Exchange Competition, Entry, and Welfare[*]

Giovanni Cespa[†] and Xavier Vives[‡]

Tuesday 3rd November, 2020

Abstract

We assess the consequences for market quality and welfare of different entry regimes and exchange pricing policies, integrating a microstructure model with a free-entry, exchange competition model where exchanges have market power in technological services. Free-entry delivers superior liquidity and welfare outcomes vis-à-vis an unregulated monopoly, but entry can be excessive or insufficient. Depending on the extent of the monopolist's technological services undersupply compared to the first best, a planner can achieve a higher welfare controlling entry or platform fees.

Keywords: Market fragmentation, welfare, endogenous market structure, platform competition, Cournot with free entry, industrial organization of exchanges.
JEL Classification Numbers: G10, G12, G14

[*] We thank Fabio Braga, Eric Budish, Thierry Foucault, Itay Goldstein (the Editor), Jiasun Li, Albert Menkveld, Duane Seppi, René Stultz, Felix Suntheim, two anonymous referees and seminar participants at ICEF (Moscow), Boston University, Universidad Carlos III (Madrid), Imperial College (London), Chicago Fed, University of Chicago, Northwestern, Tilburg University, Tsinghua University, NYU (Stern), the 2018 EFA meeting (Warsaw), the 2019 AFA meeting (Atlanta), and the 2020 Econometric Society World Congress (Milan) for useful comments and suggestions. Cespa acknowledges financial support from the Bank of England (grant no. RDC151391). Vives acknowledges the financial support of the Ministry of Science, Innovation and Universities (MCIU)-ECO2015-63711-P and PGC2018-096325-B-I00 (MCIU/AEI/FEDER, UE) and of the Department of Economy and Knowledge of the Generalitat de Catalunya (Ref. 2017 SGR 1244). Orestis Vravosinos and Barna Zsabó provided excellent research assistance.

[†] Cass Business School–City, University of London, and CEPR. 106, Bunhill Row, London EC1Y 8TZ, UK. e-mail: giovanni.cespa@gmail.com

[‡] IESE Business School, Avinguda Pearson, 21 08034 Barcelona, Spain.

> "The result is that, even while one of our fundamental mandates is to encourage competition, the SEC has stood on the sidelines while enormous market power has become concentrated in just a few players... And every time exchanges raise prices [for exchange connections], that money comes out of investors' pockets, who pay more to buy and sell stocks than they otherwise might... In a world where the costs of electronic connections are constantly falling, exchanges have asked us to raise these prices over and over again during the past three years."
> *Unfair Exchange: The State of America's Stock Markets*, SEC Commissioner Robert J. Jackson Jr., September 2018.

1 Introduction

Over the past two decades, governments and regulators moved to foster competition among trading venues. This has spurred market fragmentation, contributing to a drastic reduction in the cost of trading, which has benefited market participants. However, this has also led exchanges to heighten their reliance on revenue generating activities which price they can control better. A case in point is the provision of services such as the sale of market data, co-location space, and fast connections to matching engines.[1] As suggested by the opening quotation, US regulators have voiced their concern over the price of such "technological services," with the SEC alleging that exchanges exercise too much market power in their provision.

When is regulatory intervention warranted? Should a regulator set the price of technological services and if so, how? Do merger policy and the control of exchange entry have a role to play?

We address these issues by modelling liquidity provision as a vertical market where "upstream" exchanges supply *technological services (connectivity)* to "downstream" liquidity providers, who use them to satisfy liquidity traders' demand for immediacy.

[1] "Take, for example, our rules requiring orders to be routed to the exchange that displays the national best bid or offer [... which ensure] that all investors get the benefit of a competitive national market system. When the SEC enacted these rules, [...] there would be cases where brokers would be required to send the order to a specific exchange, leaving the broker–and [...] their customer–exposed to excessive trading fees on that exchange. So we capped the fees the exchanges can charge. But facing a limit on one kind of fee, exchanges may have simply raised other fees, like the cost of connecting to the exchange [...] For example, one exchange, EDGX, has raised the price on its standard 10GB connection five times since 2010–in total, leaving the price of the connection seven times higher than it was in that year." (Robert J. Jackson Jr., 2018)

We then put the model to work by comparing the market solution, with free entry of exchanges, with the second best solutions a regulator can implement. This allows us to analyze the Industrial Organization of stock markets and evaluate the liquidity and welfare effects of different regulatory measures.

We find that the competitive (price-taking) solution is generically *not* efficient, since exchanges only care about the welfare of market participants whose surplus they can appropriate (a vertical externality). Hence, exchanges' market power may improve or worsen welfare compared to the competitive benchmark, and regulation (conduct or structural) can improve upon the market solution. With fee (conduct) regulation it is optimal to have only one exchange; with entry (structural) regulation the market may deliver excessive or insufficient entry. In this context, a connectivity capacity increase (fee reduction) can be achieved either by fostering entry, or by directly imposing it on the regulated monopolist, and the optimal second best regulatory intervention revolves around a simple trade-off. A fee reduction depresses (increases) industry profits (liquidity) to the detriment (benefit) of exchanges (market participants). When the wedge between the first best and monopoly capacity is sufficiently large (small), entry regulation is inferior (superior) to fee regulation.

Thus, our model provides an economics backing to the logic behind the excerpt of SEC Commissioner Robert J. Jackson's speech reported in the opening quotation. Indeed, the vertical structure of the liquidity supply industry *and* exchanges' market power in offering an essential input for liquidity provision explain the *mechanism* by which "[...] every time exchanges raise prices, that money comes out of investors' pockets, who pay more to buy and sell stocks than they otherwise might." In addition, our model suggests *when* raising prices is detrimental or beneficial for overall welfare.

The profit orientation of exchanges, when they converted into publicly listed companies, led to regulatory intervention both in the US (RegNMS in 2005) and the EU (Mifid in 2007), to stem their market power over trading fees. Regulation, together with the liberalization of international capital flows and technological developments, led in turn to an increase in fragmentation and competition among trading platforms. Incumbent exchanges such as the NYSE reacted to increased competition by upgrading technology (e.g, creating, NYSE Arca), or merging with other exchanges (e.g., the NYSE merged with Archipelago in 2005 and with Euronext in 2007. See Foucault et al. (2013), Chapter 1). A relevant fact is that even though there are 13 lit stock venues in the US (and 30 alternative ones), 12 of them, which account for two-thirds of daily trading, are controlled by three major players: Intercontinental Exchange, Nasdaq, and CBOE.

As a result, the trading landscape has changed dramatically. Large-cap stocks nowadays commonly trade in multiple venues, a fact that has led to a decline in incumbents' market shares, giving rise to a "cross-sectional" dimension of market fragmentation (see Appendix B, Figure B.2). The automation of the trading process has also spurred fragmentation along a "time-series" dimension, in that some liquidity providers' market participation is limited (Duffie (2010), SEC (2010)), endogenous (Anand and Venkataraman (2016)), or impaired because of the existence of limits to the access of reliable and timely market information (Ding et al. (2014)).[2] Additionally, trading fees have declined to competitive levels (see, e.g., Foucault et al. (2013), Menkveld (2016), and Budish et al. (2019)), and exchanges have steered their business models towards the provision of technological services.[3]

Such a paradigm shift has raised regulators' and policy makers' concerns for the possibility of monopoly restrictions. Indeed, in August 2020, the SEC has rescinded the rule that allowed exchanges to unilaterally change their "core" data fees and, from September 2020, subjected such changes to public comment and regulatory approval. As a result, the SEC now holds an ex-ante control over exchanges' fee setting process.[4]

[2] Limited market participation of liquidity providers also arises because of shortages of arbitrage capital (Duffie (2010)) and/or traders' inattention or monitoring costs (Abel et al. (2013)).

[3] Increasing competition in trading services has squeezed the profit margins exchanges drew from traditional activities, leading them to gear their business model towards the provision of technological services (Cantillon and Yin (2011)). There is abundant evidence testifying to such a paradigmatic shift. For example, according to the Financial Times, "After a company-wide review Ms Friedman [Nasdaq CEO] has determined the future lies in technology, data and analytics, which collectively accounted for about 35 per cent of net sales in the first half of this year." (see, "Nasdaq's future lies in tech, data and analytics, says Nasdaq CEO" *Financial Times,* October 2017). Additionally, according to Tabb Group, in the US, exchange data, access, and technology revenues have increased by approximately 62% from 2010 to 2015 (Tabb Group, 2016).

[4] In 2018, the SEC sided with market participants over their challenge to the NYSE ARCA's and NASDAQ's decision to increase their data fees (*Statement on Market Data Fees and Market Structure,* SEC Chairman J. Clayton, October 2018, Clayton, 2018, Bloomberg, 2018). Stock exchanges subsequently appealed against the SEC's decision, and in June 2020, won the legal battle: the US Court of Appeals ruled "some fee increases can't be challenged by the government after they have taken effect." (WSJ, 2020). As a response, in August 2020, the SEC modified its regulatory framework, rescinding the rule that allowed exchanges to unilaterally change some of their fees: from September 2020, fee hikes from exchanges for "core" data require public comment and approval from the SEC (Bloomberg, 2020). Prior to this, the SEC did not have the power to approve exchanges' fees beforehand. However, market participants were permitted to challenge them, upon which challenge, the SEC could intervene–a form of *ex-post* fee control. Thus, with this new regulatory approach the SEC now holds an *ex-ante* control over exchanges' fee setting process, Bloomberg, 2020). An important element which secured the success of exchanges' appeal in June 2020, was the fact that the SEC based its 2018 action on a section of law that "makes no mention of fees at all," highlighting the lack of a proper mandate to oversee competition at the core of the US regulator's mission.

We study a framework that integrates a two-period, market microstructure model with one of exchange competition with free entry. The microstructure model defines the *liquidity determination* stage of the game. There, two classes of risk averse dealers provide liquidity to two cohorts of rational liquidity traders, who sequentially enter the market. Dealers in the first class can act at both rounds, absorbing the orders of both liquidity traders' cohorts, and are therefore called 'full' (FD); those in the second class can only act in the first round, and are called 'standard' (SD). The possibility to re-trade captures in a simple way both the limited market participation of SD, and FD's technological superiority to exploit short term return predictability. We assume that a best price rule induces the same transaction price across all competing platforms. This is the case in the US where the combination of the Unlisted Trading Privilege (which allows a security listed on any exchange to be traded by other exchanges) and RegNMS's protection against "trade-throughs" imply that, despite fragmentation, there virtually exists a unique price for each security.[5] We also assume that trading fees are set at the competitive level by the exchanges.[6]

The platform competition model features a finite number of exchanges which, upon incurring a fixed entry cost, offer technological services to FD that allow them to trade in the second round, earning an extra payoff. This defines the *inverse demand* for technological services.[7] Upon entry, each exchange incurs a constant marginal cost to produce a unit of technological service capacity, receiving the corresponding fee from the attracted FD. This defines a Cournot game with free entry which pins down the equilibrium *technological capacity* supply.

Two aspects of our model are worth noting. First, the Cournot specification of the platform game is appropriate since even if there is price competition after the capacity

[5] Price protection rules were introduced to compensate for the potential adverse effects of price fragmentation when the entry of new platforms was encouraged to limit the market power of incumbents. In particular, rule 611 of RegNMS restricts "*trade-throughs* – the execution of trades on one venue at prices that are inferior to publicly displayed quotations on another venue." Additionally, rule 610 disciplines the access to quotations, and sets a cap to the price that can be charged to access such information. The aim is to enforce price priority in all markets (see SEC). However, for large orders execution pricing may not be the same in all exchanges except if traders have in place cross-exchange order-routing technology. In Europe there is no order protection rule similar to RegNMS. Foucault and Menkveld (2008) show empirically the existence of trade-thoroughs in Amsterdam and London markets. Hendershott and Jones (2005) find that in the US price protection rules improve market quality.

[6] We therefore abstract from competition for order flow issues (see Foucault et al. (2013) for an excellent survey of the topic).

[7] Actually, FD may have to invest on their own also on items such as speed technology. In our model we will abstract from such investments.

choice, the strategic variable is costly capacity.[8] Second, ours is a Cournot model with externalities–gross welfare is *not* given by the integral below the inverse demand curve faced by exchanges. This is because platforms' capacity decisions also affect the welfare of market participants other than FD (i.e., SD, and liquidity traders).

We now describe in more detail our main results. Due to their ability to trade at both rounds, FD exhibit a higher risk bearing capacity compared to SD. As a consequence, an increase in their mass improves market liquidity, inducing two contrasting welfare effects. First, it lowers the cost of trading, which leads traders to hedge more aggressively, increasing their welfare. Second, it heightens the competitive pressure faced by SD, lowering their payoffs. As liquidity demand augments for both dealers' classes, however, SD effectively receive a *smaller* share of a *larger* pie. This contributes to make *gross* welfare increasing in the proportion of FD, making liquidity a measurable indicator of gross welfare.

Given the demand for technological services, standard Cournot analysis implies the existence and uniqueness of a symmetric equilibrium in technological capacities which, we verify, is also stable.[9] It then follows that an increase in the number of trading platforms, increases exchanges' technological capacity, lowering the price of technological services, and augmenting the mass of FD. This increases the liquidity of the market and gross welfare. Thus, when the number of platforms is *exogenous*, fostering entry is welfare increasing.

In the last part of the paper, we use our setup to compare the market solution arising with no platform competition (monopoly), and with (Cournot) free entry, with four different planner solutions which vary depending on the planner's restrictions. A planner who chooses the number of competing exchanges and the industry technological service fee, achieves the first best; a planner who can only regulate the technological service fee but not entry, achieves the Conduct second best; finally, if the planner can affect the number of market entrants but not their competitive interaction, she achieves the Structural second best solution (restricted or unrestricted, depending on whether she regulates entry making sure that platforms break even or not).

A monopolistic exchange restricts the supply of technological services to increase

[8] See Kreps and Scheinkman (1983) and Vives (1999).

[9] Our assumptions on exchanges' technology ensure this result. While the symmetry assumption is made for tractability, in light of the exchange industry's evolution over the last ten years, we think that it's not outlandish. Indeed, in 2018, the market shares (based on traded volume) of the three consolidated US exchanges were as follows: NYSE: 22.1 per cent, NASDAQ: 19.5 per cent, and CBOE: 17.8 per cent.

the fees it extracts from FD.[10] Thus, the free entry Cournot equilibrium yields a superior outcome in terms of liquidity and (generally) welfare. However, compared to the structural 2nd best, the market solution can feature excessive or insufficient entry. This is because an exchange's private entry decision does not internalize the profit reduction it imposes on its competitors. Such "profitability depression" is conducive to excessive entry. As platform entry spurs liquidity, however, it also has a positive "liquidity creation" effect which benefits traders, and can offset profitability depression, leading to insufficient entry. Our numerical simulations show that platform entry is often excessive. However, when payoff volatility is low, entry is insufficient for intermediate values of the entry cost. When the entry cost is small, the number of platforms (and the associated total capacity) is high. Thus, profitability depression dominates, and entry is excessive. As the entry cost increases, the two externalities tend to offset each other, eventually leading liquidity creation to dominate, with insufficient entry. Finally, when the entry cost is very large, entry becomes so expensive that the two externalities equilibrate again.

The optimal second best regulatory intervention revolves around a simple trade-off: increasing competition, or lowering the technological service fee, spurs technological capacity production which depresses industry profits while increasing liquidity. When the wedge between first best and monopoly capacity is sufficiently large, entry regulation is inferior. In this case, the large capacity increase required to approach the first best is cheaper to achieve by forcing the monopolistic exchange to charge the lowest, break-even compatible, technological service fee. Conversely, when the wedge between monopolist and first best capacity is small, a smaller increase in technological capacity is required to approach the first best. In this case, the planner may choose to regulate entry, since the fee ensuring a monopolist breaks even yields a large profit depression and a mild market participants' welfare gain. We show that the presence of SD committed to supplying liquidity at each round, can lead to an increase in the equilibrium supply of technological services, prompting a switch in the optimal second best regulatory approach from fee to entry regulation.

The rest of the paper is organized as follows. In the next section, we discuss the literature related to our paper. We then outline the model. In section 4, we turn our attention to study the liquidity determination stage of the game, and in section 5, we analyze the payoffs of market participants, and the demand and supply of technological

[10] In a similar vein, Cespa and Foucault (2014) find that a monopolistic exchange finds it profitable to restrict the access to price data, to increase the fee it extracts from market participants.

services. We then concentrate on the impact of platform competition with free entry, and contrast the effects of different regulatory regimes. A separate section is devoted to discuss 4 extensions of our baseline model (the related technical details are deferred to an Internet Appendix). A final section contains concluding remarks.

2 Literature review

To the best of our knowledge, this paper is the first to analyze the relative merits of different types of regulatory interventions in a single, tractable model of liquidity creation and platform competition in technological services (connectivity). Our paper is thus related to a growing literature on the effects of platform competition and investment in trading technology. Pagnotta and Philippon (2018), consider a framework where trading needs arise from shocks to traders' marginal utilities from asset holding, yielding a preference for different trading speeds. In their model, venues vertically differentiate in terms of speed, with faster venues attracting more speed sensitive investors and charging higher fees. This relaxes price competition, and the market outcome is inefficient. The entry welfare tension in their case is between business stealing and quality (speed) diversity, like in the models of Gabszewicz and Thisse (1979) and Shaked and Sutton (1982). In this paper, as argued above, the welfare tension arises instead from the profitability depression and liquidity creation effects associated with entry.[11] Biais et al. (2015) study the welfare implications of investment in the acquisition of High Frequency Trading (HFT–we will use HFT to also indicate High Frequency Traders) technology. In their model HFTs have a superior ability to match orders, and possess superior information compared to human (slow) traders. They find excessive incentives to invest in HFT technology, which, in view of the negative externality generated by HFT, can be welfare reducing. Budish et al. (2015) argue that HFT thrives in the continuous limit order book (CLOB), which is however a flawed market structure since it generates a socially wasteful arms' race to respond faster to (symmetrically observed)

[11] Pagnotta and Philippon (2018) also study the market integration impact of RegNMS. Pagnotta (2013) studies the interaction between traders' participation decisions and venues' investment in speed technology, analysing the implications of institutions' market power for market liquidity and the level of asset prices. Babus and Parlatore (2017) find that market fragmentation arises in equilibrium when the private valuations of different investors are sufficiently correlated. Malamud and Rostek (2017) and Manzano and Vives (2018) look also at whether strategic traders are better off in centralized or segmented markets. Chen and Duffie (2020) show that the fragmentation of a single asset trading activity across different venues, improves the rebalancing of traders' positions, as well as the overall informational content of the asset prices.

public signals. The authors advocate a switch to Frequent Batch Auctions (FBA) instead of a continuous market. Budish et al. (2019), introduce exchange competition in Budish et al. (2015) and analyze whether exchanges have incentives to implement the technology required to run FBA. Also building on Budish et al. (2015), Baldauf and Mollner (2017) show that heightened exchange competition has two countervailing effects on market liquidity, since it lowers trading fees, but magnifies the opportunities for cross-market arbitrage, increasing adverse selection. Menkveld and Zoican (2017) show that the impact of a speed enhancing technology on liquidity depends on the news-to-liquidity trader ratio. Indeed, on the one hand, as in our context, higher speed enhances market makers' risk sharing abilities. On the other hand, it increases liquidity providers' exposure to the risk that high frequency speculators exploit their stale quotes. Finally, Huang and Yueshen (2020) analyse speed and information acquisition decisions, assessing their impact on price informativeness, and showing that in equilibrium these can be complements or substitutes. None of the above papers contrasts the impact of different types of regulatory intervention for platforms' investment in technology, market liquidity, and market participants' welfare.

Our paper is also related to the literature on the Industrial Organization of securities' trading. This literature has identified a number of important trade-offs due to competition among trading venues. On the positive side, platform competition exerts a beneficial impact on market quality because it forces a reduction in trading fees (Foucault and Menkveld (2008) and Chao et al. (2019)), and can lead to improvements in margin requirements (Santos and Scheinkman (2001)); furthermore, it improves trading technology and increases product differentiation, as testified by the creation of "dark pools." On the negative side, higher competition can lower the "thick" market externalities arising from trading concentration (Chowdhry and Nanda (1991) and Pagano (1989)), and increase adverse selection risk for market participants (Dennert (1993)). We add to this literature, by pointing out that market incentives may be insufficient to warrant a welfare maximizing solution. Indeed, heightened competition can lead to the entry of a suboptimal number of trading venues, because of the conflicting impact of entry on profitability and liquidity.

Finally, our results speak to the theoretical Industrial Organization literature on the Cournot model with free entry. Mankiw and Whinston (1986) obtain an excessive entry result in the standard Cournot model due to a business stealing effect (i.e., individual output being decreasing in the number of firms) which leads to the profitability depressing effect of entry to dominate.[12] Ghosh and Morita (2007) obtain insufficient entry in a vertical oligopoly when the downstream sector is sufficiently imperfectly competitive. In a vertical oligopoly, increased upstream entry lowers the price of the intermediate input used by the downstream firms which, as long as these hold market power, leads to business creation and an increase in surplus. With a perfectly competitive downstream sector, such effects disappear eliminating the positive welfare externality due to upstream entry. In our model, even though liquidity providers are competitive, upstream entry induces a positive externality by increasing the mass of FD which improves risk sharing and the welfare of liquidity traders, potentially leading to insufficient entry.

3 The model

A single risky asset with liquidation value $v \sim N(0, \tau_v^{-1})$, and a risk-less asset with unit return are exchanged during two trading rounds.

Three classes of traders are in the market. First, a continuum of competitive, risk-averse, "Full Dealers" (denoted by FD) in the interval $(0, \mu)$, who are active at both rounds. Second, competitive, risk-averse "Standard Dealers" (denoted by SD) in the interval $[\mu, 1]$, who instead are active only in the first round. Finally, a unit mass of traders who enter at date 1, taking a position that they hold until liquidation. At date 2, a new cohort of traders (of unit mass) enters the market, and takes a position. The asset is liquidated at date 3.

This liquidity provision model captures in a parsimonious way a setup where FD possess an edge over SD along two related dimensions: first they trade "faster" in that they can quickly turn around their first period position, re-trading at the second round, facing no competition from SD; second, anticipating this possibility, they are able to better manage their first-round inventory, increasing their profit from liquidity supply. Both these features liken FDs to High Frequency Traders.[13] Additionally,

[12] Except for the integer problem, insufficient entry can occur by at most one firm.

[13] The literature on High Frequency Trading has identified a number of characteristics of these market participants. The SEC (2010) in a 2010 concept release on market structure argues: "Other characteristics often attributed to proprietary firms engaged in HFT are: (1) the use of extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders; (2) use of co-location services and individual data feeds offered by exchanges and others to minimize network and other types of latencies; (3) very short timeframes for establishing and liquidating positions. . . ." This view is also shared by Brogaard (2010), who defines high frequency trading as ". . . a type of strategy that is engaged in buying and selling shares rapidly, often in terms of

since in our framework all market participants' trading needs are endogenous, we are able to perform welfare analysis.

A model captures the reality we observe in a stylized manner and is thus likely to miss some of its important aspects. This is why we consider three alternative ways to model the divide between FD and SD. In section 1 of the Internet Appendix, similarly to Huang and Yueshen (2020), we assume that SD enter the market at the second round. This assumption puts front and center the speed difference between these two types of traders. In section 2, we assume that a fixed mass of SD is in the market at both rounds. This assumption relaxes the monopolistic power over liquidity supply that FD enjoy in our baseline model. In section 3, we endogenize the mass of dealers who are active in the market by studying the effect of allowing potential dealers to decide whether to enter the intermediation industry prior to making the decision to become FD. Finally, in section 4, we study the effect of having second period traders observe a noisy signal of the first period endowment shock. Section 7 summarizes the results obtained in the above extensions.

3.1 Trading venues

The organization of the trading activity depends on the competitive regime among venues. With a monopolistic exchange, both trading rounds take place on the same venue. When platforms are allowed to compete for the provision of technological services, we assume that a best price rule ensures that the price at which orders are executed is the same across all venues. We thus assume away "cross-sectional" frictions, implying that we have a virtual single platform where all exchanges provide identical access to trading, and stock prices are determined by aggregate market clearing.[14]

We model trading venues as platforms that prior to the first trading round (date 0), supply technology which offers market participants the possibility to trade in the second period. For example, it is nowadays common for exchanges to invest in the supply of co-location facilities which they rent out to traders, to store their servers and

milliseconds and seconds." See also Hasbrouck and Saar (2013) and Aït-Sahalia and Saglam (2013) for similar definitions. As will become clear in Section 3.1, we allow dealers to improve their performance via the purchase of technological services sold by exchanges, while abstracting from modeling other forms of technological investment on their part.

[14] Holden and Jacobsen (2014) find that in the US, only 3.3% of all trades take place outside the NBBO (NBBO stands for "National Best Bid and Offer," and is a SEC regulation ensuring that brokers trade at the best available ask and bid (resp. lowest and highest) prices when trading securities on behalf of customers). See also Li (2015) for indirect evidence that the single virtual platform assumption is compelling on non-announcement days.

networking equipment close to the matching engine; additionally, platforms invest in technologies that facilitate the distribution of market data feeds.[15] In the past, when trading was centralized in national venues, exchanges invested in real estate and the facilities that allowed dealers and floor traders to participate in the trading process.

At date $t = -1$, trading venues decide whether to enter and if so they incur a fixed cost $f > 0$. Suppose that there are N entrants, that each venue $i = 1, 2, \ldots, N$ produces a technological service capacity μ_i, and that $\sum_{i=1}^{N} \mu_i = \mu$, so that the proportion of FD coincides with the total technological service capacity offered by the platforms. Consistent with the evidence discussed in the introduction (see also Menkveld (2016)), we assume that trading fees are set to the competitive level.

3.2 Liquidity providers

A FD has CARA preferences, with risk-tolerance γ, and submits price-contingent orders x_t^{FD}, to maximize the expected utility of his final wealth: $W^{FD} = (v - p_2)x_2^{FD} + (p_2 - p_1)x_1^{FD}$, where p_t denotes the equilibrium price at date $t \in \{1, 2\}$.[16] A SD also has CARA preferences with risk-tolerance γ, but is in the market only in the first period. He thus submits a price-contingent order x_1^{SD} to maximize the expected utility of his wealth $W^{SD} = (v - p_1)x_1^{SD}$. Therefore, FD as SD observe p_1 at the first round; furthermore, FD also observe p_2, so that their information set at the second round is given by $\{p_1, p_2\}$.

The inability of a SD to trade in the second period is a way to capture limited market participation in our model. In today's markets, this friction could be due to technological reasons, as in the case of standard dealers with impaired access to a technology that allows trading at high frequency. In the past, two-tiered liquidity provision occurred because only a limited number of market participants could be physically present in the exchange to observe the trading process and react to demand

[15] Empirical evidence shows that co-location can have a positive impact on traders' profits and market quality. For example, according to Baron et al. (2019), HFTs that improve their latency rank due to co-location upgrades enjoy improved trading performance. The stronger performance associated with speed comes through both the short-lived information channel and the risk management channel, and speed is useful for various strategies including market making and cross-market arbitrage. Similarly, exploiting an optional colocation upgrade at NASDAQ Stockholm, Brogaard et al. (2015) show that traders who upgrade, use their enhanced speed to reduce their exposure to adverse selection and to relax their inventory constraints (reduced sensitivity to inventory position). As a result, they increase their presence at the BBO, with a beneficial effect on effective spreads.

[16] We assume, without loss of generality with CARA preferences, that the non-random endowment of FD and dealers is zero. Also, as equilibrium strategies will be symmetric, we drop the subindex i.

shocks.

3.3 Liquidity demanders

Liquidity traders have CARA preferences, with risk-tolerance γ^L. In the first period a unit mass of traders enters the market. A trader receives a random endowment of the risky asset u_1 and submits an order x_1^L in the asset that he holds until liquidation.[17] A first period trader posts a market order x_1^L to maximize the expected utility of his profit $\pi_1^L = u_1 v + (v - p_1)x_1^L$: $E[-\exp\{-\pi_1^L/\gamma^L\}|u_1]$. In period 2, a new unit mass of traders enters the market. A second period trader observes p_1 (and can thus perfectly infer u_1), receives a random endowment of the risky asset u_2, and posts a market order x_2^L to maximize the expected utility of his profit $\pi_2^L = u_2 v + (v - p_2)x_2^L$: $E[-\exp\{-\pi_2^L/\gamma^L\}|p_1, u_2]$. We assume that $u_t \sim N(0, \tau_u^{-1})$, $\text{Cov}[u_t, v] = \text{Cov}[u_1, u_2] = 0$. To ensure that the payoff functions of the liquidity demanders are well defined (see Section 5.1), we impose

$$(\gamma^L)^2 \tau_u \tau_v > 1, \tag{1}$$

an assumption that is common in the literature (see, e.g., Vayanos and Wang (2012)).

3.4 Market clearing and prices

Market clearing in periods 1 and 2 is given respectively by $x_1^L + \mu x_1^{FD} + (1-\mu)\, x_1^{SD} = 0$ and $x_2^L + \mu(x_2^{FD} - x_1^{FD}) = 0$. We restrict attention to linear equilibria where

$$p_1 = -\Lambda_1 u_1 \tag{2a}$$

$$p_2 = -\Lambda_2 u_2 + \Lambda_{21} u_1, \tag{2b}$$

where the price impacts of endowment shocks Λ_1, Λ_2, and Λ_{21} are determined in equilibrium. According to (2a) and (2b), at equilibrium, observing p_1 and the sequence $\{p_1, p_2\}$ is informationally equivalent to observing u_1 and the sequence $\{u_1, u_2\}$.

The model thus nests a standard stock market trading model in one of platform competition. Figure 1 displays the timeline of the model.

[17] Recent research documents the existence of a sizeable proportion of market participants who do not rebalance their positions at every trading round (see Heston et al. (2010), for evidence consistent with this type of behavior at an intra-day horizon).

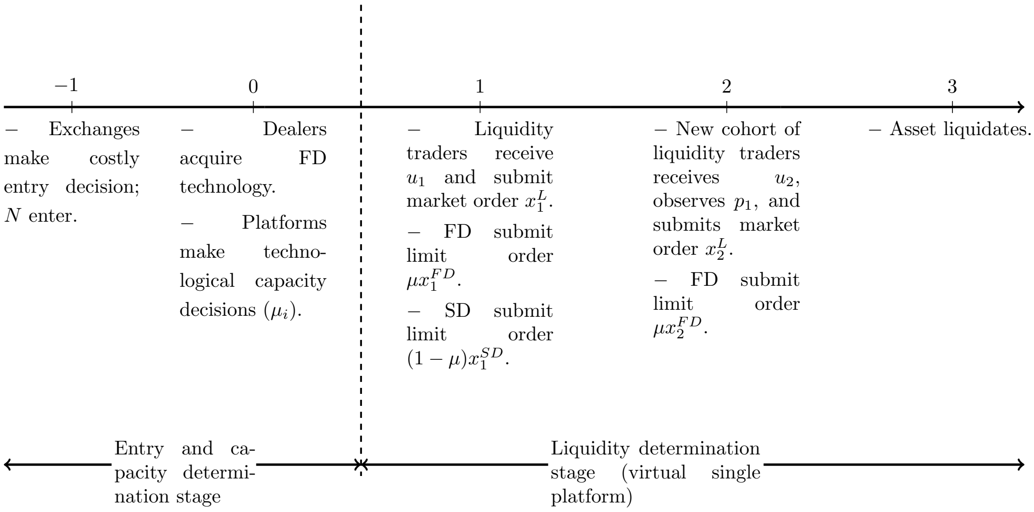

−1
0
1
2
3
− Exchanges make costly entry decision; N enter.
− Dealers acquire FD technology.
− Platforms make technological capacity decisions (μ_i).
− Liquidity traders receive u_1 and submit market order x_1^L.
− FD submit limit order μx_1^{FD}.
− SD submit limit order $(1-\mu)x_1^{SD}$.
− New cohort of liquidity traders receives u_2, observes p_1, and submits market order x_2^L.
− FD submit limit order μx_2^{FD}.
− Asset liquidates.
Entry and capacity determination stage
Liquidity determination stage (virtual single platform)

Figure 1: The timeline.

4 Stock market equilibrium

In this section we assume that a positive mass $\mu \in (0,1]$ of FD is in the market, and present a simple two-period model of liquidity provision à la Grossman and Miller (1988) where dealers only accommodate endowment shocks but where all traders are expected utility maximizers.

Proposition 1. *For $\mu \in (0,1]$, there exists a unique equilibrium in linear strategies in the stock market, where $x_1^{SD} = -\gamma\tau_v p_1$, $x_1^{FD} = \gamma\tau_u\Lambda_2^{-2}(\Lambda_{21}+\Lambda_1)u_1 - \gamma\tau_v p_1$, $x_2^{FD} = -\gamma\tau_v p_2$, $x_1^L = a_1 u_1$, $x_2^L = a_2 u_2 + b u_1$, and prices are given by* (2a) *and* (2b)*,*

$$\Lambda_1 = \left(1 - \left(1 + a_1 + \mu\gamma\tau_u \frac{\Lambda_{21}+\Lambda_1}{\Lambda_2^2}\right)\right)\frac{1}{\gamma\tau_v} > 0 \tag{3a}$$

$$\Lambda_2 = -\frac{a_2}{\mu\gamma\tau_v} > 0 \tag{3b}$$

$$\Lambda_{21} = -(1 - ((1-\mu)\gamma + \gamma^L)\tau_v\Lambda_1)\Lambda_2 < 0 \tag{3c}$$

$$a_t = \gamma^L\tau_v\Lambda_t - 1 \in (-1,0) \tag{3d}$$

$$b = -\gamma^L\tau_v\Lambda_{21} \in (0,1), \tag{3e}$$

and

$$\Lambda_{21} + \Lambda_1 > 0. \tag{4}$$

The coefficient Λ_t in (2a) and (2b) denotes the period t endowment shock's negative price impact, and is our (inverse) measure of liquidity:

$$\Lambda_t = -\frac{\partial p_t}{\partial u_t}. \tag{5}$$

As we show in Appendix A (see (A.3), and (A.14)), a trader's order is given by

$$X_1^L(u_1) = \underbrace{\gamma^L \frac{E[v - p_1|u_1]}{\text{Var}[v - p_1|u_1]}}_{\text{Speculation}} \underbrace{- u_1}_{\text{Hedging}}$$

$$X_2^L(u_1, u_2) = \underbrace{\gamma^L \frac{E[v - p_2|u_1, u_2]}{\text{Var}[v - p_2|u_1, u_2]}}_{\text{Speculation}} \underbrace{- u_2}_{\text{Hedging}}.$$

According to (3d), a trader speculates and hedges his position to avert the risk of a decline in the endowment value occurring when the return from speculation is low ($a_t \in (-1, 0)$). We will refer to $|a_t|$ as the trader's "trading aggressiveness." Additionally, according to (3e), second period traders put a positive weight b on the first period endowment shock. SD and FD provide liquidity, taking the other side of traders' orders. In the first period, standard dealers earn the spread from loading at p_1, and unwinding at the liquidation price. FD, instead, also speculate on short-term returns. Indeed,

$$x_1^{FD} = \gamma \frac{E[p_2 - p_1|u_1]}{\text{Var}[p_2|u_1]} - \gamma \tau_v p_1.$$

To interpret the above expression, suppose $u_1 > 0$. Then, liquidity traders sell the asset, depressing its price (see (2a)) and leading both FD and SD to provide liquidity, taking the other side of the trade. SD hold their position until the liquidation date, whereas FD have the opportunity to unwind it at the second round, partially unloading their inventory risk. Anticipating this, second period traders buy the security (or reduce their short-position), which explains the positive sign of the coefficient b in their strategy (see (3e)). This implies that $E[p_2 - p_1|u_1] = (\Lambda_{21} + \Lambda_1)u_1 > 0$, so that FD anticipate a positive speculative short-term return from going long in the asset.

In sum, FD supply liquidity both by posting a limit order, and a contrarian market order at the equilibrium price, to exploit the predictability of short term returns.[18] In

[18] This is consistent with the evidence on HFT liquidity supply (Brogaard et al. (2014), and Biais et al. (2015)), and on their ability to predict returns at a short term horizon based on market data (Harris and Saad (2014), and Menkveld (2016)).

view of this, Λ_1 in (3a) reflects the risk compensation dealers require to hold the portion of u_1 that first period traders hedge and FD do not absorb via speculation:

$$\Lambda_1 = \left(1 - \left(\underbrace{1 + a_1}_{\text{L1 holding of } u_1} + \underbrace{\mu\gamma\tau_u \frac{\Lambda_{21} + \Lambda_1}{\Lambda_2^2}}_{\text{FD aggregate speculative position}} \right)\right) \frac{1}{\gamma\tau_v}.$$

In the second period, liquidity traders hedge a portion a_2 of their order, which is absorbed by a mass μ of FD, thereby explaining the expression for Λ_2 in (3b).

Therefore, at both trading rounds, an increase in μ, or in dealers' risk tolerance, increases the risk bearing capacity of the market, leading to a higher liquidity:

Corollary 1. *An increase in the proportion of FD, or in dealers' risk tolerance increases the liquidity of both trading rounds: $\partial\Lambda_t/\partial\mu < 0$, and $\partial\Lambda_t/\partial\gamma < 0$ for $t \in \{1, 2\}$.*

According to (2b) and (3c), due to FD short term speculation, the first period endowment shock has a persistent impact on equilibrium prices: p_2 reflects the impact of the imbalance FD absorb in the first period, and unwind to second period traders. Indeed, substituting (3c) in (2b), and rearranging yields: $p_2 = -\Lambda_2 u_2 + \Lambda_2(-\mu x_1^{FD})$.

Corollary 2. *First period traders hedge the endowment shock more aggressively than second period traders: $|a_1| > |a_2|$. Furthermore, $|a_t|$ and b are increasing in μ.*

Comparing dealers' strategies shows that SD in the first period trade with the same intensity as FD in the second period. In view of the fact that in the first period the latter provide additional liquidity by posting contrarian market orders, this implies that $\Lambda_1 < \Lambda_2$, explaining why traders display a more aggressive hedging behavior in the first period. The second part of the above result reflects the fact that an increase in μ improves liquidity at both dates, but also increases the portion of the first period endowment shock absorbed by FD. This, in turn, leads second period liquidity traders to step up their response to u_1.

Summarizing, an increase in μ has two effects: it heightens the risk bearing capacity of the market, and it strengthens the propagation of the first period endowment shock to the second trading round. The first effect makes the market deeper, leading traders to step up their hedging aggressiveness. The second effect reinforces second period traders' speculative responsiveness. When all dealers are FD, liquidity is maximal.

5 Traders' welfare, technology demand, and exchange equilibrium

In this section we study traders' payoffs, derive demand and supply for technological services, and obtain the platform competition equilibrium.

5.1 Traders' payoffs and the liquidity externality

We measure a trader's payoff with the certainty equivalent of his expected utility: $CE^{FD} \equiv -\gamma \ln(-EU^{FD})$, $CE^{SD} \equiv -\gamma \ln(-EU^{SD})$, $CE_t^L \equiv -\gamma^L \ln(-EU_t^L)$, $t \in \{1,2\}$, where EU^j, $j \in \{SD, FD\}$ and EU_t^L, $t \in \{1,2\}$, denote respectively the unconditional expected utility of a standard dealer, a full dealer, and a first and second period trader. The following results present explicit expressions for the certainty equivalents.

Proposition 2. *The payoffs of a SD and a FD are given by*

$$CE^{SD} = \frac{\gamma}{2} \ln\left(1 + \frac{\text{Var}[E[v - p_1|p_1]]}{\text{Var}[v - p_1|p_1]}\right) \tag{6a}$$

$$CE^{FD} = \frac{\gamma}{2}\Bigg(\ln\left(1 + \frac{\text{Var}[E[v - p_1|p_1]]}{\text{Var}[v - p_1|p_1]} + \frac{\text{Var}[E[p_2 - p_1|p_1]]}{\text{Var}[p_2 - p_1|p_1]}\right) + \ln\left(1 + \frac{\text{Var}[E[v - p_2|p_1, p_2]]}{\text{Var}[v - p_2|p_1, p_2]}\right)\Bigg). \tag{6b}$$

Furthermore:

1. *For all $\mu \in (0,1]$, $CE^{FD} > CE^{SD}$.*
2. *CE^{SD} and CE^{FD} are decreasing in μ.*
3. *$\lim_{\mu\to1} CE^{FD} > \lim_{\mu\to0} CE^{SD}$.*

According to (6a) and (6b), dealers' payoffs reflect the accuracy with which these agents anticipate their strategies' unit profits. A SD only trades in the first period, and the accuracy of his unit profit forecast is given by $\text{Var}[E[v - p_1|p_1]]/\text{Var}[v - p_1|p_1]$ (the ratio of the variance explained by p_1 to the variance unexplained by p_1).

A FD instead trades at both rounds, supplying liquidity to first period traders, as a SD, but also absorbing second period traders' orders, and taking advantage of short-term return predictability. Therefore, his payoff reflects the same components of that of a SD, and also features the accuracy of the unit profit forecast from short term speculation ($\text{Var}[E[p_2 - p_1|p_1]]/\text{Var}[p_2 - p_1|p_1]$), and second period liquidity supply ($\text{Var}[E[v - p_2|p_1, p_2]]/\text{Var}[v - p_2|p_1, p_2]$). In sum, as FD can trade twice, benefiting from more opportunities to speculate and share risk, they enjoy a higher expected utility.

Substituting (3d) and (3e) in (6a) and (6b), and rearranging yields:

$$CE^{SD} = \frac{\gamma}{2}\ln\left(1 + \frac{(1+a_1)^2}{(\gamma^L)^2\tau_u\tau_v}\right) \tag{7a}$$

$$CE^{FD} = \frac{\gamma}{2}\Bigg(\ln\left(1 + \frac{(1+a_1)^2}{(\gamma^L)^2\tau_u\tau_v} + \left(\frac{1+a_1}{1+\mu\gamma\tau_u\tau_v(\mu\gamma+\gamma^L)}\right)^2\right) \tag{7b}$$
$$+ \ln\left(1 + \frac{(1+a_2)^2}{(\gamma^L)^2\tau_u\tau_v}\right)\Bigg).$$

An increase in μ has two offsetting effects on the above expressions for dealers' welfare. On the one hand, as it boosts market liquidity, it leads traders to hedge more, increasing dealers' payoffs (Corollaries 1 and 2). On the other hand, as it induces more competition to supply liquidity it lowers them. The latter effect is stronger than the former. Importantly, even in the extreme case in which $\mu = 1$, a FD receives a higher payoff than a SD in the polar case $\mu \approx 0$.

Proposition 3. *The payoffs of first and second period traders are given by*

$$CE_1^L = \frac{\gamma^L}{2}\ln\left(1 + \frac{\text{Var}[E[v-p_1|p_1]]}{\text{Var}[v-p_1|p_1]} + 2\frac{\text{Cov}[p_1, u_1]}{\gamma^L}\right) \tag{8a}$$

$$CE_2^L = \frac{\gamma^L}{2}\ln\Bigg(1 + \frac{\text{Var}[E[v-p_2|p_1,p_2]]}{\text{Var}[v-p_2|p_1,p_2]} + \tag{8b}$$
$$2\frac{\text{Cov}[p_2, u_2|p_1]}{\gamma^L} + \frac{\text{Var}[E[v-p_2|p_1]]}{\text{Var}[v]} - \left(\frac{\text{Cov}[p_2,u_1]}{\gamma^L}\right)^2\Bigg).$$

Furthermore:

1. *CE_1^L and CE_2^L are increasing in μ.*
2. *For all $\mu \in (0, 1]$, $CE_1^L > CE_2^L$.*

Similarly to SD, liquidity traders only trade once (either at the first, or at the second round). This explains why their payoffs reflect the precision with which they can anticipate the unit profit from their strategy (see (8a) and (8b)). Differently from SD, these traders are however exposed to a random endowment shock. As a less liquid market increases hedging costs, it negatively affects their payoff ($\text{Cov}[p_1, u_1] = -\Lambda_1 \tau_u^{-1}$, and $\text{Cov}[p_2, u_2|p_1] = -\Lambda_2 \tau_u^{-1}$). Finally, (8b) shows that a second period trader benefits when the return he can anticipate based on u_1 is very volatile compared to v ($\text{Var}[E[v - p_2|p_1]]/\text{Var}[v]$), since this indicates that he can speculate on the propagated endowment shock at favorable prices. However, a strong speculative activity reinforces the relationship between p_2 and u_1, ($\text{Cov}[p_2, u_1]^2$), leading a trader to hedge little of his endowment shock u_2, and keep a large exposure to the asset risk, thereby reducing his payoff.

Substituting (3d) and (3e) in (8a) and (8b), and rearranging yields:

$$CE_1^L = \frac{\gamma^L}{2} \ln \left(1 + \frac{a_1^2 - 1}{(\gamma^L)^2 \tau_u \tau_v} \right) \tag{9}$$

$$CE_2^L = \frac{\gamma^L}{2} \ln \left(1 + \frac{a_2^2 - 1}{(\gamma^L)^2 \tau_u \tau_v} + \frac{b^2((\gamma^L)^2 \tau_u \tau_v - 1)}{(\gamma^L)^4 \tau_u^2 \tau_v^2} \right). \tag{10}$$

An increase in the proportion of FD makes the market more liquid and leads traders to hedge and speculate more aggressively (Corollary 2), benefiting first period traders (Proposition 3). At the same time, it heightens the competitive pressure faced by SD, lowering their payoffs (Proposition 2). As liquidity demand augments for both dealers' classes, however, SD effectively receive a *smaller* share of a *larger* pie. This mitigates the negative impact of increased competition, implying that on balance the positive effect of the increased liquidity prevails:

Corollary 3. *The positive effect of an increase in the proportion of FD on first period traders' payoffs is stronger than its negative effect on SD' welfare:*

$$\frac{\partial CE_1^L}{\partial \mu} > -\frac{\partial CE^{SD}}{\partial \mu}, \tag{11}$$

for all $\mu \in (0, 1]$.

Aggregating across market participants' welfare yields the following Gross Welfare function:

$$\begin{aligned} GW(\mu) &= \mu CE^{FD} + (1-\mu)CE^{SD} + CE_1^L + CE_2^L \\ &= \underbrace{\mu(CE^{FD} - CE^{SD})}_{\text{Surplus earned by FD}} + \underbrace{CE^{SD} + CE_1^L + CE_2^L}_{\text{Welfare of other market participants}} \end{aligned} \tag{12}$$

Corollary 4.

1. *The welfare of market participants other than FD is increasing in μ.*
2. *Gross welfare is higher at $\mu = 1$ than at $\mu \approx 0$.*

The first part of the above result is a direct consequence of Corollary 3: as μ increases, SD's losses due to heightened competition are more than compensated by traders' gains due to higher liquidity. The second part, follows from Proposition 2 (part 3), and Proposition 3. Note that it rules out the possibility that the payoff decline experienced by FD as μ increases, leads gross welfare to be higher at $\mu \approx 0$. Therefore, a solution that favors liquidity provision by FD is also in the interest of all market participants. Finally, we have:

Numerical Result 1. *Numerical simulations show that $GW(\mu)$ is monotone in μ. Therefore, $\mu = 1$ is the unique maximum of the gross welfare function $GW(\mu)$.*

In view of Corollary 1, gross welfare is maximal when liquidity is at its highest level.[19] Furthermore, because of monotonicity, the above market quality indicator, becomes "measurable" welfare indexes.

[19] Numerical simulations where conducted using the following grid: $\gamma, \mu \in \{0.01, 0.02, \ldots, 1\}$, $\tau_u, \tau_v \in \{1, 2, \ldots, 10\}$, and $\gamma^L \in \{1/\sqrt{\tau_u\tau_v} + 0.001, 1/\sqrt{\tau_u\tau_v} + 0.101, \ldots, 1\}$, in order to satisfy (1).

5.2 The demand for technological services

We define the value of becoming a FD as the extra payoff that such a dealer earns compared to a SD. According to (6a) and (6b), this is given by:

$$\begin{aligned}\phi(\mu) &\equiv CE^{FD} - CE^{SD} \qquad (13)\\ &= \frac{\gamma}{2}\Bigg(\underbrace{\ln\left(1+\frac{\text{Var}[E[v-p_1|p_1]]}{\text{Var}[v-p_1|p_1]}+\frac{\text{Var}[E[p_2-p_1|p_1]]}{\text{Var}[p_2-p_1|p_1]}\right)-\ln\left(1+\frac{\text{Var}[E[v-p_1|p_1]]}{\text{Var}[v-p_1|p_1]}\right)}_{\text{Competition}}\\ &\qquad +\underbrace{\ln\left(1+\frac{\text{Var}[E[v-p_2|p_1,p_2]]}{\text{Var}[v-p_2|p_1,p_2]}\right)}_{\text{Liquidity supply}}\Bigg).\end{aligned}$$

FD rely on two sources of value creation: first, they compete business away from SD, extracting a larger rent from their trades with first period traders (since they can supply liquidity and speculate on short-term returns); second, they supply liquidity to second period traders.

We interpret the function $\phi(\mu)$ as the (inverse) demand for technological services as it is the willingness to pay to become a FD.[20]

Corollary 5. *The inverse demand for technological services $\phi(\mu)$ is decreasing in μ.*

A marginal increase in μ heightens the competition FD face among themselves, and vis-à-vis SD. The former effect lowers the payoff of a FD. In Appendix A, we show that the same holds also for the latter effect. Thus, an increase in the mass of FD erodes the rents from competition, implying that $\phi(\mu)$ is decreasing in μ.[21]

5.3 The supply of technological services and exchange equilibrium

Depending on the industrial organization of exchanges, the supply of technological services is either controlled by a single platform, acting as an "incumbent monopolist,"

[20] It formalizes in a simple manner the way in which Lewis (2014) describes Larry Tabb's estimation of traders' demand for the high speed, fiber optic connection that Spread laid down between New York and Chicago in 2009.

[21] Numerical simulations show that when μ, τ_u, and τ_v are sufficiently large and γ is large above γ^L, $\phi(\mu)$ is log-convex in μ: $(\partial^2 \ln\phi(\mu)/\partial\mu^2) \geq 0$. When this occurs, the price reduction corresponding to an increase in μ becomes increasingly smaller as μ increases. In these conditions, we find that for $N = 2$ exchanges' best response functions can become upward sloping, differently from what happens in standard Cournot models.

or by $N \geq 2$ venues who compete à la Cournot in technological capacities. In the former case, the monopolist profit is given by

$$\pi(\mu) = (\phi(\mu) - c)\mu - f, \tag{14}$$

where c and f, respectively denote the marginal and fixed cost of supplying a capacity μ. We denote by μ^M the optimal capacity of the monopolist exchange: $\mu^M \in \arg\max_{\mu \in (0,1]} \pi(\mu)$. In the latter case, denoting by μ_i and $\mu_{-i} = \sum_{j \neq i}^{N} \mu_j$, respectively the capacity installed by exchange i and its rivals, and by f and c the fixed and marginal cost incurred by an exchange to enter and supply capacity μ_i, an exchange i's profit function is given by

$$\pi(\mu_i, \mu_{-i}) = (\phi(\mu) - c)\mu_i - f. \tag{15}$$

With $N \geq 2$ venues we may assume that dealers are distributed uniformly across the exchanges and that competition among exchanges proceeds in a two-stage manner. First each exchange sets its capacity (and this determines how many dealers become FD in the venue) and then exchanges compete in prices. This two stage game is known to deliver Cournot outcomes under some mild conditions (Kreps and Scheinkman (1983)). We define a symmetric Cournot equilibrium as follows:

Definition 1. *A symmetric Cournot equilibrium in technological service capacities is a set of capacities $\mu_i^C \in (0,1]$, $i = 1, 2, \ldots, N$, such that (i) each μ_i^C maximizes (15), for given capacity choice of other exchanges μ_{-i}^C: $\mu_i^C \in \arg\max_{\mu_i} \pi(\mu_i, \mu_{-i}^C)$, (ii) $\mu_1^C = \mu_2^C = \cdots = \mu_N^C$, and (iii) $\sum_{i=1}^{N} \mu_i^C = \mu^C(N)$.*

We have the following result:

Proposition 4. *There exists at least one symmetric Cournot equilibrium in technological service capacities and no asymmetric ones.*

Proof. See Amir (2018), Proposition 7, and Vives (1999), Section 4.1. □

Numerical simulations show that the equilibrium is unique and stable.[22] As a consequence, standard comparative statics results apply (see, e.g., Section 4.3 in Vives (1999)).

[22] In our setup, a sufficient condition for stability (Section 4.3 in Vives (1999)) is that the elasticity of the slope of the FD inverse demand function is bounded by the number of platforms (plus one): $\mathcal{E}|_{\mu=\mu^C(N)} \equiv -\mu\phi''(\mu)/\phi'(\mu)|_{\mu=\mu^C(N)} < 1 + N$.

In particular, an increase in the number of exchanges leads to an increase in the aggregate technological service capacity, and a decrease in each exchange profit:

$$\frac{\partial \mu^C(N)}{\partial N} \geq 0 \tag{16a}$$

$$\left.\frac{\partial \pi_i(\mu)}{\partial N}\right|_{\mu=\mu^C(N)} \leq 0. \tag{16b}$$

If the number of competing platforms is exogenously determined, condition (16a) implies that spurring competition in the intermediation industry has positive effects in terms of liquidity and gross welfare (Proposition 1 and Numerical Result 1):

Corollary 6. *At a stable Cournot equilibrium, an exogenous increase in the number of competing exchanges has a positive impact on liquidity and gross welfare:* $\partial\Lambda_t/\partial N < 0$, $\partial GW/\partial N > 0$.

Degryse et al. (2015) study 52 Dutch stocks in 2006-2009 (listed on Euronext Amsterdam and trading on Chi-X, Deutsche Börse, Turquoise, BATS, Nasadaq OMX and SIX Swiss Exchange) and find a positive relationship between market fragmentation (in terms of a lower Herfindhal index, higher dispersion of trading volume across exchanges) and the *consolidated liquidity* of the stock. Foucault and Menkveld (2008) also find that consolidated liquidity increased when in 2004 the LSE launched EuroSETS, a new limit order market to allow Dutch brokers to trade stocks listed on Euronext (Amsterdam).

6 Endogenous platform entry and welfare

In this section we endogenize platform entry, and study its implications for welfare and market liquidity.[23] Assuming that platforms' technological capacities are identical ($\mu = N\mu_i$), a social planner who takes into account the costs incurred by the exchanges

[23] For example, according to the UK Competition Commission (2011), a platform entry fixed cost covers initial outlays to acquire the matching engine, the necessary IT architecture to operate the exchange, the contractual arrangements with connectivity partners that provide data centers to host and operate the exchange technology, and the skilled personnel needed to operate the business. The Commission estimated that in 2011 this roughly corresponded to £10-£20 million.

faces the following objective function:

$$\begin{aligned}\mathcal{P}(\mu, N) &\equiv GW(\mu) - c\mu - fN \\ &= \pi(\mu_i)N + \psi(\mu).\end{aligned} \tag{17}$$

Expression (17) is the sum of two components. The first component reflects the profit generated by competing platforms, who siphon out FD surplus, and incur the costs associated with running the exchanges: $\pi(\mu_i)N = ((\phi(\mu) - c)\mu_i - f)N$, implying that FD surplus only contributes indirectly to the planner's function, via platforms' total profit. The second component in (17) reflects the welfare of other market participants: $\psi(\mu) = CE^{SD} + CE_1^L + CE_2^L$, and highlights the welfare effect of technological capacity choices via the liquidity externality.[24] We consider five possible outcomes:

1. Cournot with free entry (CFE). In this case, we look for a symmetric Cournot equilibrium in μ, as in Definition 1, and impose the free entry constraint:

$$(\phi(\mu^C(N)) - c)\frac{\mu^C(N)}{N} \geq f > (\phi(\mu^C(N+1)) - c)\frac{\mu^C(N+1)}{N+1}, \tag{18}$$

   which pins down N. We denote by μ^{CFE}, and N^{CFE} the pair that solves the Cournot case. Note that, given Proposition 4 and (16b), a unique Cournot equilibrium with free entry obtains in our setup if (16b) holds and for a given N the equilibrium is unique.

2. Structural second best (ST). In this case we posit that the planner can determine the number of exchanges that operate in the market. As exchanges compete à la Cournot in technological capacities, we thus look for a solution to the following problem: $\max_{N \geq 1} \mathcal{P}(\mu^C(N), N)$ s. t. $\mu^C(N)$ is a Cournot equilibrium with $\pi_i^C(N) \geq 0$, and denote by μ^{ST}, and N^{ST} the pair that solves the planner's problem.

3. Unrestricted Structural second best (UST). In this case we relax the non-negativity constraint in the STR, assuming that the planner can make side-payments to exchanges if they do not break-even, and look for a solution to the following prob-

[24] Even incumbent exchanges may have to incur an entry cost to supply liquidity in the second round. For example, faced with increasing competition from alternative trading venues, in 2009 LSE decided to absorb Turquoise, a platform set up about a year before by nine of the world's largest banks. (See "LSE buys Turquoise share trading platform," *Financial Times,* December 2009).

lem: $\max_{N\geq 1}\mathcal{P}(\mu^C(N),N)$ s. t. $\mu^C(N)$ is a Cournot equilibrium, and denote by μ^{UST}, and N^{UST} the pair that solves the planner's problem.

4. Conduct second best (CO). In this case, we let the planner set the fee that exchanges charge to FD, assuming free entry of platforms. Because of Corollary 5, $\phi(\mu)$ is invertible in μ, implying that setting the fee is equivalent to choosing the aggregate technological capacity μ. Thus, we look for a solution to the following problem:

$$\max_{\mu\in(0,1]}\mathcal{P}(\mu,N) \text{ s.t. } (\phi(\mu)-c)\frac{\mu}{N}\geq f\geq(\phi(\mu)-c)\frac{\mu}{N+1}, \tag{19}$$

   and denote by μ^{CO} and N^{CO} the pair that solves (19).[25]

5. first best (FB). In this case, we assume that the planner can regulate the market choosing the fee and the number of competing platforms: $\max_{\mu\in(0,1],N\geq 1}\mathcal{P}(\mu,N)$. We denote by μ^{FB} and N^{FB} the pair that solves the planner's problem.

We contrast the above four cases with the "Unregulated Monopoly" outcome (M) defined in Section 5.3, focusing on welfare and market liquidity (we state our results in terms of aggregate technological capacity μ, with the understanding that based on Corollary 1, a higher technological capacity implies a higher liquidity). We make the maintained assumptions that both the monopoly profit and $\mathcal{P}(\mu,1)$ are single-peaked in μ, with maximum monopoly profit being positive, and that the Cournot equilibrium is stable.[26]

6.1 First best vs. market solutions

We start by comparing the first best (FB) outcome with the two polar market solutions of Monopoly (M) and Cournot Free Entry (CFE). We obtain the following result:

Proposition 5. *At the first best the planner sets $N^{FB}=1$. Furthermore:*

1. *If the monopoly solution is interior ($\mu^M\in(0,1)$), then $\mu^{FB}>\mu^M$.*

2. *If at the first best the monopoly profit is non-positive ($\pi^M(\mu^{FB})\leq 0$), then $\mu^{FB}>\mu^{CFE}\geq\mu^M$.*

[25] We assume for simplicity that if the second inequality holds with equality, then only N firms enter.

[26] These conditions are satisfied in all of our simulations (see Table 3).

3. *Welfare comparison:* $\mathcal{P}^{FB}$ *is larger than both* $\mathcal{P}^{CFE}$ *and* $\mathcal{P}^{M}$.

At the first best, the planner minimizes entry costs by letting a single exchange satisfy the industry demand for technological services. If at the first best the monopoly profit is non-positive, then aggregate capacity must be strictly larger than the one at the Cournot Free Entry outcome since otherwise platforms would make negative profits. Furthermore, the capacity supplied at the monopoly outcome can be no larger than the one obtained with Cournot Free Entry since, under Cournot stability, increased platform entry leads to increased technological service capacity. Finally, with higher technological service capacity, and minimized fixed costs, the first best solution is superior to either market outcome.

We can compare μ^{FB} with the capacity that obtains if the fixed cost tends to zero, and thus the number of platforms grows unboundedly at the CFE. In this case, platforms become price takers (PT), and the implied aggregate capacity is implicitly defined by: $\phi(\mu^{PT}) = c$. Since $\mathcal{P} = (\phi(\mu) - c)\mu + \psi(\mu)$, we have that $\mathcal{P}' = \phi(\mu) - c + \mu\phi'(\mu) + \psi'(\mu)$ and therefore:

$$\left.\frac{\partial \mathcal{P}}{\partial \mu}\right|_{\mu=\mu^{PT}} = \mu^{PT}\phi'(\mu^{PT}) + \psi'(\mu^{PT}), \tag{20}$$

which will be positive or negative depending on whether the cost to the industry of marginally increasing capacity $(-\mu^{PT}\phi'(\mu^{PT}))$ is smaller or larger than the marginal benefit to the other market participants $(\psi'(\mu^{PT}))$. At μ^{PT} exchanges do not internalize either effect and only in knife-edge cases we will have that $\mu^{FB} = \mu^{PT}$.[27]

6.2 Fee regulation

We now compare the constrained second best optimum the planner achieves with conduct (fee) regulation (CO) with the two polar market solutions. Under the assumption that the monopoly profit is negative at the first best solution, which implies that at the CO profits are exactly zero (see Lemma A.2 in Appendix A), we obtain the following result:

Proposition 6. *When the planner regulates the technological service fee, if at the first best the monopoly profit is negative,*

[27] Note, however, that if the welfare of other market participants is constant, then the monopoly solution implements the first best.

1. *$N^{CO} = 1$.*

2. *The technological service capacity supplied at the CO is lower than at the FB but higher than at the CFE: $\mu^{FB} > \mu^{CO} > \mu^{CFE}$.*

3. *Welfare ranking: $\mathcal{P}^{FB} > \mathcal{P}^{CO} > \mathcal{P}^{CFE}$.*

Suppose that at the first best the monopoly profit is negative.[28] Then with fee regulation (CO), the aggregate technological capacity should be smaller than at the first best since otherwise the platforms would make negative profits. As for a given (aggregate) μ the profit of an exchange is decreasing in N, for given μ the maximum profit obtains when $N = 1$ implying that $N^{CO} = 1$. Furthermore, given that $\mathcal{P}$ is single peaked in μ, it is optimal for μ^{CO} to be set as large as possible so that profits are zero. Finally, with fee regulation, one platform breaks even, while at a Cournot Free Entry (i) a single platform makes a positive profit (recall that monopoly profits are assumed to be positive), and (ii) if more than one platform is in the market, then platforms lose money when offering a capacity larger or equal to the one obtained with fee regulation. In either case, $\mu^{CFE} < \mu^{CO}$, and we have: $\mu^{CFE} < \mu^{CO} < \mu^{FB}$.[29]

Remark 1. *If at the first best the monopoly profit is positive, two cases can arise. First, we can have that $(\phi(\mu^{FB}) - c)\mu^{FB}/2 - f \leq 0$, in which case both constraints of the Conduct second best problem are satisfied at $(\mu^{CO}, N^{CO}) = (\mu^{FB}, 1)$, and the Conduct second best implements the first best outcome. If, on the other hand, two platforms earn a positive profit at the first best—$(\phi(\mu^{FB}) - c)\mu^{FB}/2 - f > 0$—then at a Conduct second best the planner needs to set a lower fee for technological services compared to the one of the first best and/or allow more than one platform to enter the market. Indeed, if $N^{CO} = 1$, then by construction μ^{CO} cannot be set smaller than μ^{FB} since this would violate the right constraint of the Conduct second best problem.*

[28] We have numerically verified the above sufficient condition for $N^{CO} = 1$, and found that in our simulations it is always satisfied. In the reverse order of actions model, in some cases $\pi^M(\mu^{CO}) > 0$, but the planner still sets $N^{CO} = 1$. See Table 2 for details.

[29] More precisely, notice that μ^{CFE} cannot be higher than μ^{CO}, as at μ^{CO} one firm makes zero profit. Thus, given single-peakedness of the monopoly profit, if there is either one or more than one firm in the CFE with $\mu^{CFE} > \mu^{CO}$, profits will be negative. Similarly, it cannot be $\mu^{CFE} = \mu^{CO}$ because if $N^{CFE} = 1$ then $\mu^{CO} = \mu^{CFE} = \mu^M$, and by assumption the monopoly profit is positive; if, instead, $N^{CFE} > 1$, then more than one firm shares the revenue that one firm has in the CO solution, so that its profit must be negative.

6.3 Entry regulation

Regulating the fee can however be complicated, as our discussion in the introduction suggests. Thus, we now focus on the case in which the planner can only decide on the number of competing exchanges. In the absence of regulation, a Cournot equilibrium with free entry arises (see (18)), and we compare this outcome to the Structural second best, in both the unrestricted and restricted cases. Evaluating the first order condition of the planner at $N = N^{CFE}$ (ignoring the integer constraint) yields:

$$\begin{aligned} \left.\frac{\partial \mathcal{P}(\mu^C(N), N)}{\partial N}\right|_{N=N^{CFE}} = {} & \left.\underbrace{\pi_i(\mu^C(N), N)}_{=0}\right|_{N=N^{CFE}} \\ & + N^{CFE} \left.\underbrace{\frac{\partial \pi_i(\mu^C(N), N)}{\partial N}}_{\text{Profitability depression} < 0}\right|_{N=N^{CFE}} \\ & + \left.\underbrace{\psi'(\mu)\frac{\partial \mu^C(N)}{\partial N}}_{\text{Liquidity creation} > 0}\right|_{N=N^{CFE}} . \end{aligned} \tag{21}$$

According to (21), at a stable Cournot equilibrium, entry has two countervailing welfare effects.[30] The first one is a "profitability depression" effect, and captures the profit decline associated with the demand reduction faced by each platform as a result of entry. This effect is conducive to excessive entry, as each competing exchange does not internalize the negative impact of its entry decision on competitors' profits. The second one is a "liquidity creation" effect and reflects the welfare creation of an increase in N via the liquidity externality. This effect is conducive to insufficient entry since each exchange does not internalize the positive impact of its entry decision on other market participants' payoffs.[31]

The above effects are related but distinct to the ones arising in a Cournot equilibrium with free entry (Mankiw and Whinston (1986) and in the vertical oligopoly of Ghosh and Morita (2007)).

[30] This is because at a stable equilibrium (16a) and (16b) hold, see section 4.3 in Vives (1999).

[31] As clarified by condition (21), the necessary conditions for insufficient entry are: (1) that the total technological capacity installed by entering platforms is increasing in the number of entrants–a property of all stable Cournot equilibria–and (2) that the gross welfare of all agents except for FD is increasing in the total capacity of technological services that platforms supply at equilibrium–a result that holds in our baseline model, and that we formally prove in Corollary 4 (part 1).

In our setup, when we compare N^{CFE} with N^{ST}, entry is always excessive (as in Mankiw and Whinston (1986)); however, when N^{CFE} is stacked against N^{UST}, this conclusion does not necessarily hold. More in detail, N^{CFE} is the the largest N so that platforms break even at a Cournot equilibrium. At the STR solution, platforms break even too, but the planner internalizes the profitability depression effect of entry. Thus, we have that $N^{CFE} \geq N^{ST}$. Conversely, removing the break even constraint, the planner achieves the Unrestricted STR and, depending on which of the effects outlined above prevails, both excessive or insufficient entry can occur:

Proposition 7. *When the planner regulates entry, for stable Cournot equilibria:*

1. $N^{CFE} \geq N^{ST}$, $\mu^{CFE} \geq \mu^{ST}$.
2. *When the profitability depression effect is stronger than the liquidity creation effect,* $N^{CFE} \geq N^{UST}$, $\mu^{CFE} \geq \mu^{UST}$. *Otherwise, the opposite inequalities hold.*
3. *Both* μ^{ST} *and* μ^{UST} *are no smaller than* μ^{M}.
4. *Welfare ranking:* $\mathcal{P}^{UST} \geq \mathcal{P}^{ST} \geq \mathcal{P}^{CFE}$.

The first two items in the proposition reflect our previous discussion. Item 3 shows that while the technological capacity offered with free platform entry (CFE) is higher than at the Structural second best (a natural consequence of excessive entry with respect to the STR benchmark), when the planner relaxes the break-even constraint (UST), the comparison is inconclusive. Indeed, as explained above, to exploit the positive liquidity externality, the planner may favor entry beyond the break-even level–subsidising the loss-making platforms. Thus, while entry regulation implies that liquidity maximization is generally at odds with welfare maximization, the two may be aligned when the planner is ready to make up for platforms' losses. Finally, as at the UST the non-negativity constraint of the exchanges' profit is relaxed, $\mathcal{P}^{UST} \geq \mathcal{P}^{ST}$ must hold.

To verify the possibility of excessive or insufficient entry compared to the UST, we run numerical simulations.[32] We assume $\gamma = 0.5$, $\gamma^L = 0.25$, a 10% annual volatility

[32] We have also extended the parameter space to account for a case with "low" risk aversion ($\gamma = 25$, $\gamma^L = 12$) which is consistent with the literature on price pressure, and recent results on the structural estimation of risk aversion based on insurance market data (see, respectively, Hendershott and Menkveld (2014), and Cohen and Einav (2007)). In this case, we set $\tau_v = \tau_u = 0.1$ (corresponding to a 316% annual volatility for both the endowment shock and the liquidation value), $f \in \{1 \times 10^{-2}, 1.1 \times 10^{-2}, \ldots, 3.1 \times 10^{-2}\}$, and $c = 2$.

for the endowment shock, and consider a “high” and a “low” payoff volatility scenario (respectively, $\tau_v = 3$, which corresponds to a 60% annual volatility for the liquidation value, and $\tau_v = 25$ which corresponds to a 20% annual volatility). Platform costs are set to $f \in \{1 \times 10^{-6}, 2 \times 10^{-6}, \ldots, 31 \times 10^{-6}\}$, and $c = 0.002$.[33] With this set of parameters, we solve for the technological capacity and the number of platforms, and perform robustness analysis (see Tables 2 and 3). In all simulations we obtain $\pi^M(\mu^{FB}) < 0$.

Numerical Result 2. *The results of our numerical simulations are as follows:*

1. *With high payoff volatility, entry is excessive:* $N^{CFE} > N^{UST}$, *and* $\mu^{CFE} > \mu^{UST}$.

2. *With low payoff volatility and for intermediate values of the entry cost, entry is insufficient:* $N^{CFE} < N^{UST}$ *and* $\mu^{CFE} < \mu^{UST}$.

Furthermore, at all solutions N *and* μ *are decreasing in* f.[34]

Figure 2 (panels (a) and (c)) illustrates the output of two simulations in which insufficient entry occurs. Intuitively, the combination of a high entry cost, and low payoff volatility, work to reduce exchanges’ profit margins. A high entry cost, makes it harder for platforms to break-even; a low payoff volatility, reduces traders’ needs to hedge the endowment shock, lowering the rents from liquidity supply. In these conditions, decentralizing entry decisions yields an outcome where liquidity is too low compared to the planner’s solution. According to the figure, insufficient entry occurs.

Let us examine Figure 2(c). When f is small, both N^{CFE} and μ^{CFE} are high. Then, the profitability depression effect

$$N^{CFE} \left. \frac{\partial \pi_i(\mu^C(N), N)}{\partial N} \right|_{N=N^{CFE}},$$

dominates the liquidity creation effect

$$\psi'(\mu) \left. \frac{\partial \mu^C(N)}{\partial N} \right|_{N=N^{CFE}}.$$

[33] Analyzing the US market, Jones (2018) argues that barriers to entry to the intermediation industry are very low, a consideration that is corroborated by the current state of the market, where 13 cash equity exchanges compete with over 30 ATS. This suggests that the entry cost must be low.

[34] Assuming $\gamma = 0.25 < \gamma^L = 0.5$ yields qualitatively similar results in the high volatility case, whereas in the low volatility case insufficient entry disappears. Additionally, with low risk aversion, we find that entry is insufficient for all levels of f, and all levels of payoff volatility.

In these conditions, further entry would have a limited impact on the welfare of other market participants, consistent with the fact that in the simulations $\psi(\mu)$ is concave. The result is thus excessive entry. As f grows, N^{CFE} diminishes and the two effects equilibrate. For larger values of f and smaller N^{CFE}, encouraging entry generates large liquidity creation benefits and there is insufficient entry at the CFE. For very large values of f, entry is so restricted that the two forces equilibrate again since to foster more entry is very expensive (because fN becomes very high).

6.4 Comparing all solutions

The previous sections have shown that either fee regulation (Section 6.2) or entry/merger policy (Section 6.3) can be used as a tool to correct platforms' market power, and improve aggregate welfare. The following result assesses which one of such tools works best:

Proposition 8. *Comparing solutions when $\pi^M(\mu^{FB}) < 0$:*

1. $\mu^{FB} > \mu^{CO} > \mu^{CFE} \geq \mu^{ST} \geq \mu^M$.
2. *The number of exchanges entering the market with Cournot free entry or with entry regulation is no lower than with fee regulation ($N^{CO} = 1$).*
3. *Welfare comparison:*

$$\mathcal{P}^{FB} > \mathcal{P}^{CO} > \mathcal{P}^{ST} \geq \max\left\{\mathcal{P}^{CFE}, \mathcal{P}^{M}\right\} \geq \min\left\{\mathcal{P}^{CFE}, \mathcal{P}^{M}\right\}, \tag{22a}$$

$$\mathcal{P}^{FB} \geq \mathcal{P}^{UST} \geq \mathcal{P}^{ST}, \tag{22b}$$

where if the FB solution is interior, then $\mathcal{P}^{FB} > \mathcal{P}^{UST}$.

The first two items in the proposition respectively follow from Propositions 5, 6, and 7, and from Propositions 6, and 7.

In terms of welfare, due to Propositions 5, and 6, the first best outcome is superior to the one achieved with fee regulation, which is in turn preferred to the monopoly outcome. Since $\mu^{ST} \leq \mu^{CFE} < \mu^{CO} < \mu^{FB}$—so that we are in the increasing (in μ) part of the planner's objective function—and $N^{ST} \geq 1 = N^{CO}$, we have that fee regulation is also superior to entry/merger policy (constrained by the break even condition). In words: with entry policy the planner allows platforms to retain some market power, to make up for the entry cost. However, if the planner can regulate the fee, provided that aggregate capacity at a constrained second best solution falls short of that implied by the first best, a superior outcome can be achieved in terms of liquidity, which also allows to save on setup costs. Finally, entry policy (with the break-even constraint) yields an outcome that is never inferior in terms of welfare to the polar market solutions (CFE and M). Indeed, the latter are always available to the planner.

The results under the assumption of Proposition 8 imply that, if unregulated, the monopoly outcome yields lower liquidity compared to any other alternative. Furthermore, in our simulations, the planner's objective function evaluated at μ^M is always the lowest compared to the other five alternatives. Thus, both from a liquidity, and welfare point of view the unregulated monopoly solution is the worst possible.

7 Extensions

In this section, we summarize the results of four extensions to the baseline model. In Section 7.1, we assume that SD enter the market at the second round; in section 7.2 we suppose that a sector of "committed" SD is in the market at both rounds; in section 7.3 we allow potential dealers to decide whether to enter the intermediation industry, thereby endogenizing the mass of liquidity providers; finally, in section 7.4, we consider the case in which second period traders have access to a noisy signal about the first period endowment shock (see the online appendix for the details of these extensions). We find in general that the results of the baseline model are robust with one important proviso in the case of "committed" SD where entry regulation may be welfare superior. Furthermore, we also find that for baseline parameters at the CFE entry is always excessive both with SD in the second round, and with asymmetric information.

7.1 SD at the second round

In this section, we assume that SD enter the market at the second round (similarly to Huang and Yueshen (2020)). In this case (denoted RO), we find that all of the analytical results of the baseline analysis carry through. More in detail, in relation to the liquidity provision stage of the game, we obtain the following.

Proposition 9. *For $\mu \in (0, 1]$, there exists a unique equilibrium in linear strategies in the stock market where SD enter at the second round. Denoting by $\tilde{\Lambda}_t$, and $\tilde{\Lambda}_{21}$ respectively the weight that p_t assigns to u_t, $t \in \{1, 2\}$, and the one p_2 assigns to u_1, compared to the baseline case, the following holds: $\tilde{\Lambda}_1 > \Lambda_1$, $\Lambda_1 < \tilde{\Lambda}_2 < \Lambda_2$, and $|\tilde{\Lambda}_{21}| > |\Lambda_{21}|$.*

Thus, SD entry at the second round reduces (increases) the competitive pressure faced by FD at the first (second) round, explaining the decrease (increase) in first (second) period liquidity. Comparing dealers' payoffs across the two models (maintaining the convention of using $\tilde{}$ to denote the variables obtained in the RO case), we obtain the following result.

Proposition 10. $\widetilde{CE}^{FD} > \widetilde{CE}^{SD}$*, and SD have a higher payoff when entering in the second round, whereas the result for FD is ambiguous:* $\widetilde{CE}^{SD} > CE^{SD}$*, and* $\widetilde{CE}^{FD} \gtrless CE^{FD}$*.*

In the baseline model, in the first round FD supply liquidity, anticipating the possibility to rebalance their position at the second round. This heightens the competitive pressure they exert on SD compared to the model studied in this section, explaining why $\widetilde{CE}^{SD} > CE^{SD}$. Conversely, the payoff comparison for FD is less clear cut. Indeed, compared to the baseline model, liquidity is lower (higher) at the first (second) round. Now the (inverse) demand for technological services is given by $\tilde{\phi}(\mu) = \widetilde{CE}^{FD} - \widetilde{CE}^{SD}$, which is also decreasing in μ.

Table 3 displays the results for the simulations conducted on the baseline (OO) and the RO models. The table shows that in the RO case, entry is excessive at all τ_v. Additionally, conduct (fee) regulation yields $N^{CO} = 1$ and is welfare superior to structural (entry) regulation.

7.2 The effect of "committed" SD

In the liquidity provision model of Section 4, only FD can supply liquidity to second period traders. This is a convenient assumption which however fails to recognize that in actual markets liquidity provision is ensured by a multitude of market participants. In this section we check how the introduction of a mass of "committed" SD, present in the market at each trading round, can alter the risk-sharing properties of the market, affecting exchanges' technology supply, and the welfare ranking among different forms of regulatory intervention.

Suppose that a mass ϵ of SD is unable to become FD and is in the market at each trading round, so that in total, a mass 2ϵ SD are committed. Thus, at the first round liquidity is supplied by μ FD, and a total mass $1-\mu$ of SD (ϵ of which are committed, and $1-\epsilon-\mu$ are as in the baseline model). At the second round, instead, we have a total mass of μ FD and ϵ SD liquidity suppliers. In the following, we summarize the effect of committed dealers in the model, and refer the interested reader to the Internet Appendix for a detailed analysis. With committed dealers the market clearing equations in Section 3.4 are replaced by: $\mu x_1^{FD} + (1-\mu)x_1^{SD} + x_1^L = 0$, and $(x_2^{FD} - x_1^{FD})\mu + \epsilon x_2^{SD} + x_2^L = 0$, where x_2^{SD} denotes the position of a committed dealer at the second round. Thus, denoting by $\tilde{\Lambda}_t$, $\tilde{\Lambda}_{21}$, $\tilde{a}_t$, and $\tilde{b}$ the coefficients of the linear equilibrium with committed dealers (which replace the corresponding price coefficients in (2a), (2b), and traders' strategies in Proposition 1), in the Internet Appendix, we prove the following result:

Proposition 11. *For $\mu \in [0,1]$, there exists a unique equilibrium in linear strategies in the stock market where a mass ϵ of SD is in the market at both rounds. At equilibrium, the sign of the comparative statics effect of μ, as well as the ranking between liquidity traders' hedging aggressiveness is preserved, while*

1. *$\tilde{\Lambda}_t$, $|\tilde{a}_t|$ are respectively decreasing and increasing in ϵ.*
2. *$|\tilde{\Lambda}_{21}|$ and $\tilde{b}$ are decreasing in ϵ.*

Committed dealers improve the risk bearing capacity of the market, increasing its liquidity at both rounds, and leading traders to hedge a larger fraction of their endowment shock. As second period traders face heightened competition in speculating against the propagated endowment shock, a larger ϵ reduces their response to u_1 ($\tilde{b}$).

To measure the impact on the welfare of market participants and the market for technological services, we appropriately replace the equilibrium coefficients with their tilde-ed counterparts in the expressions for the market participants' payoffs (see (6a), (6b), (8a), and (8b)), and compute the payoff for second period committed dealers which, given Proposition 11, is decreasing in μ. Finally, defining the inverse demand for technological services as the payoff difference between FD and first period dealers: $\phi(\mu) \equiv CE^{FD} - CE_1^{SD}$, we also obtain the following result:

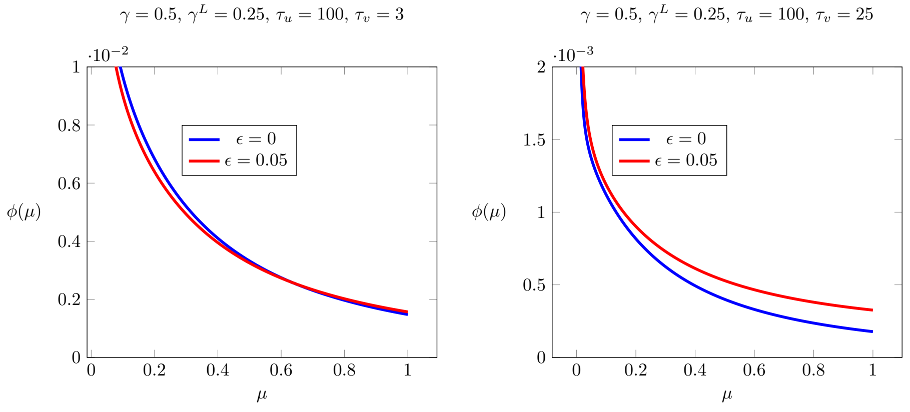
Corollary 7. *With committed SD:*

1. *The comparative statics effect of μ on dealers' and traders' payoffs, and the inverse demand for technological services, are as in Propositions 2, 3, and Corollary 5. Furthermore, the payoff of second period committed dealers is decreasing in μ.*

2. *Standard dealers' payoffs are decreasing in ϵ.*

Committed dealers heighten competition in the provision of liquidity, explaining the second part of Corollary 7. Numerical simulations show that an increase in ϵ increases the payoff of liquidity traders at both rounds as well as that of FD. The former effect is in line with the improved liquidity provision enjoyed by liquidity traders. The intuition for the latter is that besides the competitive effect, committed dealers also improve FD ability to share risk when they retrade at the second period. This also explains why in our simulations, the demand for technological services can be non-monotone in ϵ, as shown in Figure 3.

More in detail, the demand for technological services can increase (decrease) with a higher ϵ in the low (high) payoff volatility scenario. This is consistent with the fact that an increase in τ_v leads first period traders to hedge a larger portion of their endowment, increasing FD risk exposure, and thereby increasing the value of technological services to this class of liquidity providers. Indeed, in the Internet Appendix, we show that $\tilde{a}_1 < 0$, and also that $\partial\tilde{a}_1/\partial\tau_v < 0$.



(a) (b)

Figure 3: Comparative statics effect of an increase in ϵ on the demand for technological services.

Our simulations also confirm that as in Numerical Result 1, gross welfare: $GW(\mu) =$

$\mu(CE^{FD} - CE_1^{SD}) + CE_1^{SD} + \epsilon CE_2^{SD} + CE_1^L + CE_2^L$, is increasing in μ. Furthermore, we also find that $GW(\mu)$ is increasing in ϵ.

We can now use the model to rank market outcomes against the different welfare benchmarks introduced in Section 6:[35]

Numerical Result 3. *With committed dealers, we obtain the same Numerical Result for ϵ small as when $\epsilon = 0$. Otherwise:*

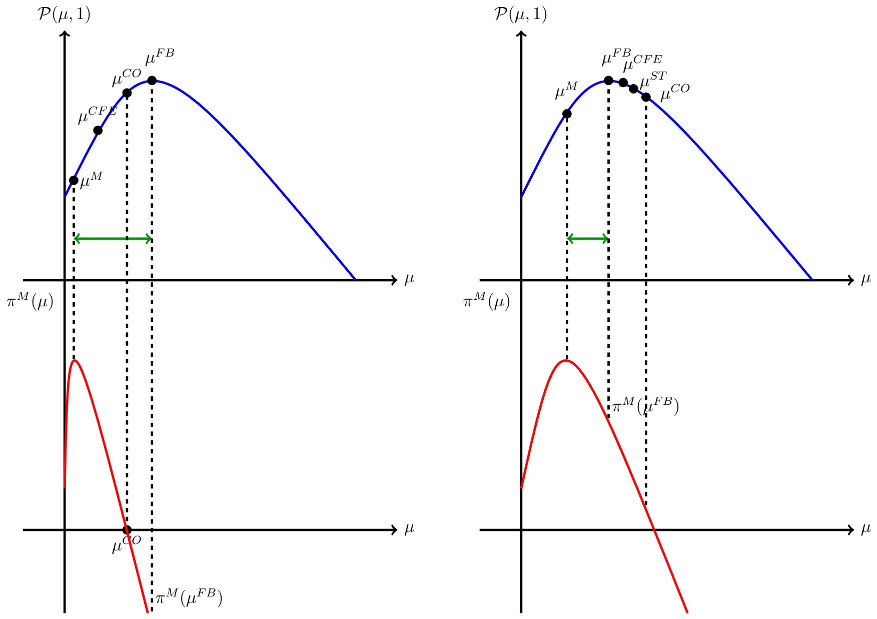
1. *With high payoff volatility ($\tau_v = 3$), $\pi^M(\mu^{FB}) < 0$, and the general ranking result of Proposition 8 applies.*
2. *With low payoff volatility ($\tau_v = 25$), we have that $\pi^M(\mu^{FB}) > 0$, and:*
   (a) *$\mu^{ST} = \mu^{UST}$, and $\mu^{CO} > \mu^{CFE} \geq \max\{\mu^{FB}, \mu^{ST}\}$;*
   (b) *entry regulation can yield a higher welfare than fee regulation when fixed costs are small.*

Furthermore, at all solutions N and μ are decreasing in f.

In Figure 4 we present a case in which, for a small entry cost, entry regulation yields a higher welfare than fee regulation, a result that is at odds with Proposition 8. The reason for this finding is as follows. The presence of committed dealers boosts the demand for technological services, making the monopoly solution "closer" to the first best (with $\epsilon > 0$, $\pi^M(\mu^{FB})$ can be positive). In this case, increasing welfare via fee regulation, requires the planner to set μ very high (much higher than at FB), substantially depressing industry profits for mild liquidity gains. Thus, for a small entry cost, controlling μ by choosing N can be better. Summarizing:

- When μ^M and μ^{FB} are far apart ($\pi^M(\mu^{FB}) < 0$), a very high N is needed to increase capacity via entry, which is very expensive in terms of fixed costs (see Figure 5, panel (a)), and it is optimal to control μ to induce $N = 1$.
- When μ^M and μ^{FB} are closer ($\pi^M(\mu^{FB}) > 0$), increasing welfare with CO substantially depresses industry profits for mild liquidity gains. In this case, it may be better to control μ by choosing N (see Figure 5, panel (b)).

[35] We have run our numerical simulations assuming $\epsilon \in \{0.01, 0.03, 0.05\}$, and standard risk aversion. When $\epsilon \in \{0.01, 0.03\}$ and $\tau_v = 25$, insufficient entry obtains for intermediate values of f, as in the baseline model. When $\epsilon = 0.05$, only excessive entry obtains.



(a)

(b)

Figure 5: The figure illustrates the effectiveness of alternative regulatory tools to approach μ^{FB}. In panel (a) $\epsilon = 0$, $\pi^M(\mu^{FB}) < 0$, and fee regulation dominates. In panel (b) $\epsilon > 0$, and entry regulation can dominate.

7.3 Free entry into market making

In this section, we allow potential dealers to decide whether to enter the intermediation industry prior to making the decision to become FD, thereby endogenizing the mass of dealers who are active in the market. Importantly, this extension adds a further equilibrium restriction to the model. Indeed, in this case we have:

1. *Equilibrium conditions implied by the liquidity provision stage of the game.* Assuming that a mass of dealers $\hat{\mu}$ enters the market, this condition pins down the coefficients of the equilibrium prices at the two trading rounds $\Lambda_t^*, \Lambda_{21}^*$, $t \in \{1,2\}$ as we have done in Section 4.

2. *Equilibrium condition implied by the capacity determination stage of the game.* As done in section 5, based on the equilibrium coefficients obtained at the liquidity provision stage of the game, we determine traders' payoffs and the demand for technological services. With this, we obtain the functional relationship $\mu(\hat{\mu})$ which yields the mass of FD as a function of the total mass of dealers who entered the liquidity provision industry. Note that the specific form of this function

depends on the capacity competition by exchanges.

3. *Equilibrium condition implied by the entry decision of dealers.* Finally, replacing the equilibrium mapping $\mu(\hat{\mu})$ obtained at the previous step in SD's payoffs, we pin down the mass of dealers who enter the intermediation industry solving

$$CE^{SD}(\mu(\hat{\mu});\hat{\mu}) = K, \tag{23}$$

where K represent the reservation utility of staying out of the liquidity provision industry, and where

$$CE^{SD}(\mu(\hat{\mu});\hat{\mu}) = \frac{\gamma}{2}\ln\left(1+\frac{(\Lambda_1^*)^2\tau_v}{\tau_u}\right). \tag{24}$$

This extension is more complicated than our baseline model. Indeed, the mapping $\mu(\hat{\mu})$ (which is obtained at the second step of the algorithm described above) becomes a crucial ingredient of equilibrium determination (i.e., the solution of (23)). Given that the condition yielding $\mu(\hat{\mu})$ is non-linear, we cannot rely on closed form solutions. Our numerical simulations confirm the findings of Numerical Result 2:

1. With low payoff volatility ($\tau_v = 25$), for intermediate values of f, entry is insufficient.

2. With high payoff volatility ($\tau_v = 3$), entry is always excessive.

Additionally, in all our results $N^{CO} = 1$, and fee regulation welfare dominates entry regulation.

7.4 Asymmetric information

In this section, we assume that at the second round a liquidity trader receives an imperfect signal about the first period endowment shock: $s_i = u_1 + \epsilon_i$, where $\epsilon_i \sim N(0, \tau_\epsilon^{-1})$, i.i.d. across traders and orthogonal to all the other random variables in the model. With this assumption, the second period information set of a liquidity trader becomes $\Omega_2^L = \{u_2, s_i\}$. We solve for the equilibrium of the model, and find that depending on the precision of second period traders' signal, three different scenarios arise:

1. *Perfect signal.* In this case, we assume that $\tau_\epsilon \to \infty$, and obtain the result of the baseline model's analysis.

2. *Imprecise signal.* In this case, we assume that $0 < \tau_\epsilon < \infty$, and find that the equilibrium at the liquidity determination stage of the game obtains as a solution to a 7-degree equation in Λ_2 which, based on numerical analysis, admits up to three real roots.

3. *Uninformative signal.* In this case, we assume that $\tau_\epsilon \to 0$, and find that a unique equilibrium obtains at the liquidity determination stage of the game.

Based on our results, it is possible to show that when $0 < \tau_\epsilon < \infty$, the second period equilibrium illiquidity is straddled by the two limit cases of perfect and uninformative second period signal. Other things equal, starting from a parametrization yielding 3 equilibria, as μ or γ increase, a unique equilibrium obtains where Λ_2 tends to the value it obtains when the signal is uninformative. Conversely, as γ^L, τ_u, or τ_v increase, uniqueness obtains with Λ_2 tending to the value it has in the baseline model. This suggests that, to simplify the analysis and gauge the effect of asymmetric information, we can concentrate on the case of an uninformative second period signal.

With this in mind, we replicate the numerical simulations of our baseline model (i.e., using the parameter values of Table 2), assuming that second period traders observe an uninformative signal. The result of our simulations confirm all our findings in Numerical result 2–except for the presence of insufficient entry. Thus, when the signal is uninformative, with the parameter values of Table 2, entry is always excessive. Additionally, in this case too we find that $N^{CO} = 1$ and that fee regulation is always superior to entry regulation.

8 Concluding remarks

We provide a model where both supply and demand for liquidity arise endogenously, with the former depending on exchange competition. Exchanges compete in the provision of technological services (connectivity) which improve the participation of (full) dealers and allow them to absorb more of the net order flow, enhancing the risk bearing capacity of the market. At equilibrium, the mass of full dealers matches the industry's technological service capacity. As a consequence, as exchange competition heightens, the mass of full dealers increases, improving market liquidity and traders' welfare. We use the model to analyze the welfare effects of different entry regimes. A monopolistic exchange exploits its market power, and under-supplies technological services, thereby negatively affecting liquidity and welfare. Allowing competition among trading platforms is beneficial for market quality and (generally) for welfare. However, the market outcome can overprovide or underprovide technological capacity with the corresponding effects on liquidity. If the regulator cannot make transfers to platforms, then entry is never insufficient and the market never underprovides capacity when the benchmark is regulated entry. If, on the other hand, side payments are possible, depending on parameter values entry can also be insufficient.

The optimal second best regulatory approach turns out to depend on the magnitude of the wedge between the technological capacity produced by a monopolistic exchange and the one a first best planner would implement. When such a wedge is large, approaching the first best by spurring entry involves high total fixed costs. In this case, then, the planner limits market power by setting a fee low enough so that only one platform can break even and provide a larger (and cheaper) capacity than the market outcome. When, on the other hand, the wedge is small, the incremental welfare gain achieved through fee regulation is small compared to the industry profit depression this generates. In this case, the planner may find it better to control industry capacity by spurring entry.

Our results suggest that exchanges' technological capacity decisions can be an important *driver* of market liquidity, adding to the usual, demand-based factors highlighted by the market microstructure literature (e.g., arbitrage capital, risk bearing capacity of the market). Thus, any argument about market liquidity should take into account the framework in which exchanges interact. Furthermore, they provide a justification for a number of recent regulatory interventions. First, the need of an "Office of Competition Economics within [the SEC's] Division of Economic Research and Analysis" as advocated by SEC's commissioner Robert J. Jackson Jr., something which, given the latest developments in the debate over the cost of technological services in the US, seems particularly relevant.[36] In this respect, our numerical results can offer guidance as to the type of regulatory intervention. Indeed, we find that when risk-sharing concerns are mild, platform entry is below the planner's desired level. This may seem surprising since with mild risk sharing concerns one would say that there is no need for more entry. But with endogenous (platform) entry, this means that the number of firms that enter the market is very small (low risk sharing concerns means less business for them, because liquidity providers face a lower demand for immediacy, which depresses their demand for technological services). In turn, this means that

[36] *Competition: The Forgotten Fourth Pillar of the SEC's Mission*, Washington D.C., October 2018.

the positive liquidity externality due to an increase in the mass of FD is going to be larger compared to profitability depression effect of new entry. In these conditions, pro-competitive intervention is warranted precisely because the welfare gains that accrue to liquidity traders more than offset the losses incurred by exchanges who end up supplying technological services at a loss. Second, with regard to the SEC's push to upgrade the architecture for the consolidation and dissemination of NMS market data.[37] Through the lenses of our model such a decision corresponds to conditioning the technological capacity offered by exchanges, in order to increase the liquidity creation effect. Finally, our results show the limits of the view aligning liquidity to welfare: with excessive entry, even though the market is more liquid, a social planner chooses to restrict competition, in this way reducing market liquidity.

Our modelling has integrated industrial organization and market microstructure methods taking technological services as homogeneous. An extension of our approach is to consider that exchanges offer differentiated capacities and introduce asymmetries among exchanges. Differentiation could be both in terms of quality (e.g., speed of connection) and horizontal attributes (e.g., lit vs. dark venues).[38]

[37] See *SEC, Proposed Release No. 34-88216; File No. S7-03-20.*

[38] This would also allow to more directly contrast our results with the differentiated approach of Pagnotta and Philippon (2018).

References

Abel, A. B., J. C. Eberly, and S. Panageas (2013). Optimal inattention to the stock market with information costs and transactions costs. *Econometrica 81*(4), 1455–1481.

Aït-Sahalia, Y. and M. Saglam (2013). High frequency traders: Taking advantage of speed. *NBER Working Paper 19531*.

Amir, R. (2018). On the Cournot and Bertrand oligopolies and the theory of supermodular games: A survey. *Handbook of Game Theory and Industrial Organization 1*, 40–65.

Anand, A. and K. Venkataraman (2016). Market conditions, fragility, and the economics of market making. *Journal of Financial Economics 121*(2), 327–349.

Babus, A. and C. Parlatore (2017). Strategic fragmented markets. *Working paper*.

Baldauf, M. and J. Mollner (2017). Trading in fragmented markets. *Working Paper*.

Baron, M., J. Brogaard, B. Hagströmer, and A. Kirilenko (2019). Risk and return in high-frequency trading. *Journal of Financial and Quantitative Analysis 54*(3), 993–1024.

Biais, B., F. Declerck, and S. Moinas (2015). Who supplies liquidity, how and when? *Working paper*.

Biais, B., T. Foucault, and S. Moinas (2015). Equilibrium fast trading. *Journal of Financial Economics 116*, 292–313.

Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan (2015, 08). Trading fast and slow: Colocation and liquidity. *The Review of Financial Studies 28*(12), 3407–3443.

Brogaard, J., T. Hendershott, and R. Riordan (2014). High frequency trading and price discovery. *Review of Financial Studies 27*, 2267–2306.

Brogaard, J. A. (2010). High frequency trading and its impact on market quality. *Working Paper*.

Budish, E., P. Cramton, and J. Shim (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics 139*, 1547–1621.

Budish, E., R. Lee, and J. Shim (2019). Will the market fix the market? a theory of stock exchange competition and innovation. *Working Paper*.

Cantillon, E. and P.-L. Yin (2011). Competition between exchanges: A research agenda. *International Journal of Industrial Organization 29*(3), 329–336.

Cespa, G. and T. Foucault (2014). Sale of price information by exchanges: Does it promote price discovery? *Management Science 60*(1), 148–165.

Chao, Y., C. Yao, and M. Ye (2019). Why discrete pricing fragments the us stock exchanges and disperses their fee structures. *Review of Financial Studies* (32), 1068–1101.

Chen, D. and D. Duffie (2020). Market fragmentation. *Working Paper*.

Chowdhry, B. and V. Nanda (1991). Multimarket trading and market liquidity. *Review of Financial Studies 4*(3), 483–511.

Cohen, A. and L. Einav (2007). Estimating risk preferences from deductible choice. *American Economic Review 97*(3), 745–788.

Degryse, H., F. de Jong, and V. van Kervel (2015). The impact of dark trading and visible fragmentation on market quality. *Review of Finance 19*, 1587–1622.

Dennert, J. (1993). Price competition between market makers. *Review of Economic Studies* (60), 735–751.

Ding, S., J. Hanna, and T. Hendershott (2014). How slow is the NBBO? a comparison with direct exchange feeds. *Financial Review 49*, 313–332.

Duffie, D. (2010). Presidential address: Asset price dynamics with slow-moving capital. *Journal of Finance 65*(4), 1237–1267.

Foucault, T. and A. J. Menkveld (2008). Competition for order flow and smart order routing systems. *The Journal of Finance 63*(1), 119–158.

Foucault, T., M. Pagano, and A. Röell (2013). *Market Liquidity.* Oxford University Press.

Gabszewicz, J. and J. Thisse (1979). Price competition, quality and income disparities. *Journal of Economic Theory 20*(3), 340–359.

Ghosh, A. and H. Morita (2007). Free entry and social efficiency under vertical oligopoly. *RAND Journal of Economics 38*(3), 541–554.

Grossman, S. J. and M. H. Miller (1988). Liquidity and market structure. *Journal of Finance 43*(3), 617–37.

Harris, J. H. and M. Saad (2014). The sound of silence. *The Financial Review 49*, 203–230.

Hasbrouck, J. and G. Saar (2013). Low-latency trading. *Journal of Financial Markets 16*(4), 646–679.

Hendershott, T. and C. M. Jones (2005). Trade-through prohibitions and market quality. *Journal of Financial Markets 8*, 1–23.

Hendershott, T. and A. J. Menkveld (2014). Price pressures. *Journal of Financial Economics 114*(3), 405 – 423.

Heston, S. L., R. A. Korajczyk, and R. Sadka (2010). Intraday patterns in the cross-section of stock returns. *The Journal of Finance 65*(4), 1369–1407.

Holden, C. W. and S. Jacobsen (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance 69*(4), 1747–1785.

Huang, S. and B. Yueshen (2020). Speed acquisition. *Management Science (Forthcoming).*

Jones, C. M. (2018). Understanding the market for U.S. equity market data. *Working paper*.

Kreps, D. and J. Scheinkman (1983). Quantity precommitment and bertrand competition yield cournot outcomes. *Bell Journal of Economics 14*(2), 326–337.

Lewis, M. (2014). *Flash Boys.* London: Allen Lane.

Li, J. (2015). Slow price adjustment to public news in after-hours trading. *Journal of Trading 11*(3), 16–31.

Malamud, S. and M. Rostek (2017). Decentralized exchange. *American Economic Review 107*, 3320–3362.

Mankiw, N. G. and M. D. Whinston (1986). Free entry and social inefficiency. *RAND Journal of Economics 17*(1), 48–58.

Manzano, C. and X. Vives (2018). Market power and welfare in asymmetric divisible good auctions. *Theoretical Economics (Forthcoming).*

Menkveld, A. J. (2016). The economics of high-frequency trading: Taking stock. *Annual Review of Financial Economics 8*, 1–24.

Menkveld, A. J. and M. A. Zoican (2017). Need for speed? Exchange latency and liquidity. *Review of Financial Studies 30*(4), 1188–1228.

Pagano, M. (1989). Trading volume and asset liquidity. *Quarterly Journal of Economics 104*, 255–274.

Pagnotta, E. (2013). Speed, fragmentation, and asset prices. *Working paper*.

Pagnotta, E. and T. Philippon (2018). Competing on speed. *Econometrica 86*(3), 1067–115.

Santos, T. and J. A. Scheinkman (2001). Competition among exchanges. *The Quarterly Journal of Economics 116*(3), 1027–1061.

SEC (2010). Concept release on equity market structure. *Federal Register*.

Shaked, A. and J. Sutton (1982). Relaxing price competition through product differentiation. *Review of Economic Studies 49*(1), 3–13.

UK Competition Commission (2011, November). *BATS Global Markets, Inc/Chi-X Europe Limited, merger inquiry.* UK Competition Commission.

Vayanos, D. and J. Wang (2012). Liquidity and asset returns under asymmetric information and imperfect competition. *Review of Financial Studies 25*(5), 1339–1365.

Vives, X. (1999). *Oligopoly Pricing: old ideas and new tools.* MIT Press.

Vives, X. (2008). *Information and learning in markets: the impact of market microstructure*. Princeton University Press.

A Appendix

The following is a standard result (see, e.g., Vives (2008), Technical Appendix, pp. 382–383) that allows us to compute the unconditional expected utility of market participants.

Lemma A.1. *Let the n-dimensional random vector $z \sim N(0, \Sigma)$, and $w = c + b'z + z'Az$, where $c \in \mathbb{R}$, $b \in \mathbb{R}^n$, and A is a $n \times n$ matrix. If the matrix $\Sigma^{-1} + 2\rho A$ is positive definite, and $\rho > 0$, then*

$$E[-\exp\{-\rho w\}] = -|\Sigma|^{-1/2}|\Sigma^{-1} + 2\rho A|^{-1/2} \exp\{-\rho(c - \rho b'(\Sigma^{-1} + 2\rho A)^{-1}b)\}.$$

Proof of Proposition 1

We start by assuming that at a linear equilibrium prices are given by

$$p_2 = -\Lambda_2 u_2 + \Lambda_{21} u_1 \tag{A.1a}$$

$$p_1 = -\Lambda_1 u_1, \tag{A.1b}$$

with Λ_1, Λ_{21}, and Λ_2 to be determined in equilibrium. In the second period a new mass of liquidity traders with risk-tolerance coefficient $\gamma^L > 0$ enter the market. Because of CARA and normality, the objective function of a second period liquidity trader is given by

$$E[-\exp\{-\pi_2^L/\gamma^L\}|\Omega_2^L] = -\exp\left\{-\frac{1}{\gamma^L}\left(E[\pi_2^L|\Omega_2^L] - \frac{1}{2\gamma^L}\text{Var}[\pi_2^L|\Omega_2^L]\right)\right\}, \tag{A.2}$$

where $\Omega_2^L = \{u_1, u_2\}$, and $\pi_2^L \equiv (v - p_2)x_2^L + u_2 v$. Maximizing (A.2) with respect to x_2^L, yields:

$$X_2^L(u_1, u_2) = \gamma^L \frac{E[v - p_2|\Omega_2^L]}{\text{Var}[v - p_2|\Omega_2^L]} - \frac{\text{Cov}[v - p_2, v|\Omega_2^L]}{\text{Var}[v - p_2|\Omega_2^L]} u_2. \tag{A.3}$$

Using (A.1a):

$$E[v - p_2|\Omega_2^L] = \Lambda_2 u_2 - \Lambda_{21} u_1 \tag{A.4a}$$

$$\text{Var}[v - p_2|\Omega_2^L] = \text{Cov}[v - p_2, v|\Omega_2^L] = \frac{1}{\tau_v}. \tag{A.4b}$$

Substituting (A.4a) and (A.4b) in (A.3) yields

$$X_2^L(u_1, u_2) = a_2 u_2 + b u_1, \tag{A.5}$$

where

$$a_2 = \gamma^L \tau_v \Lambda_2 - 1 \tag{A.6a}$$
$$b = -\gamma^L \tau_v \Lambda_{21}. \tag{A.6b}$$

Consider the sequence of market clearing equations

$$\mu x_1^{FD} + (1-\mu) x_1^{SD} + x_1^L = 0 \tag{A.7a}$$
$$\mu(x_2^{FD} - x_1^{FD}) + x_2^L = 0. \tag{A.7b}$$

Condition (A.7b) highlights the fact that since first period liquidity traders and SD only participate at the first trading round, their positions do not change across dates. Rearrange (A.7a) as follows:

$$(1-\mu) x_1^{SD} + x_1^L = -\mu x_1^{FD}.$$

Substitute the latter in (A.7b):

$$\mu x_2^{FD} + x_2^L + (1-\mu) x_1^{SD} + x_1^L = 0. \tag{A.8}$$

To pin down p_2, we need the second period strategy of FD and the first period strategies of SD and liquidity traders. Starting from the former, because of CARA and normality, the expected utility of a FD is given by:

$$E\Big[-\exp\Big\{ -\frac{1}{\gamma}\Big((p_2 - p_1) x_1^{FD} + (v - p_2) x_2^{FD} \Big) \Big\} | p_1, p_2 \Big] = \tag{A.9}$$
$$= \exp\Big\{ -\frac{1}{\gamma}(p_2 - p_1) x_1^{FD} \Big\} \Big(-\exp\Big\{ -\frac{1}{\gamma}\Big(E[v - p_2 | p_1, p_2] x_2^{FD} - \frac{(x_2^{FD})^2}{2\gamma} \mathrm{Var}[v - p_2 | p_1, p_2] \Big) \Big\} \Big),$$

For given x_1^{FD} the above is a concave function of x_2^{FD}. Maximizing with respect to x_2^{FD} yields:

$$X_2^{FD}(p_1, p_2) = -\gamma \tau_v p_2. \tag{A.10}$$

Similarly, due to CARA and normality, in the first period a traditional market maker maximizes

$$E\Big[-\exp\Big\{-\frac{1}{\gamma}(v-p_1)x_1^{SD}\Big\}|p_1\Big] = \tag{A.11}$$
$$-\exp\Big\{-\frac{1}{\gamma}\Big(E[v-p_1|p_1]x_1^{SD}-\frac{(x_1^{SD})^2}{2\gamma}\mathrm{Var}[v-p_1|p_1]\Big)\Big\}.$$

Hence, his strategy is given by

$$X_1^{SD}(p_1) = -\gamma\tau_v p_1. \tag{A.12}$$

Finally, consider a first period liquidity trader. CARA and normality imply

$$E[-\exp\{-\pi_1^L/\gamma^L\}] = -\exp\Big\{-\frac{1}{\gamma}\Big(E[\pi_1^L|u_1]-\frac{1}{2\gamma^L}\mathrm{Var}[\pi_1^L|u_1]\Big)\Big\}, \tag{A.13}$$

where $\pi_1^L \equiv (v-p_1)x_1^L + u_1 v$. Maximizing (A.13) with respect to x_1^L, and solving for the optimal strategy, yields

$$X_1^L(u_1) = \gamma^L\frac{E[v-p_1|u_1]}{\mathrm{Var}[v-p_1|u_1]} - \frac{\mathrm{Cov}[v-p_1,v|u_1]}{\mathrm{Var}[v-p_1|u_1]}u_1. \tag{A.14}$$

Using (A.1b):

$$E[v-p_1|u_1] = \Lambda_1 u_1 \tag{A.15a}$$
$$\mathrm{Cov}[v-p_1,v|u_1] = \frac{1}{\tau_v}. \tag{A.15b}$$

Substituting the above in (A.14) yields

$$X_1^L(u_1) = a_1 u_1, \tag{A.16}$$

where

$$a_1 = \gamma^L\tau_v\Lambda_1 - 1. \tag{A.17}$$

Substituting (A.5), (A.10), (A.12), and (A.16) in (A.8) and solving for p_2 yields

$$p_2 = -\underbrace{\frac{1-\gamma^L\tau_v\Lambda_2}{\mu\gamma\tau_v}}_{\Lambda_2}u_2 + \underbrace{\frac{((1-\mu)\gamma+\gamma^L)\tau_v\Lambda_1 - 1 - \gamma^L\tau_v\Lambda_{21}}{\mu\gamma\tau_v}}_{\Lambda_{21}}u_1. \tag{A.18}$$

Identifying the price coefficients:

$$\Lambda_2 = \frac{1}{(\mu\gamma + \gamma^L)\tau_v} \tag{A.19a}$$

$$\Lambda_{21} = \Lambda_2 \left(((1-\mu)\gamma + \gamma^L)\tau_v \Lambda_1 - 1\right). \tag{A.19b}$$

Substituting the above expressions in (A.18), and using (A.12) yields:

$$p_2 = -\Lambda_2 u_2 + \Lambda_2 \left((1-\mu)x_1^{SD} + x_1^L\right).$$

Consider now the first period. We start by characterizing the strategy of a FD. Substituting (A.10) in (A.9), rearranging, and applying Lemma A.1 yields the following expression for the first period objective function of a FD:

$$E[U((p_2 - p_1)x_1^{FD} + (v - p_2)x_2^{FD})|u_1] = -\left(1 + \frac{\text{Var}[p_2|u_1]}{\text{Var}[v]}\right)^{-1/2} \times \tag{A.20}$$
$$\exp\left\{-\frac{1}{\gamma}\left(\frac{\gamma\tau_v}{2}\nu^2 + (\nu - p_1)x_1^{FD} - \frac{(x_1^{FD} + \gamma\tau_v\nu)^2}{2\gamma}\left(\frac{1}{\text{Var}[p_2|u_1]} + \frac{1}{\text{Var}[v]}\right)^{-1}\right)\right\},$$

where, due to (A.1a) and (A.1b)

$$\nu \equiv E[p_2|u_1] = \Lambda_{21}u_1 \tag{A.21a}$$

$$\text{Var}[p_2|u_1] = \frac{\Lambda_2^2}{\tau_u}. \tag{A.21b}$$

Maximizing (A.20) with respect to x_1^{FD} and solving for the first period strategy yields

$$\begin{aligned} X_1^{FD}(p_1) &= \gamma\frac{E[p_2|u_1]}{\text{Var}[p_2|u_1]} - \gamma\left(\frac{1}{\text{Var}[p_2|u_1]} + \frac{1}{\text{Var}[v]}\right)p_1 \\ &= \gamma\frac{\Lambda_{21}\tau_u}{\Lambda_2^2}u_1 - \gamma\frac{\tau_u + \Lambda_2^2\tau_v}{\Lambda_2^2}p_1. \end{aligned} \tag{A.22}$$

Substituting (A.12), (A.16), and (A.22) in (A.7a) and solving for the price yields $p_1 = -\Lambda_1 u_1$, where

$$\Lambda_1 = \left(\left(1 + \frac{\mu\gamma^L\tau_u}{\Lambda_2 + \mu\gamma\tau_u}\right)\gamma + \gamma^L\right)^{-1}\frac{1}{\tau_v}. \tag{A.23}$$

The remaining equilibrium coefficients are as follows:

$$
\begin{aligned}
a_1 &= \gamma^L \Lambda_1 \tau_v - 1 && \text{(A.24)}\\
a_2 &= -\frac{\mu\gamma}{\mu\gamma + \gamma^L} && \text{(A.25)}\\
b &= -\gamma^L \tau_v \Lambda_{21} && \text{(A.26)}\\
\Lambda_{21} &= -\frac{\mu\gamma(\Lambda_2^2 \tau_v + \tau_u)}{\mu\gamma\tau_u + \Lambda_2}\Lambda_1 && \text{(A.27)}\\
\operatorname{Var}[p_2|u_1] &= \frac{\Lambda_2^2}{\tau_u}, && \text{(A.28)}
\end{aligned}
$$

where Λ_2 is given by (A.19a). An explicit expression for Λ_1 can be obtained substituting (A.19a) into (A.23):

$$
\Lambda_1 = \frac{1 + (\mu\gamma + \gamma^L)\mu\gamma\tau_u\tau_v}{(\gamma + \gamma^L + (\gamma + 2\gamma^L)(\mu\gamma + \gamma^L)\mu\gamma\tau_u\tau_v)\tau_v}. \tag{A.29}
$$

Finally, substituting (A.19a) and (A.29) in (A.27) yields

$$
\Lambda_{21} + \Lambda_1 = \frac{\gamma^L}{\tau_v(\gamma\mu + \gamma^L)(\gamma\mu\tau_u\tau_v(\gamma + 2\gamma^L)(\gamma\mu + \gamma^L) + \gamma + \gamma^L)} > 0. \tag{A.30}
$$

□

Proof of Corollary 1

The comparative static effect of μ and γ on Λ_2 follows immediately from (A.19a). For Λ_1, differentiating (A.29) with respect to μ and γ yields:

$$
\begin{aligned}
\frac{\partial \Lambda_1}{\partial \mu} &= -\frac{(2\mu\gamma + \gamma^L)\gamma\gamma^L\tau_u}{(\gamma + \gamma^L + (\gamma + 2\gamma^L)(\mu\gamma + \gamma^L)\mu\gamma\tau_u\tau_v)^2} < 0\\
\frac{\partial \Lambda_1}{\partial \gamma} &= -\frac{1 + (\gamma^2\mu\tau_u\tau_v(\gamma\mu + \gamma^L)^2 + 2\gamma^2\mu + 2\gamma\gamma^L\mu + 2\gamma\gamma^L + (\gamma^L)^2)\mu\tau_u\tau_v}{\tau_v(\gamma\mu\tau_u\tau_v(\gamma + 2\gamma^L)(\gamma\mu + \gamma^L) + \gamma + \gamma^L)^2} < 0,
\end{aligned}
$$

which proves our result. □

Proof of Corollary 2

The first part of the corollary follows from the fact that $\Lambda_1 < \Lambda_2$. Also, since Λ_t is decreasing in μ, because of (3d), $|a_t|$ is increasing in μ. Finally, substituting (A.27) in (A.26) and rearranging yields

$$b = \frac{\mu\gamma\gamma^L(1+(\mu\gamma+\gamma^L)^2\tau_u\tau_v)}{(\mu\gamma+\gamma^L)(\gamma+\gamma^L+(\gamma+2\gamma^L)\mu\gamma\tau_u\tau_v)},$$

which is increasing in μ. □

Proof of Proposition 2

We start by obtaining an expression for the unconditional expected utility of SD and FD. Because of CARA and normality, a dealer's conditional expected utility evaluated at the optimal strategy is given by

$$\begin{aligned} E[U((v-p_1)x_1^{SD})|p_1] &= -\exp\left\{-\frac{(E[v|p_1]-p_1)^2}{2\text{Var}[v]}\right\} \\ &= -\exp\left\{-\frac{\tau_v\Lambda_1^2}{2}u_1^2\right\}. \end{aligned} \tag{A.31}$$

Thus, traditional dealers derive utility from the expected, long term capital gain obtained supplying liquidity to first period hedgers.

$$\begin{aligned} EU^{SD} \equiv E\left[U\left((v-p_1)x_1^{SD}\right)\right] &= -\left(1+\frac{\text{Var}[p_1]}{\text{Var}[v]}\right)^{-1/2} \\ &= -\left(\frac{\tau_{u_1}}{\tau_{u_1}+\tau_v\Lambda_1^2}\right)^{1/2}, \end{aligned} \tag{A.32}$$

and

$$CE^{SD} = \frac{\gamma}{2}\ln\left(1+\frac{\text{Var}[p_1]}{\text{Var}[v]}\right). \tag{A.33}$$

Differentiating CE^{SD} with respect to μ yields:

$$\begin{aligned} \frac{\partial CE^{SD}}{\partial\mu} &= \frac{\gamma\tau_v}{2}\left(1+\frac{\text{Var}[p_1]}{\text{Var}[v]}\right)^{-1}\frac{\partial\text{Var}[p_1]}{\partial\mu} \\ &= \frac{\gamma\tau_v}{2\tau_{u_1}}\left(1+\frac{\text{Var}[p_1]}{\text{Var}[v]}\right)^{-1}2\Lambda_1\frac{\partial\Lambda_1}{\partial\mu} < 0, \end{aligned} \tag{A.34}$$

since Λ_1 is decreasing in μ.

Turning to FD. Replacing (A.22) in (A.20) and rearranging yields

$$E[U((p_2 - p_1)x_1^{FD} + (v - p_2)x_2^{FD})|u_1] = -\left(1 + \frac{\text{Var}[p_2|u_1]}{\text{Var}[v]}\right)^{-1/2} \times \exp\left\{-\frac{g(u_1)}{\gamma}\right\}, \tag{A.35}$$

where

$$g(u_1) = \frac{\gamma}{2}\left(\frac{(E[p_2|p_1] - p_1)^2}{\text{Var}[p_2|p_1]} + \frac{(E[v|p_1] - p_1)^2}{\text{Var}[v]}\right).$$

The argument of the exponential in (A.35) is a quadratic form of the first period endowment shock. We can therefore apply Lemma A.1 and obtain

$$\begin{aligned} EU^{FD} &\equiv E[U((p_2 - p_1)x_1^{FD} + (v - p_2)x_2^{FD})] = \\ &= -\left(1 + \frac{\text{Var}[p_2|p_1]}{\text{Var}[v]}\right)^{-1/2}\left(1 + \frac{\text{Var}[p_1]}{\text{Var}[v]} + \frac{\text{Var}[E[p_2|p_1] - p_1]}{\text{Var}[p_2|p_1]}\right)^{-1/2}, \end{aligned} \tag{A.36}$$

where, because of (A.21a),

$$\text{Var}\left[E[p_2 - p_1|p_1]\right] = \left(\Lambda_{21} + \Lambda_1\right)^2 \tau_u^{-1}, \tag{A.37}$$

so that:

$$\frac{\text{Var}[E[p_2 - p_1|u_1]]}{\text{Var}[p_2|u_1]} = \left(\frac{\Lambda_{21} + \Lambda_1}{\Lambda_2}\right)^2.$$

Therefore, we obtain

$$CE^{FD} = \frac{\gamma}{2}\left\{\ln\left(1 + \frac{(\Lambda_2)^2\tau_v}{\tau_u}\right) + \ln\left(1 + \frac{(\Lambda_1)^2\tau_v}{\tau_u} + \left(\frac{\Lambda_{21} + \Lambda_1}{\Lambda_2}\right)^2\right)\right\}. \tag{A.38}$$

Computing,

$$\frac{\Lambda_{21} + \Lambda_1}{\Lambda_2} = \frac{\gamma^L}{\gamma + \gamma^L + (\gamma + 2\gamma^L)(\mu\gamma + \gamma^L)\mu\gamma\tau_u\tau_v}. \tag{A.39}$$

Thus, the arguments of the logarithms in (A.38) are decreasing in μ, which proves that CE^{FD} is decreasing in μ.

Finally, note that taking the limits for $\mu \to 0$ and $\mu \to 1$ in (A.33) and (A.38) yields

$$\lim_{\mu\to 0} CE^{SD} = \frac{\gamma}{2}\ln\left(1+\frac{1}{(\gamma+\gamma^L)^2\tau_u\tau_v}\right)$$

$$\lim_{\mu\to 1} CE^{FD} = \frac{\gamma}{2}\left\{\ln\left(1+\frac{1}{(\gamma+\gamma^L)^2\tau_u\tau_v}\right)+\ln\left(1+\frac{(\Lambda_1)^2\tau_v}{\tau_u}+\left(\frac{\Lambda_{21}+\Lambda_1}{\Lambda_2}\right)^2\right)\right\},$$

which proves the last part of the corollary. □

Proof of Proposition 3

Consider now first period liquidity traders. Evaluating the objective function at optimum and rearranging yields

$$-\exp\left\{-\frac{1}{\gamma^L}\left(E[\pi_1^L|u_1]-\frac{1}{2\gamma^L}\mathrm{Var}[\pi_1^L|u_1]\right)\right\} = -\exp\left\{-\frac{u_1^2}{\gamma^L}\left(\frac{a_1^2-1}{2\gamma^L\tau_v}\right)\right\},$$

where $u_1 \sim N(0,\tau_{u_1}^{-1})$. The argument of the exponential is a quadratic form of a normal random variable. Therefore, applying again Lemma A.1 yields

$$E\left[-\exp\left\{\frac{\pi_1^L}{\gamma^L}\right\}\right] = -\left(\frac{(\gamma^L)^2\tau_u\tau_v}{(\gamma^L)^2\tau_u\tau_v-1+a_1^2}\right)^{1/2}, \tag{A.40}$$

so that

$$CE_1^L = \frac{\gamma^L}{2}\ln\left(1+\frac{a_1^2-1}{(\gamma^L)^2\tau_u\tau_v}\right). \tag{A.41}$$

Note that a higher a_1^2 increases traders' expected utility, and thus increases their payoff.

Next, for second period liquidity traders, substituting the optimal strategy (A.3) in the objective function (A.2) yields

$$\begin{aligned} E\left[-\exp\left\{-\frac{\pi_2^L}{\gamma^L}\right\}|\Omega_2^L\right] &= -\exp\left\{-\frac{1}{\gamma^L}\left(\frac{(x_2^L)^2-u_2^2}{2\gamma^L\tau_v}\right)\right\} \\ &= -\exp\left\{-\frac{1}{\gamma^L}\begin{pmatrix} x_2^L & u_2\end{pmatrix}\left(\frac{1}{2\gamma_2^L\tau_v}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\right)\begin{pmatrix} x_2^L\\ u_2\end{pmatrix}\right\}. \end{aligned} \tag{A.42}$$

The argument of the exponential is a quadratic form of the normally distributed random vector

$$\begin{pmatrix} x_2^L & u_2\end{pmatrix} \sim N\left(\begin{pmatrix} 0 & 0\end{pmatrix},\Sigma\right),$$

where

$$\Sigma \equiv \begin{pmatrix} \mathrm{Var}[x_2^L] & a_2\mathrm{Var}[u_2] \\ a_2\mathrm{Var}[u_2] & \mathrm{Var}[u_2] \end{pmatrix}. \tag{A.43}$$

Therefore, we can again apply Lemma A.1 to (A.42), obtaining

$$E\left[E\left[-\exp\left\{-\frac{\pi_2^L}{\gamma^L}\right\}|\Omega_2^L\right]\right] = -|I+(2/\gamma^L)\Sigma A|^{-1/2}, \tag{A.44}$$

where

$$A \equiv \frac{1}{2\gamma^L\tau_v}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{A.45}$$

$$\mathrm{Var}[x_2^L] = \frac{a_2^2+b^2}{\tau_u}. \tag{A.46}$$

Substituting (A.43), (A.45), and (A.46) in (A.44) and computing the certainty equivalent, yields:

$$CE_2^L = \frac{\gamma^L}{2}\ln\left(1+\frac{a_2^2-1}{(\gamma^L)^2\tau_u\tau_v}+\frac{b^2((\gamma^L)^2\tau_u\tau_v-1)}{(\gamma^L)^4\tau_u^2\tau_v^2}\right). \tag{A.47}$$

For $\mu=0$, $b=0$ and, in view of Corollary 2, $CE_1^L > CE_2^L$. The same condition holds when evaluating (A.41) and (A.47) at $\mu=1$. As CE_t^L is increasing in μ, we have that for all $\mu \in (0,1]$, $CE_1^L(\mu) > CE_2^L(\mu)$. □

Proof of Corollary 3

We need to prove that:

$$\frac{\partial CE_1^L(\mu)}{\partial\mu} > -\frac{\partial CE^{SD}(\mu)}{\partial\mu}.$$

Computing:

$$\frac{\partial CE_1^L(\mu)}{\partial\mu} = \frac{\gamma^L a_1 a_1'}{(\gamma^L)^2\tau_u\tau_v-1+a_1^2}, \tag{A.48}$$

where a_1' the partial derivative of a_1 with respect to μ, and

$$\frac{\partial CE^{SD}(\mu)}{\partial\mu} = \frac{\gamma(1+a_1)a_1'}{(\gamma^L)^2\tau_u\tau_v+(1+a_1)^2}. \tag{A.49}$$

First, note that the denominator in (A.49) is higher than the one in (A.48), and they are both positive. Next, comparing the numerators in the above expressions yields:

$$\gamma^L a_1 a_1' > -\gamma(1+a_1)a_1' \iff \underbrace{(\gamma^L a_1 + \gamma(1+a_1))}_{<0} \underbrace{a_1'}_{<0} > 0,$$

as can be checked by substituting (A.24) in the above. Thus, the LHS of the inequality to be proved has a higher (positive) numerator and a lower (positive) denominator compared to the (positive) numerator and denominator of the RHS, and the inequality follows. □

Proof of Corollary 4

The first part of the result follows immediately from (12), and Corollary 3. Next, because of Propositions 2 and 3, $GW(1) > \lim_{\mu\to 0} GW(\mu)$, which rules out the possibility that gross welfare is maximized at $\mu \approx 0$. □

Proof of Corollary 5

Note that because of (A.39), we can write

$$\frac{\Lambda_{21}+\Lambda_1}{\Lambda_2} = \frac{\Lambda_1 \gamma^L \tau_v}{1+\mu\gamma(\mu\gamma+\gamma^L)\tau_u\tau_v}.$$

Therefore, substituting the expressions for dealers' payoffs in (13), we have:

$$\begin{aligned} \phi(\mu) &= CE^{FD} - CE^{D} \qquad\qquad (A.50)\\ &= \frac{\gamma}{2}\left\{ \ln\left(1+\frac{\Lambda_2^2\tau_v}{\tau_u}\right) + \ln\left(1+\frac{\Lambda_1^2\tau_v}{\tau_u}K\right) - \ln\left(1+\frac{\Lambda_1^2\tau_v}{\tau_u}\right)\right\} > 0. \end{aligned}$$

where $K = 1+(\gamma^L/(1+\mu\gamma(\mu\gamma+\gamma^L)\tau_u\tau_v))^2\tau_u\tau_v > 1$, and decreasing in μ. The first term inside curly braces in the above expression is decreasing in μ since Λ_2 is decreasing in μ. The difference between the second and third terms can be written as follows:

$$\ln\left(1+\frac{\Lambda_1^2\tau_v}{\tau_u}K\right) - \ln\left(1+\frac{\Lambda_1^2\tau_v}{\tau_u}\right) = \ln\left(\frac{\tau_u+\Lambda_1^2\tau_v K}{\tau_u+\Lambda_1^2\tau_v}\right).$$

Differentiating the above logarithm and rearranging yields:

$$\frac{\tau_v\Lambda_1}{(\tau_u+\Lambda_1^2\tau_v K)(\tau_u+\Lambda_1^2\tau_v)}\left(2(K-1)\tau_u\frac{\partial\Lambda_1}{\partial\mu} + (\tau_u+\Lambda_1^2\tau_u)\Lambda_1\frac{\partial K}{\partial\mu}\right) < 0,$$

since $K > 1$, and both Λ_1 and K are decreasing in μ. □

We now state and prove a lemma which will be useful for some of the proofs that follow:

Lemma A.2. $\pi^M(\mu^{FB}) \leq 0 \Longrightarrow \pi^M(\mu^{CO}) = 0$, *and the converse is also true generically.*

Proof of Lemma A.2

We first we prove the direction $\Longrightarrow$. Since $\pi^M(\mu^{FB}) \leq 0$ then, given that the monopoly profit is single-peaked, the CO constraints can only be satisfied for $\mu \leq \mu^{FB}$. Note that for a given (aggregate) μ, the profit of an exchange (given that it is non-negative) is non-increasing in N; so for a given μ, $N = 1$ maximizes profit. Then, given that $\mathcal{P}(\mu)$ is single-peaked at μ^{FB}, it is optimal for μ^{CO} to be set as large as possible with $N^{CO} = 1$, so that $\pi^M(\mu^{CO}) = 0$.

Next we prove the opposite direction ($\Longleftarrow$) generically by proving the contrapositive. Suppose that at μ^{FB} the monopoly profit is positive, that is $(\phi(\mu^{FB}) - c)\mu^{FB} > f$, then:

1. If $(\phi(\mu^{FB}) - c)\mu^{FB}/2 \leq f$, then $\mu^{CO} = \mu^{FB}$, $N^{CO} = N^{FB} = 1$ and thus $\pi^M(\mu^{CO}) > 0$.

2. If $(\phi(\mu^{FB}) - c)\mu^{FB}/2 > f$, then given that from Proposition 5 we know that $\mu^{CO} \geq \mu^M$, and the monopoly profit is single peaked at μ^M (thus, we work in the decreasing part of monopoly profit), we only need to examine whether it is optimal to choose $N^{CO} > 1$ and/or $\mu^{CO} > \mu^{FB}$ in order to satisfy the right CO constraint.

   (a) Assume that for $N > 2$, we do not have that $(\phi(\mu^{FB}) - c)\mu^{FB}/N = f$. We prove that it cannot be $N^{CO} > 1$ with $\mu^{CO} \leq \mu^{FB}$. Suppose by contradiction that the latter holds. Then with $\mu^{CO} = \mu^{FB}$, the left CO constraint cannot bind and $\pi^{CO}(\mu^{CO}) > 0$. If $\mu^{CO} < \mu^{FB}$, then the left CO constraint must bind (and the right not): $(\phi(\mu^{CO}) - c)\mu^{CO}/N^{CO} = f > (\phi(\mu^{CO}) - c)\mu^{CO}/(N^{CO} + 1)$. (To see this, observe that if the left did not bind, we could increase μ^{CO} to bring it closer to μ^{FB} with both constraints still satisfied.) But then consider a new candidate CO solution resulting from reducing N^{CO} by one and increasing μ^{CO} to $\mu^{CO'} > \mu^{CO}$. From the previous left CO constraint we know that the new right CO constraint will not bind. Thus, it has to either be that

$\mu^{CO'} = \mu^{FB}$, in which case (μ^{CO}, N^{CO}) is rejected as a solution and we have a contradiction, or that the new left CO constraint will bind—to see the latter, it suffices to observe that if neither constraint binds and $\mu^{CO'} \neq \mu^{FB}$, there is $\epsilon > 0$ small enough such that either $\mu^{CO'} + \epsilon$ or $\mu^{CO'} - \epsilon$ increases the planner's function. In the case that the new left constraint binds, we have that $(\phi(\mu^{CO'}) - c)\mu^{CO'}/(N^{CO} - 1) = f > (\phi(\mu^{CO'}) - c)\mu^{CO'}/(N^{CO})$, so $\mu^{CO'} < \mu^{FB}$ (consider a similar argument of reducing $\mu^{CO'}$ by ϵ to exclude $\mu^{CO'} > \mu^{FB}$). This case also induces $P(\mu^{CO'}, N^{CO} - 1) > P(\mu^{CO}, N^{CO})$, as $\mu^{CO} < \mu^{CO'} < \mu^{FB}$. We conclude that it cannot be that $N^{CO} > 1$ with some $\mu^{CO} < \mu^{FB}$.

(b) Now consider the case $N^{CO} \geq 1$ with $\mu^{CO} > \mu^{FB}$. Then the right CO constraint must bind (and the left not): $(\phi(\mu^{CO}) - c)\mu^{CO}/N^{CO} > f = (\phi(\mu^{CO}) - c)\mu^{CO}/(N^{CO} + 1)$. To see this, observe that if the right did not bind, we could reduce μ^{CO} to bring it closer to μ^{FB} with both constraints still satisfied. Thus, $\pi^M(\mu^{CO}) = (\phi(\mu^{CO}) - c)\mu^{CO}/N^{CO} - f > 0$. □

Proof of Proposition 5

In the first best case, for given μ, the objective function (17) is decreasing in N. Thus, to economise on fixed costs, the planner allows a monopolistic exchange to provide trading services, and $N^{FB} = 1$. From $\pi^M(\mu^{FB}) \leq 0$ it follows that:

$$(\phi(\mu^{FB}) - c)\mu^{FB} \leq f \implies \frac{(\phi(\mu^{FB}) - c)\mu^{FB}}{N} < f, \ \forall N \in \mathbb{N} \setminus \{1\}.$$

We now establish the technological capacity and liquidity ranking. First, note that it cannot be $N^{CFE} = 1$ and $\mu^{CFE} \geq \mu^{FB}$. This is because for $N^{CFE} = 1$, $\mu^{CFE} = \mu^M$ and by assumption the monopolist makes positive profits. Thus, it will be $\mu^{CFE} < \mu^{FB}$. Finally, by Cournot stability, $\mu^{CFE} \geq \mu^M$ (see also the proof of Proposition 7), with the final implication that $\mu^{FB} > \mu^{CFE} \geq \mu^M$.

We now prove the welfare ranking. Since $\mu^{FB} > \mu^M$ and $\mathcal{P}(\mu, 1)$ is single-peaked at μ^{FB}, it follows that $\mathcal{P}^{FB} > \mathcal{P}^M$. Also, we have that $\mathcal{P}(\mu, 1) = GW(\mu) - c\mu - f$ is single-peaked in μ at μ^{FB}, which means that $GW(\mu) - c\mu$ is also single-peaked. Thus, since $\mu^{FB} > \mu^{CFE}$ we have:

$$\begin{aligned}\mathcal{P}^{FB} = GW(\mu^{FB}) - c\mu^{FB} - f &> GW(\mu^{CFE}) - c\mu^{CFE} - f \\ &\geq GW(\mu^{CFE}) - c\mu^{CFE} - fN^{CFE} = \mathcal{P}^{CFE},\end{aligned}$$

and so overall we have $\mathcal{P}^{FB} > \max\left\{\mathcal{P}^{CFE}, \mathcal{P}^{M}\right\} \geq \min\left\{\mathcal{P}^{CFE}, \mathcal{P}^{M}\right\}$. □

Proof of Proposition 6

Given that the monopoly profit is negative at μ^{FB} (so that it cannot be $\mu^{FB} \leq \mu^{CO}$), it will be $N^{CO} = 1$ and $\mu^{FB} > \mu^{CO}$ so that profits are zero at the CO solution (look also at the proof of Lemma A.2).

We now prove that $\mu^{CO} > \mu^{CFE}$. Suppose, by contradiction, that $\mu^{CFE} \geq \mu^{CO}$. From $\pi^{M}(\mu^{FB}) < 0$ we know that at the Conduct second best $N^{CO} = 1$ and the exchange breaks even, so given that $\psi'(\mu) > 0$ we have:

$$(\phi(\mu^{CFE}) - c)\mu^{CFE} \leq f. \tag{A.51}$$

We first deal with the case where only one firm enters at CFE, and then with the one where $N^{CFE} > 1$. At a CFE with $N = 1$ exchanges, we have $\mu^{CFE} = \mu^{M}$ and thus the monopolist profit is positive by assumption:

$$(\phi(\mu^{CFE}) - c)\frac{\mu^{CFE}}{N} = (\phi(\mu^{CFE}) - c)\mu^{CFE} = (\phi(\mu^{M}) - c)\mu^{M} > f. \tag{A.52}$$

Putting together (A.51) and (A.52) leads to a contradiction.

At a CFE with $N > 1$ exchanges, we need to have:

$$(\phi(\mu^{CFE}) - c)\frac{\mu^{CFE}}{N} \geq f. \tag{A.53}$$

Putting together (A.51) and (A.53) yields

$$f \leq (\phi(\mu^{CFE}) - c)\frac{\mu^{CFE}}{N} < (\phi(\mu^{CFE}) - c)\mu^{CFE} \leq f,$$

a contradiction. Thus, we must have $\mu^{CO} > \mu^{CFE}$.

Now, since $N^{FB} = N^{CO} = 1 \leq N^{CFE}$, it follows that $\mathcal{P}^{FB} > \mathcal{P}^{CO} > \mathcal{P}^{CFE}$ since $\mu^{FB} > \mu^{CO} > \mu^{CFE}$ and $\mathcal{P}(\mu, 1)$ is single-peaked at μ^{FB} and, all else constant, $\mathcal{P}(\mu, N)$ is decreasing in N. □

Proof of Proposition 7

Let $\mu^C(N)$ denote the total co-location capacity at a symmetric Cournot equilibrium for a given number of exchanges N. The objective function of a planner that controls entry can be written as follows:

$$\mathcal{P}(\mu^C(N), N) = N\pi_i(\mu^C(N)) + \psi(\mu^C(N)), \tag{A.54}$$

where $\psi(\mu^C(N))$ denotes the welfare of other market participants at the Cournot solution:

$$\psi(\mu^C(N)) = CE^{SD}(\mu^C(N)) + CE_1^L(\mu^C(N)) + CE_2^L(\mu^C(N)).$$

Consider now the derivative of the planner's objective function with respect to N, and evaluate it at N^{CFE}:

$$\begin{aligned} \left.\frac{\partial \mathcal{P}(\mu^C(N), N)}{\partial N}\right|_{N=N^{CFE}} &= \left.\underbrace{\pi_i(\mu^C(N), N)}_{=0}\right|_{N=N^{CFE}} \\ &+ N^{CFE} \left.\underbrace{\frac{\partial \pi_i(\mu^C(N), N)}{\partial N}}_{<0}\right|_{N=N^{CFE}} + \psi'(\mu^C(N)) \left.\underbrace{\frac{\partial \mu^C(N)}{\partial N}}_{>0}\right|_{N=N^{CFE}} . \end{aligned} \tag{A.55}$$

The first term on the right hand side of (A.55) is null at N^{CFE} (modulo the integer constraint). At a stable, symmetric Cournot equilibrium, an increase in N has a negative impact on the profit of each exchange, and a positive impact on the aggregate technological capacity (see, e.g., Vives (1999)). Therefore, the second and third terms are respectively negative and positive. Given our definitions, N^{CFE} is the largest N such that platforms break even. N^{ST}, instead, reflects the planner's choice of N in Cournot equilibria that keep exchanges from making negative profits and maximizes welfare. Hence, it can only be that

$$N^{CFE} \geq N^{ST} \text{ and } \mu^{CFE} \geq \mu^{ST},$$

since a planner can decide to restrict entry. At a UST, the planner can make side payments to an unprofitable exchange. This has two implications: first, the planner can push entry beyond the level at which platforms break even, so that

$$N^{UST} \geq N^{ST} \text{ and } \mu^{UST} \geq \mu^{ST}.$$

Additionally, depending on which of the two terms in (A.55) prevails, we have

$$\left.\frac{\partial \mathcal{P}(\mu^C(N), N)}{\partial N}\right|_{N=N^{CFE}} \gtrless 0 \implies N^{CFE} \lesseqgtr N^{UST}.$$

Finally, $\mu^C(N) \geq \mu^M$, for $N \geq 1$ because at a stable CFE the total capacity is an increasing function of the number of platforms. A similar argument holds at both the STR and UST, since in this case the planner picks N subject to μ being a Cournot equilibrium

We have that $\mathcal{P}^{UST} \geq \mathcal{P}^{ST}$, because STR imposes an additional constraint on the planner's objective function compared to STR. Finally, $\mathcal{P}^{ST} \geq \mathcal{P}^{CFE}$, because CFE does not account for other traders' welfare, and the planner may choose to favour these market participants when at the margin this creates a larger increase in $GW(\mu)$. □

Proof of Proposition 8

From Propositions 6 and 7 we have $N^{CO} = 1$, $\mu^{FB} > \mu^{CO} > \mu^{CFE} \geq \mu^{ST}$, $\mathcal{P}^{FB} > \mathcal{P}^{CO} > \mathcal{P}^M$, and by Cournot stability $\mu^{ST} \geq \mu^M$. Also, we have that $\mathcal{P}(\mu, 1) = GW(\mu) - c\mu - f$ is single-peaked in μ at μ^{FB}, which means that $GW(\mu) - c\mu$ is so. Thus, since $\mu^{CO} > \mu^{CFE} \geq \mu^{ST}$ we have:

$$\begin{aligned}\mathcal{P}^{CO} = GW(\mu^{CO}) - c\mu^{CO} - f &> GW(\mu^{ST}) - c\mu^{ST} - f \\ &\geq GW(\mu^{ST}) - c\mu^{ST} - fN^{ST} = \mathcal{P}^{ST}\end{aligned}$$

and so $\mathcal{P}^{CO} > \mathcal{P}^{ST} \geq \mathcal{P}^{CFE}$, where the weak inequality follows from the fact that the CFE solution is always available in solving the STR problem. For the same reason, $\mathcal{P}^{FB} \geq \mathcal{P}^{UST} \geq \mathcal{P}^{ST} \geq \mathcal{P}^M$. Thus, overall we have:

$$\mathcal{P}^{FB} > \mathcal{P}^{CO} > \mathcal{P}^{ST} \geq \max\left\{\mathcal{P}^{CFE}, \mathcal{P}^M\right\} \geq \min\left\{\mathcal{P}^{CFE}, \mathcal{P}^M\right\}.$$

Last, evaluate the derivative of welfare with respect to μ at the UST solution:

$$\left.\frac{\partial \mathcal{P}(\mu, N)}{\partial \mu}\right|_{(\mu,N)=(\mu^{UST},N^{UST})} = \left.[\phi'(\mu)\mu + \phi(\mu) - c]\right|_{\mu=\mu^{UST}}$$

The FOC of a firm i in UST reads:

$$\left[\frac{\phi'(\mu)\mu}{N} + \phi(\mu) - c\right]\Bigg|_{(\mu,N)=(\mu^{UST},N^{UST})} = 0.$$

Combining this with the above we have:

$$\frac{\partial\mathcal{P}(\mu,N)}{\partial\mu}\Bigg|_{(\mu,N)=(\mu^{UST},N^{UST})} = \phi'(\mu)\mu\frac{N-1}{N}\Bigg|_{(\mu,N)=(\mu^{UST},N^{UST})} < 0$$

so the UST solution does not maximize welfare given that the FB solution is interior (and thus, for the UST solution to maximize welfare the derivative above should have been zero), so it must be $\mathcal{P}^{FB} > \mathcal{P}^{UST}$. □

B Market shares

Figures B.1 and B.2 replicate the ones that appear in the *OECD Business and Finance Outlook 2016* (Ch. 4, respectively Figure 4.4 and 4.6), with data that reflect the volume for 2018. The purpose of such figures is to illustrate the degree of market fragmentation, by showing the fraction of volume for securities that are listed on a given exchange that is captured by the listing and competing venues. For Europe, the lack of a public, consolidated tape, implies that one has to resort to data provided by private institutions. Specifically, as in the OECD report, we used data provided by "BATS for stocks listed on 12 European major exchanges" in 2018. To interpret: for the UK, BATS data shows that about 63% of the trading in stocks listed at the LSE takes place on that exchange, while about 22% occurs on BATS (CBOE), 9% on Turquoise, and the remaining 6% is split between other lit and dark venues.

For the US, the exercise is facilitated by the existence of a consolidated tape, which offers aggregate information on traded volume on- and off-exchange, and allows to use venue "ownership" as a classification criterion (the "Off-exchange" category includes dark-pools, crossing-networks, systemic internalizers and OTC trading).

(a) Classification of European trading venues

Trading venue	Venue category
CBOE Europe	CBOE
CBOE Europe APA	CBOE
Instinet Blockmatch	Dark volume
ITG Posit	Dark volume
Liquidnet	Dark volume
SIGMA X MTF	Dark volume
UBS MTF	Dark volume
Aquis	Other lit venue
Equiduct	Other lit venue
Bolsa de Madrid	Primary
Euronext	Primary
LSE Group	Primary
Nasdaq OMX	Primary
Oslo	Primary
SIX Swiss Exchange	Primary
Wiener Börse	Primary
Xetra	Primary
Turquoise	Turquoise

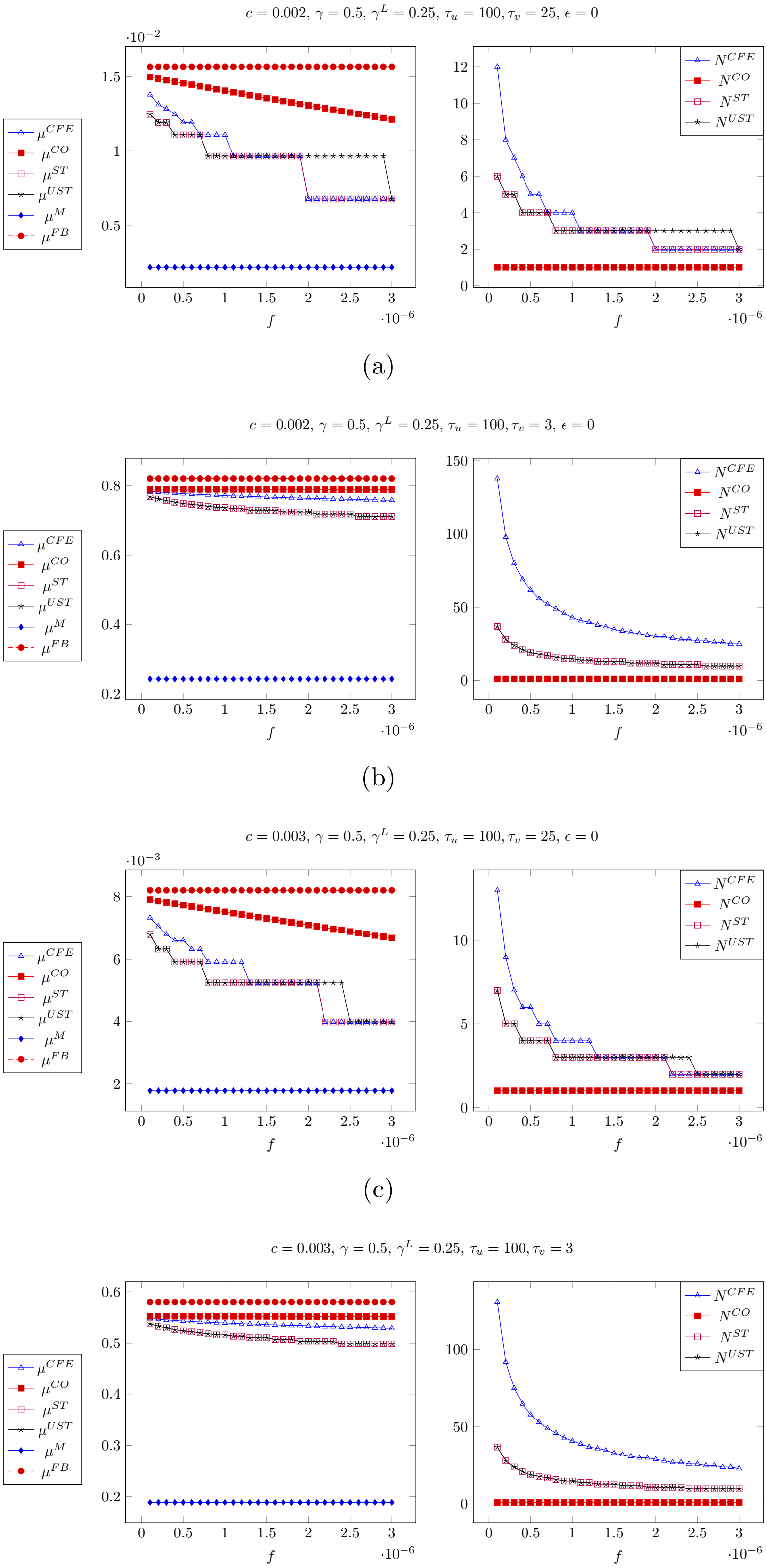
(b) Classification of US trading venues

Trading venue	Venue category
EDGX Equities	CBOE
BZX Equities	CBOE
BYX Equities	CBOE
EDGA Equities	CBOE
IEX	IEX
NASDAQ	NASDAQ
NASDAQ BX	NASDAQ
NASDAQ PSX	NASDAQ
NYSE	NYSE
NYSE Arca	NYSE
NYSE Chicago	NYSE
NYSE American	NYSE
NYSE National	NYSE
NASDAQ TRF Carteret	Off-exchange
NYSE TRF	Off-exchange
NASDAQ TRF Chicago	Off-exchange

Table 1: Trading venues classification for Figures B.1 and B.2.

Initial parametrization	Alternative parameter values		
	c	γ	γ_L
$c = 0.002$, $\gamma = 0.5$, $\gamma_L = 0.25$, $\tau_u = 100$, $\tau_v = 25$	0.001	$\{0.45, 0.35, 0.3, 0.25\}$	0.15
$c = 0.002$, $\gamma = 0.5$, $\gamma_L = 0.25$, $\tau_u = 100$, $\tau_v = 3$	0.003	$\{0.45, 0.35, 0.3, 0.25\}$	0.15
$c = 2$, $\gamma = 25$, $\gamma_L = 12$, $\tau_u = 0.1$, $\tau_v = 0.1$	2.5	$\{22.5, 17.5, 15, 12.5\}$	18

Table 2: Parametrizations used in the simulations. The third row of the table refers to a parametric extension that we have employed in the simulation of the baseline case and the one where SD enter at the second round.


$c = 0.002$, $\gamma = 0.5$, $\gamma^L = 0.25$, $\tau_u = 100$, $\tau_v = 25$, $\epsilon = 0$

(a)

$c = 0.002$, $\gamma = 0.5$, $\gamma^L = 0.25$, $\tau_u = 100$, $\tau_v = 3$, $\epsilon = 0$

(b)

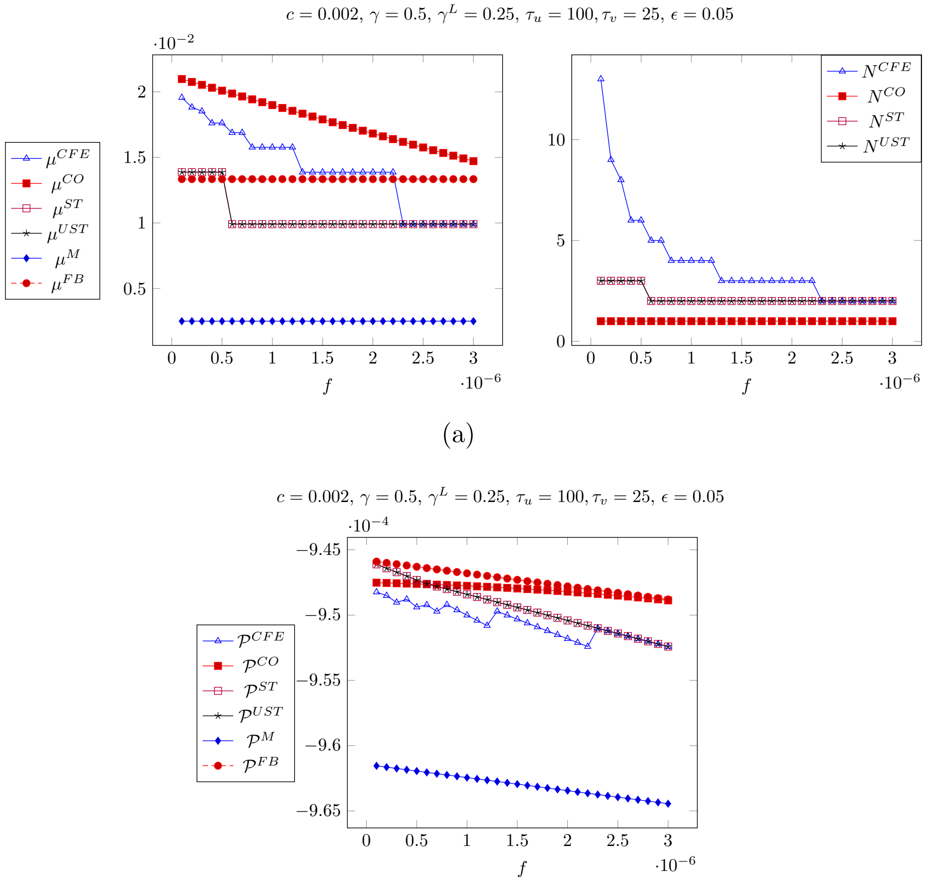
$c = 0.003$, $\gamma = 0.5$, $\gamma^L = 0.25$, $\tau_u = 100$, $\tau_v = 25$, $\epsilon = 0$

(c)

$c = 0.003$, $\gamma = 0.5$, $\gamma^L = 0.25$, $\tau_u = 100$, $\tau_v = 3$

(d)

Figure 2: Panels (a) and (c) illustrate two cases in which insufficient entry occurs. In Panel (b) and (d), entry is always excessive.



$c = 0.002$, $\gamma = 0.5$, $\gamma^L = 0.25$, $\tau_u = 100, \tau_v = 25$, $\epsilon = 0.05$

(a)

$c = 0.002$, $\gamma = 0.5$, $\gamma^L = 0.25$, $\tau_u = 100, \tau_v = 25$, $\epsilon = 0.05$

(b)

Figure 4: The effect of committed dealers on μ^{FD} (panel (a), left), the number of platforms (panel (a), right), and the welfare of market participants (panel (b)).

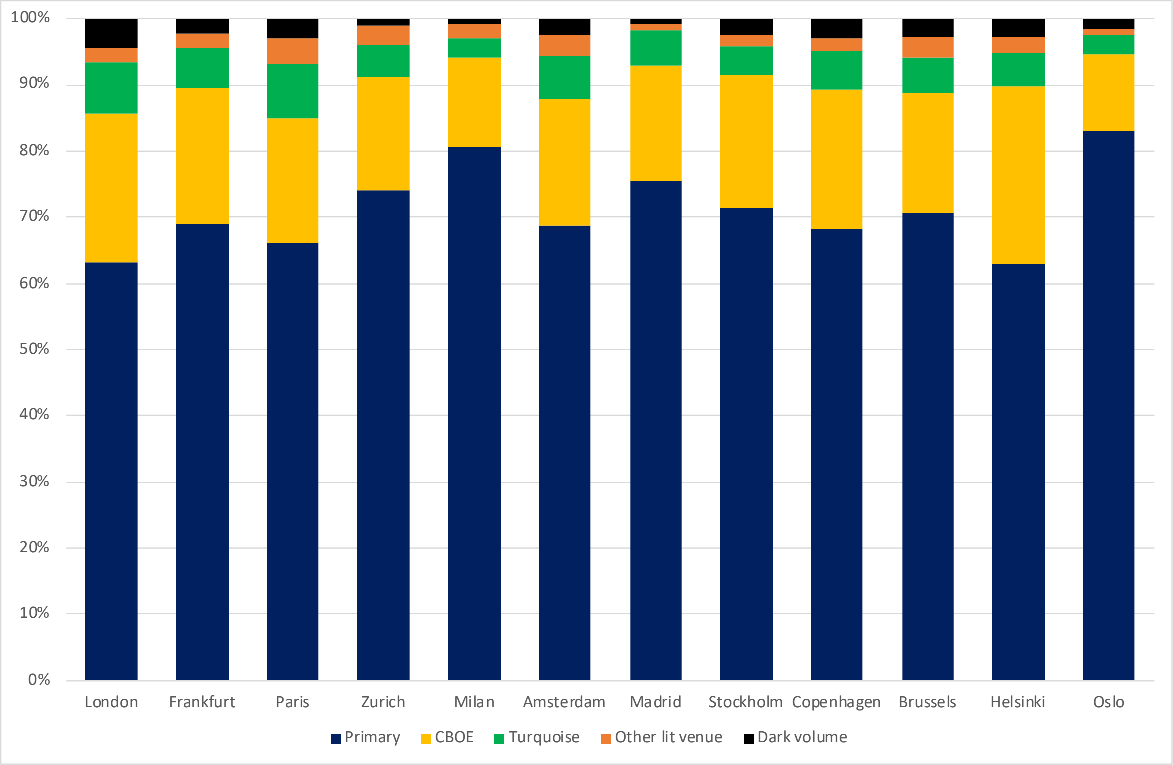


Figure B.1: Market shares among trading venues in Europe in 2018. Source: CBOE Global Markets, own calculations (See Table 1 (Panel (a)) for venues' classification).

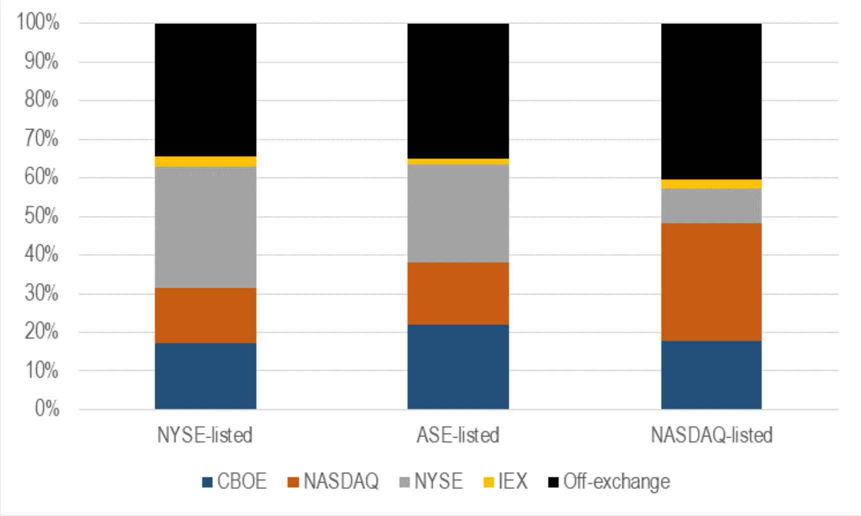


Figure B.2: Market shares among trading venues in the US in 2018. Source: CBOE Global Markets, own calculations (See Table 1 (Panel (b)) for venues' classification).

Result	Order of action	Parametrization
$N^{CFE} < N^{UST}$	OO	$c = 0.002$, $\gamma \in \{0.45, 0.5\}$, $\gamma_L = 0.25$, $\tau_u = 100$, $\tau_v = 25$ $c = 2$, $\gamma \in \{25, 22.5, 17.5, 15, 12.5\}$, $\gamma_L = 12$, $\tau_u = 0.1$, $\tau_v = 0.1$ $c = 2$, $\gamma = 25$, $\gamma_L = 18$, $\tau_u = 0.1$, $\tau_v = 0.1$
	RO	$c = 2$, $\gamma \in \{25, 22.5, 17.5, 15, 12.5\}$, $\gamma_L = 12$, $\tau_u = 0.1$, $\tau_v = 0.1$ $c = 2$, $\gamma = 25$, $\gamma_L = 18$, $\tau_u = 0.1$, $\tau_v = 0.1$ $c = 2.5$, $\gamma = 25$, $\gamma_L = 12$, $\tau_u = 0.1$, $\tau_v = 0.1$
$N^{CFE} < N^{UST} - 1$	OO	$c = 2$, $\gamma \in \{25, 22.5, 17.5, 15, 12.5\}$, $\gamma_L = 12$, $\tau_u = 0.1$, $\tau_v = 0.1$
	RO	$c = 2$, $\gamma \in \{17.5, 15, 12.5\}$, $\gamma_L = 12$, $\tau_u = 0.1$, $\tau_v = 0.1$
$\pi^M\left(\mu^{FB}\right) > 0$, $\mu^{CO} > \mu^{CFE} > \mu^{FB}$	RO	$c = 0.002$, $\gamma = 0.5$, $\gamma_L = 0.25$, $\tau_u = 100$, $\tau_v \in \{3, 25\}$ and all shifts around presented in Table 2
$\pi^M(\mu^{FB}) \leq 0$	OO	$\{(\gamma, \gamma_L, \tau_u, \tau_v, c, f) : \gamma, \gamma_L \in \{1, 2, \ldots, 25\}, \tau_u, \tau_v \in \{0.1, 0.2, \ldots, 10\}, c \in \{0.01, 0.02, \ldots, 0.5\}, f \in \{0.001, 0.002, \ldots, 0.1\}\}$
$N^{CO} = 1$	OO RO	The three parametrizations in Table 2.
Single-peakedness of $\mathcal{P}(\mu, 1)$ and $\pi(\mu)$	OO RO	The three parametrizations in Table 2.

Table 3: Cases where various phenomena are observed. OO and RO refer to the original and reverse order of action models, respectively.