreasing, and therefore the inequality flips. Further, by envelope theorem, we know that $\frac{\mathrm{d}}{\mathrm{d}\kappa}\Big(\pi_j^{\mathrm{M}}\big(\hat{r}^*; \hat{R}^*_{-j}\big) - \kappa\hat{r}^*\mu\Big) = -\hat{r}^*\mu < 0$. Therefore, there exists a unique threshold of $\bar{\kappa}$ ($< \bar{\bar{\kappa}}(m, n)$), defined by

$$\pi_j^{\mathrm{M}}\big(\hat{r}^*; \hat{R}^*_{-j}\big) - \bar{\kappa}\hat{r}^*\mu = \pi_j^{\mathrm{M}}\big(0; \hat{R}^*_{-j}\big), \tag{C.11}$$

so that the interior symmetric solution $\hat{r}^*$ characterized by (C.9) is indeed the global optimal. □

Corollary 2 and 3

Proof. Following the proof of Proposition 2, in the symmetric-strategy equilibrium, the first-order condition becomes (C.9) and must hold if the equilibrium is interior. To sustain the equality, a higher κ (a larger m) must be accompanied by a higher (lower) $g(\zeta) = \frac{2\zeta-1}{\zeta-1}f(\zeta)$. Recall that in the interior equilibrium, $f(\zeta)$ must be decreasing in ζ. Note also that $\frac{2\zeta-1}{\zeta-1}$ is also decreasing in ζ (> 1). Therefore, $g(\cdot)$ is also decreasing in ζ, and a higher κ (a larger m) must result in a strictly lower (higher) ζ. If instead the equilibrium is cornered at $\hat{r}_j = 0$, then ζ does not change with small changes in κ or in m. Therefore, $\frac{\mathrm{d}\zeta}{\mathrm{d}\kappa} \leq 0$ and $\frac{\mathrm{d}\zeta}{\mathrm{d}m} \geq 0$. Since $\zeta = 1 + 1/((m-1)R - 1)$, $\frac{\mathrm{d}R}{\mathrm{d}\kappa} \leq 0$ and $\frac{\mathrm{d}R}{\mathrm{d}m} \geq 0$. Finally, Corollary 1 entails $\frac{\mathrm{d}w}{\mathrm{d}\kappa} \geq 0$ and $\frac{\mathrm{d}w}{\mathrm{d}m} \leq 0$. □

Corollary 4

Proof. Suppose m market makers have entered. We first derive their symmetric net expected profit $\Pi_j^{\mathrm{M}} := \pi_j^{\mathrm{M}} - \kappa\hat{r}_j\mu$, where π_j^{M} is given by (18), as charcterized by Proposition 2. In this symmetric-strategy equilibrium, $\hat{R}_{-j} = (m-1)\hat{r}_j = \frac{m-1}{m}(1-R)$, where $R = \frac{1}{m-1}\frac{\zeta}{\zeta-1}$ as implied by (14). Substituting these into the profit, we obtain

$$\Pi_j^{\mathrm{M}} = \frac{\mu}{\rho\tau_v}\frac{\zeta\left(\zeta^2\phi - \zeta(2\phi+1) + 2\kappa\rho\tau\mathrm{v} + \phi + 2\right) - \kappa\rho\tau\mathrm{v} + (1-\zeta)m((\zeta-1)\zeta\phi - \zeta + \kappa\rho\tau\mathrm{v} + 1) - 1}{(\zeta-1)(m-1)m}.$$

Recall that $\zeta \in (1, 1/\phi)$ is always finite. Hence, in the limit of $m \to \infty$, $\Pi_j^{\mathrm{M}}(m)$ converges to zero. In the other limit, $\Pi_j^{\mathrm{M}}(m_\circ)$ is strictly positive because $\kappa < \bar{\kappa}(m_\circ)$ (Proposition 2). By continuity, therefore, $\max \Pi_j^{\mathrm{M}} > 0$ exists.

A potential entrant expects $\Pi_j^{\mathrm{M}}(m+1) - \kappa_{\mathrm{e}}$. If $\kappa_{\mathrm{e}} \geq \max \Pi_j^{\mathrm{M}}$, then no one will enter, and the equilibrium $m^* = m_\circ$. If instead $\kappa_{\mathrm{e}} < \max \Pi_j^{\mathrm{M}}$, then, by continuity, $\Pi_j^{\mathrm{M}}(m) - \kappa_{\mathrm{e}}$ as a function of m crosses zero from above to below finitely many times (at least once). Let m^* be the largest m at which $\Pi_j^{\mathrm{M}}(m) - \kappa_{\mathrm{e}} = 0$. Then such an m^* constitutes an equilibrium, as no potential entrant would enter. □

Proposition 3

Proof. Absent regulatory requirement, once an order z_i is routed off exchange to a market maker j, the market maker has all the incentive to charge infinity prices to maximize her expected profit. Therefore, that investor i's (expected) trading cost is also infinity, and she is strictly better off directing, instead, the order to the exchange. Therefore, price referencing of $\hat{p}_{ij} = p$ cannot be an equilibrium if there are non-zero off-exchange orders.

The above argument implies that if there is an equilibrium, it must be $\hat{r}_{ij} = 0$ for all $i \in [0, \mu]$ and for all $j \in \{1, ..., m\}$; and consequently, $R = 1$. Proposition 1 then applies. The only remaining equilibrium object is $\hat{p}_{ij}$. To be consistent with $\hat{r}_{ij} = 0$, these off-exchange pricing rules must be worse than the on-exchange price p, i.e., $(\hat{p}_{ij} - p)z_i > 0$. Otherwise, the investor i will have incentive to choose $\hat{r}_{ij} > 0$. □

Corollary 5

Proof. The investors' expected payoff is given by (17), which is monotone decreasing in $\zeta \in (1, \frac{1}{\phi})$. Since illiquidity ζ is the lowest without price referencing, the investors strictly prefer so.

Consider next the market makers. We compare their expected payoffs with vs. without price referencing. **With price referencing** (Proposition 2), they each expect $\pi_j^{\mathrm{M}}(\zeta) - \kappa \hat{r}_j \mu$, with

$$\pi_j^{\mathrm{M}}(\zeta) = \left(1 - \hat{R}_{-j}\right) \frac{(\zeta-1)^2(1-\zeta\phi)}{2\zeta - 1} \frac{\mu}{\rho\tau_v} = (\zeta - 1)(1 - \zeta\phi)\frac{\mu}{m\rho\tau_v},$$

where the first equality uses the functional form of $\pi_j^{\mathrm{M}}(\cdot)$ given by (18); and the second equality uses $\hat{R}_{-j} = \frac{m-1}{m}\hat{R}$ (the symmetry of the equilibrium), $\hat{R} = 1 - R$ (fragmentation identity), and the definition of ζ in (14). We observe three properties (i)–(iii) from the above: Clearly, (i) the function $\pi_j^{\mathrm{M}}(\cdot)$ is hump-shaped in $\zeta \in (1, 1/\phi)$. The market maker then chooses her $\hat{r}_j$ to maximize $\pi_j^{\mathrm{M}}(\zeta) - \kappa\hat{r}_j\mu$, and, assuming $\kappa < \bar{\kappa}$ (Proposition 2), the first-order condition holds:

$$\frac{\partial \pi_j^{\mathrm{M}}}{\partial \zeta}\frac{\mathrm{d}\zeta}{\mathrm{d}R}\frac{\mathrm{d}R}{\mathrm{d}\hat{r}_j} - \kappa\mu = 0 \implies \frac{\partial \pi_j^{\mathrm{M}}}{\partial \zeta} = \kappa\mu\frac{((m-1)R-1)^2}{m-1} > 0;$$

that is, in the equilibrium with price referencing, (ii) the market maker's expected trading profit π_j^{M} must be (locally) increasing with ζ. Finally, by envelope theorem, (iii) $\pi_j^{\mathrm{M}}(\zeta) - \kappa\hat{r}_j\mu$ is strictly decreasing in κ. **Without price referencing** (Proposition 3), each market maker j expects $\pi_j^{\mathrm{M}}(\zeta_\circ)$, where $\zeta_\circ := 1 + 1/(m-2)$ is the lowest level of illiquidity. **To compare the two expected payoffs**, we use the following:

$$\lim_{\kappa \to 0}\left(\pi_j^{\mathrm{M}}(\zeta) - \kappa\hat{r}_j\mu\right) = \lim_{\kappa\to 0}\left(\pi_j^{\mathrm{M}}(\zeta)\right) > \pi_j^{\mathrm{M}}(\zeta) > \pi_j^{\mathrm{M}}(\zeta_\circ),$$

where the first equality holds because $\lim_{\kappa\to 0} \kappa\hat{r}_j\mu = 0$; the first ">" holds because of (iii); and the second ">" holds because of (i) and (ii). Therefore, by continuity, for sufficiently small κ, the expected payoff without price referencing is always higher. □

Propositions 4 and 5

Proof. We prove the more general Proposition 5. Proposition 4 is then obtained as a special case by setting $\hat{\ell} = 0$. We proceed by solving investors', cross market makers', and local market makers' first-order conditions separately, and then jointly determining the linear coefficients. We then determine the cross market makers' Lagrangian coefficients. Finally, we check the second-order conditions.

Investors' market orders. Denote the average coefficients by $\beta := \frac{1}{\mu}\int_0^\mu \beta_i \mathrm{d}i$ and $\gamma := \frac{1}{\mu}\int_0^\mu \gamma_i \mathrm{d}i$. Consider an investor i. From her perspective, the market clearing prices under the linear conjecture (26) are given by

$$p = \bar{v} + (B+C)^{-1}R\mu\beta t \text{ and } \hat{p} = \bar{v} + (\hat{B}+\hat{C})^{-1}\hat{R}\beta t$$

where $t = v - \bar{v} - \gamma u$. Note that since the investor observes both u and v, she equivalently knows p and $\hat{p}$. Therefore, her objective in (23) becomes $(z_i+u)v - (Rp+\hat{R}\hat{p})z_i - \frac{\rho}{2}(z_i+u)^2$, which is concavely quadratic in her order size z_i. The first-order condition thus gives the unique optimal order of

$$z_i = \frac{1}{\rho}(v-\bar{v}) - \left(\frac{R^2}{B+C} + \frac{\hat{R}^2}{\hat{B}+\hat{C}}\right)\mu\beta t - u,$$

thus verifying the linear conjecture. Matching the coefficients, by symmetry, we obtain

$$\beta_i = \beta = \Big((B+C)^{-1}R^2\mu + (\hat{B}+\hat{C})^{-1}\hat{R}^2\mu + \rho\Big)^{-1} \text{ and } \gamma_i = \gamma = \rho. \tag{C.12}$$

Cross market makers' supply schedules. Following the discussion preceding the proposition, a cross market maker j's first-order conditions are given by (24) and (25), with price impacts $\frac{\mathrm{d}p}{\mathrm{d}x_j} = -1/(B_{-j}+C)$ and $\frac{\mathrm{d}\hat{p}}{\mathrm{d}\hat{x}_j} = -1/(\hat{B}_{-j}+\hat{C}_{-j})$. (For now we assume the second-order conditions hold and verify them later.) To verify the linear conjecture, we substitute $\hat{Z} = \hat{R}\mu\beta t$ and then, using the respective market clearing conditions, $t = \frac{B+C}{R\mu\beta}(p-\bar{v})$ for (24) and $t = \frac{\hat{B}+\hat{C}}{\hat{R}\mu\beta}(\hat{p}-\bar{v})$ for (25). This is because we require the supply schedules to be uncontingent, i.e., the on-exchange supply $x_j(\cdot)$ can only depend on p and the off-exchange $\hat{x}_j(\cdot)$ only on $\hat{p}$. This gives

$$x_j = -\frac{(B_{-j}+C)\phi - (R-\omega_j\hat{R})\mu\beta}{(B_{-j}+C)\phi + (R+\omega_j\hat{R})\mu\beta}(B_{-j}+C)(p-\bar{v}) \text{ and } x_j = -\frac{(\hat{B}_{-j}+\hat{C})\phi - (\hat{R}+\omega_j\hat{R})\mu\beta}{(\hat{B}_{-j}+\hat{C})\phi + (\hat{R}-\omega_j\hat{R})\mu\beta}(\hat{B}_{-j}+\hat{C})(\hat{p}-\bar{v}),$$

thus verifying the conjectured linear strategies of $x_j = b_j(p-\bar{v})$ and $\hat{x}_j = \hat{b}_j(\hat{p}-\bar{v})$. Match the coefficients, replace $B_{-j} = B - b_j$ and $\hat{B}_{-j} = \hat{B} - \hat{b}_j$, and solve for $\{b_j, \hat{b}_j\}$ to get

$$b_j = \frac{(B+C)\phi - (R-\omega_j\hat{R})\mu\beta}{(B+C)\phi - 2R\mu\beta}(B+C) \text{ and } \hat{b}_j = \frac{(\hat{B}+\hat{C})\phi - (\hat{R}-\omega_j\hat{R})\mu\beta}{(\hat{B}+\hat{C})\phi - 2\hat{R}\mu\beta}(\hat{B}+\hat{C}).$$

Aggregating the above across all ℓ cross market makers, we obtain

$$B = \frac{(B+C)\phi\ell - (R\ell - \Omega\hat{R})\mu\beta}{(B+C)\phi - 2R\mu\beta}(B+C) \text{ and } \hat{B} = \frac{(\hat{B}+\hat{C})\phi\ell - (\hat{R}\ell - \Omega\hat{R})\mu\beta}{(\hat{B}+\hat{C})\phi - 2\hat{R}\mu\beta}(\hat{B}+\hat{C}), \tag{C.13}$$

where we write $\Omega := \sum_{j=1}^{n} \omega_j$.

Local market makers' supply schedules. The local market makers' problems can be analyzed analogously to the above. Consider an on-exchange local market maker j, who understands her supply y_j affects the on-exchange price p via market clearing according to

$$y_j + \big(B + C_{-j}\big)(p - \bar{v}) = Z \iff p(y_j) = \bar{v} + \big(B + C_{-j}\big)^{-1}\big(Z - y_j\big).$$

Since p perfectly reveals Z, she also infers that $\mathbb{E}[v|\,p] = \mathbb{E}[v|\,Z] = \phi Z$. She then maximizes her objective as in (23) by choosing y_j according to the first-order condition, which yields $y_j = \big(B+C_{-j}\big)(\bar{v}+\phi t - p)$. (Again we assume for now that the second-order condition holds and verify this later.) Substitute $t = \frac{B+C}{R\mu\beta}(p - \bar{v})$ into the above to get

$$y_j = -\frac{(B + C_{-j})\phi - R\mu\beta}{(B + C_{-j})\phi + R\mu\beta}(B + C_{-j})(p - \bar{v}),$$

thus verifying the linear conjecture of $y_j = c_j(p - \bar{v})$. Match the coefficient, replace $C_{-j} = C - c_j$, and solve for c_j to get

$$c_j = \frac{(B+C)\phi - R\beta\mu}{(B+C)\phi - 2R\mu\beta}(B+C) = \frac{C}{\ell}, \tag{C.14}$$

where the last equality holds because c_j is the same across all on-exchange local market makers. Doing the same for off-exchange local market makers, we obtain

$$\hat{c}_j = \frac{(\hat{B}+\hat{C})\phi - \hat{R}\beta\mu}{(\hat{B}+\hat{C})\phi - 2\hat{R}\mu\beta}(\hat{B}+\hat{C}) = \frac{\hat{C}}{\hat{\ell}}. \tag{C.15}$$

Determine all coefficients. We can now jointly solve all coefficients via the system (C.12)–(C.15). There is only one non-zero solution for $\{\beta, \gamma, B, C, \hat{B}, \hat{C}\}$, using which we then also find $\{b_j, c_j, \hat{b}_j, \hat{c}_j\}$ as stated in the proposition. The market clearing conditions (21) then generate the expressions of the two prices.

Check whether (22) is binding. Using the equilibrium price expressions, we obtain

$$(\hat{p} - p)\hat{Z} = \frac{(n + R\ell + \hat{R}\hat{\ell} - 1)\Omega - (\ell - \hat{\ell})R}{(n + \hat{\ell} - 2 + \Omega)((n + \ell - 2)R - \hat{R}\Omega)}\hat{R}\mu\beta\phi t^2.$$

If the constraint is not binding, i.e., if $(\hat{p} - p)\hat{Z} < 0$, then all $\omega_j = 0$ and $\Omega = 0$, so that the above becomes

$$(\hat{p} - p)\hat{Z} = \frac{(\hat{\ell} - \ell)R}{(n + \hat{\ell} - 2)(n + \ell - 2)R}\hat{R}\mu\beta\phi t^2 < 0,$$

which holds if and only if $\ell < \hat{\ell}$, thus yielding the two scenarios given in the proposition. In particular, when $\hat{\ell} < \ell$, i.e., when the constraint binds, setting $\hat{p} - p = 0$ to the above yields (28).

Market makers' second-order conditions. Consider a cross market maker j, whose second-order condition requires the Hessian matrix of her objective function to be negative-definite. Note that both off-diagonal terms are zero (because her demand schedules are uncontingent; see also Rostek and Yoon, 2021). Referring to her first-order conditions (24) and (25), we can compute the two diagonal terms as $-2(B_{-j} + C)^{-1}$ and $-2(\hat{B}_{-j} + \hat{C})^{-1}$, which, under the coefficients found above, become

$$-\frac{1}{B_{-j} + C} = -\frac{n + R\ell + \hat{R}\hat{\ell}}{(n + R\ell + \hat{R}\hat{\ell} - 2)(R + \hat{R}\omega_j)}\frac{\phi}{\mu\beta} \text{ and } -\frac{1}{\hat{B}_{-j} + \hat{C}} = -\frac{n + R\ell + \hat{R}\hat{\ell}}{(n + R\ell + \hat{R}\hat{\ell} - 2)(\hat{R} - \hat{R}\omega_j)}\frac{\phi}{\mu\beta}.$$

The second-order condition requires that both these terms be negative. Note that $\beta > 0$ in equilibrium and recall (c). Therefore, it is equivalent to require

$$R + \hat{R}\omega_j > 0 \text{ and } \hat{R} - \hat{R}\omega_j > 0 \Longrightarrow -\frac{R}{\hat{R}} < \omega_j < 1.$$

But the lower bound is irrelevant because $\omega_j \geq 0$ (for it is the Lagrangian coefficient for the inequality constraint). Hence, the cross market maker's second-order condition holds whenever $0 \leq \omega_j < 1$. Doing the same for local market makers shows that their second-order conditions always hold. □

Corollary 6

Proof. Taking R as an exogenous parameter, (27) implies that $\frac{d\zeta}{dR} < 0$. Substituting $n = m - \ell$, we then have $\frac{d\zeta}{d\ell} > 0$. □

Corollary 7

Proof. Directly comparing the illiquidity levels from (14) and (27) under the respective propositions, replacing $m = n + \ell$, gives

$$\left(1 + \frac{1}{(n+\ell-1)R-1}\right) - \left(1 + \frac{1}{n+R\ell-2}\right) = \frac{(1-R)(n-1)}{((n+\ell-1)R-1)(n+R\ell-2)} > 0.$$

Therefore, the illiquidity is always higher without (Proposition 1) than with off-exchange competition (Proposition 4). As shown in the proof of Corollary 1, w decreases in ζ. Hence, it also follows that w is lower without than with off-exchange competition. □

Corollary 8

Proof. We directly compute the agents' expected trading gains, π_i^{I}, π_{yj}^{M}, $\pi_{\hat{y}j}^{\mathrm{M}}$, and π_{xj}^{M} from their objectives (23) using the equilibrium results from Proposition 4. The total gains from trade is then $w = \ell\pi_{yj}^{\mathrm{M}} + \hat{\ell}\pi_{\hat{y}j}^{\mathrm{M}} + n\pi_{xj}^{\mathrm{M}} + \mu\pi_i^{\mathrm{I}}$. After careful calculation, we obtain

$$\pi_i^{\mathrm{I}} = \frac{1}{2\rho\tau_v\phi}(1-\bar{\zeta}\phi)^2 \text{ and } w = \frac{\mu}{2\rho\tau_v\phi}(1-\bar{\zeta}\phi)(1+(\bar{\zeta}-2)\phi),$$

where $\bar{\zeta} = R\zeta + \hat{R}\hat{\zeta}$. Note that both π_i^{I} and w are strictly decreasing in the average illiquidity $\bar{\zeta}$. The difference now is that the order routing $\{R, \hat{R}\}$ are endogenous. In particular, we know that $R_{\text{no-OPR}} = 1$. Therefore, recalling that $h(\cdot)$ is monotone decreasing, we have $\bar{\zeta}_{\text{no-OPR}} = h(n+\ell) \geq h\Big(n + R_{\text{OPR}}\ell + \big(1 - R_{\text{OPR}}\big)\hat{\ell}\Big) = \bar{\zeta}_{\text{OPR}}$, where the inequality holds for all $R_{\text{OPR}} \in [0,1]$ because $\hat{\ell} < \ell$. We then obtain the stated rankings of the gains from trade. Clearly, the equality holds if and only if $R_{\text{OPR}} = 1$. □

References

Anand, Amber, Mehrdad Samadi, Jonathan Sokobin, and Kumar Venkataraman. 2021. "Institutional Order Handling and Broker-Affiliated Trading Venues." *The Review of Financial Studies* 34 (7):3364–3402.

Antill, Samuel and Darrell Duffie. 2021. "Augmenting Markets with Mechanisms." *Review of oEconomic Studies* 88:1665–1719.

Aramian, Fatemeh and Lars Norden. 2021. "High-Frequency Traders and Single-Dealer Platforms." Working paper.

Babus, Ana and Cecilia Parlatore. 2021. "Strategic Fragmented Markets." *Journal of Financial Economics* Forthcoming.

Baldauf, Markus and Joshua Mollner. 2021. "Trading in Fragmented Markets." *Journal of Financial and Quantitative Analysis* 56 (1):93–121.

Baldauf, Markus, Joshua Mollner, and Bart Zhou Yueshen. 2024. "Siphoned apart: A portfolio perspective on order flow segmentation." *Journal of Financial Economics* 154:153807.

Battalio, Robert, Shane A. Corwin, and Robert Jennings. 2016. "Can Brokers Have It All? On the Relation between Make-Take Fees and Limit Order Execution Quality." *The Journal of Finance* 71 (5):2193–2237.

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 and Robert Jennings. 2023. "Wholesaler Execution Quality." Working paper.

Biais, Bruno and Theirry Foucault. 2014. "HFT and Market Quality." *Bankers, Markets & Investors* (128).

Bryzgalova, Svetlana, Anna Pavlova, and Taisiya Sikorkaya. 2023a. "Retail Trading in Options and the Rise of the Big Three." *The Journal of Finance* 78 (6):3465–3514.

———. 2023b. "Strategic Arbitrage in Segmented Markets." 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.

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

Colliard, Jean-Edouard and Thierry Foucault. 2012. "Trading Fees and Efficiency in Limit Order Markets." *The Review of Financial Studies* 25:3389–3421.

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.

Degryse, Hans, Ivan Markovic, and Gunther Wuyts. 2023. "To protect and serve? Order protection rule, market quality and trading gains." Working paper.

Ding, Shengwei, John Hanna, and Terrence Hendershott. 2014. "How Slow Is the NBBO? A Comparison with Direct Exchange Feeds." *The Financial Review* 49:313–332.

Duffie, Darrell, Piotr Dworczak, and Haoxiang Zhu. 2017. "Benchmarks in Search Markets." *The Journal of Finance* 72 (5):1983–2044.

Duffie, Darrell and Haoxiang Zhu. 2017. "Size Discovery." *The Review of Financial Studies* 30 (4):1095–1150.

Dutt, Hans R. and Lawrence E. Harris. 2005. "Position limits for cash-settled derivative contracts." *Journal of Futures Markets* 25 (10):945–965. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/fut.20179.

Dyhrberg, Anne Haubo, Andriy Shkilko, and Ingrid M. Werner. 2023. "The Retail Execution Quality Landscape." 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.

Ernst, Thomas, Chester Spatt, Andrey Malenko, and Jian Sun. 2024. "What Does Best Execution Look Like?" Working paper.

Ernst, Thomas, Chester Spatt, and Jian Sun. 2022. "Would Order-By-Order Auctions Be Competitive?" Working paper.

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

Financial Times. 2024. "High-speed trader Jane Street raked in $4.4bn at start of 2024." https://www.ft.com/content/7bdd0bc1-5284-4d52-93aa-ddfb609b0bd6.

FINRA. 2014. "Rule 5310 Best Execution and Interpositioning." https://www.finra.org/rules-guidance/rulebooks/finra-rules/5310.

Foley, Sean and Talis J. Putniņš. 2016. "Should we be afraid of the dark? Dark trading and market quality." *Journal of Financial Economics* 122 (3):456–481.

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

Foucault, Thierry, Marco Pagano, and Ailsa Roell. 2013. *Market Liquidity: Theory, Evidence, and Policy*. OUP Catalogue, Oxford University Press.

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

Glosten, Lawrence R. 1994. "Is the Electronic Limit Order Book Inevitable?" *The Journal of Finance* 49 (4):1127–1161.

Griffin, John M. and Amin Shams. 2018. "Manipulation in the VIX?" *The Review of Financial Studies* 31 (4):1376–1417.

Hendershott, Terrence and Charles M. Jones. 2005. "Trade-through prohibitions and market quality."

Journal of Financial Markets 8 (1):1–23.

Hendershott, Terrence and Haim Mendelson. 2000. "Crossing Networks and Dealer Marekts: Competition and Performance." *Journal of Finance* 55 (5):2071–2115.

Hochfelder, David. 2006. ""Where the Common People Could Speculate": The Ticker, Bucket Shops, and the Origins of Popular Participation in Financial Markets, 1880-1920." *The Journal of American History* 93 (2):335–358. URL http://www.jstor.org/stable/4486233.

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

Huang, Xing, Philippe Jorion, Jeongmin Lee, and Christopher Schwarz. 2023. "Who Is Minding the Store? Order Routing and Competition in Retail Trade Execution." Working paper.

Kumar, Praveen and Duane J. Seppi. 1992. "Futures Manipulation with "Cash Settlement"." *The Journal of Finance* 47 (4):1485–1502. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1992.tb04666.x.

Kyle, Albert S. 1989. "Informed Speculation with Imperfect Competition." *Review of Economic Studies* 56 (3):317–356.

Li, Sida, Mao Ye, and Miles Zheng. 2023. "Refusing the best price?" *Journal of Financial Economics* 147 (2):317–337.

Malamud, Semyon and Marzena Rostek. 2017. "Decentralized Exchange." *American Economic Review* 107 (11):3320–3362.

Menkveld, Albert J., Bart Zhou Yueshen, and Haoxiang Zhu. 2017. "Shades of Darkness: A Pecking Order of Trading Venues." *Journal of Financial Economics* 124 (1):503–534.

O'Hara, Maureen and Mao Ye. 2011. "Is Market Fragmentation Harming Market Quality?" *Journal of Financial Economics* 100 (3):459–474.

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.

Parlour, Christine A. and Duane J. Seppi. 2003. "Liquidity-Based Competition for Order Flow." *The Review of Financial Studies* 16 (2):301–343.

Peirce, Hester M. 2022. "Statement on Ordering Competition." https://www.sec.gov/news/statement/peirce-order-competition-20221214.

Rostek, Marzena and Marek Weretka. 2012. "Price Inference in Small Makets." *The Review of Financial Studies* 28 (10):2946–2992.

Rostek, Marzena and Ji Hee Yoon. 2021. "Exchange Design and Efficiency." *Econometrica* 89 (6):2887–2928.

Salop, Steven C. 1986. “Practices that (credibly) facilitate oligopoly co-ordination.” In *New Developments in the analysis of market structure: Proceedings of a conference held by the International Economic Association in Ottawa, Canada*. Springer, 265–294.

Schwarz, Christopher, Brad M. Barber, Xing Huang, Philippe Jorion, and Terrance Odean. 2023. “The “Actual Retail Price” of Equity Trades.” Working paper.

SEC. 2022a. “Fact Sheet: Proposed Rule to Enhance Order Competition.” https://www.sec.gov/files/34-96495-fact-sheet.pdf.

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

Shapiro, Carl. 1989. “Chapter 6 Theories of oligopoly behavior.” Elsevier, 329–414. URL https://www.sciencedirect.com/science/article/pii/S1573448X89010095.

TD Cowen. 2023. “TD Cowen Market Structure: An Update on Liquidity.” Report. https://www.cowen.com/insights/td-cowen-market-structure-an-update-on-liquidity/.

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

Vayanos, Dimitri and Jiang Wang. 2012. “Liquidity and Asset Returns Under Asymmetric Information and Imperfect Competition.” *The Review of Financial Studies* 25 (5):1339–1365.

Virtu. 2021. “Measuring Real Execution Quality: Benefits to Retail Are Significantly Understated.” Report. https://www.sec.gov/comments/265-28/26528-8901054-242178.pdf.

Vives, Xavier. 2011. “Strategic Supply Function Competition with Private Information.” *Econometrica* 79 (6):1911–1966.

Wittwer, Milena. 2021. “Connecting Disconnected Financial Markets?” *American Economic Journal: Microeconomics* 13 (1):252–282.

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.

Ye, Mao and Wei Zhu. 2020. “Strategic Informed Trading and Dark Pools.” Working paper.

Zhang, Anthony Lee. 2022. “Competition and manipulation in derivative contract markets.” *Journal of Financial Economics* 144:396–413.

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