Asset Pricing Frictions in Fragmented Markets*

Emiliano Pagnotta†

January 21, 2013

Abstract

We study the consequences of trading fragmentation and speed on liquidity and asset prices. Exchanges invest in speed-enhancing technologies and price trading services to attract investors. Investors trade due to idiosyncratic preference shocks. We show how the resulting market organization affects asset liquidity and the composition of participating investors. In a consolidated market, speed investments raise liquidity and prices. When markets fragment, liquidity and asset prices can move in opposite directions. We also show how mechanisms that protect execution prices, such as the SEC's trade-through rule, can decrease price levels and trading volume relative to unregulated markets. Our results suggest that recent regulatory reforms in secondary markets may have unintended negative consequences for public corporations.

JEL Codes: G12, G15, G18, D40, D43, D61.

*Earlier versions of this paper circulated under the title "Long-Run Liquidity". I thank Yakov Amihud, Bruce Carlin, Xavier Gabaix, Andrea Eisfeldt, Joel Hasbrouck, Albert Menkveld, Thomas Philippon, Guillaume Rocheteau, Dimitri Vayanos, Pierre-Oliver Weill, Brian Weller, and Finance Dept. seminar participants at the New York University Stern School of Business and the UCLA Anderson School of Management for comments and suggestions. I also thank Joseba Martinez for excellent computational assistance. I acknowledge support form the Smith Richardson Foundation. The usual disclaimers apply.

†New York University Department of Finance. Email: epagnott@stern.nyu.edu

Economists widely accept that market frictions can significantly affect liquidity, and thus have effects on asset prices, aggregate investments and welfare. Technological frictions, for example, can prevent buyers and sellers from being continuously matched over time and trading institutions' market power can limit investor participation in financial markets. During any given day, traders (e.g., households, dealers, or hedge funds) must incorporate these frictions as a constraint in their investment plans. However, one can expect that, over a longer period, the existence of such frictions can provide incentives to developing institutions that would profit by alleviating them.[1] In fact, consistent with this intuition, the last decade has witnessed profound transformations in secondary markets for securities: The speed at which investors trade has greatly increased (see Figure 2), and the number of new stock and derivatives exchanges in the U.S. and Europe has grown significantly, rendering markets more "fragmented" (see Figure 3). Do these transformations enhance liquidity and trading activity? What are the consequences for asset prices? The answers to these questions are the subject of heated debates in academic and policy circles.[2]

We address these issues by studying a tractable framework that characterizes why and how investors with liquidity needs value trading speed, how investment in speed affects competition between trading institutions and investor market participation, and how the interaction between investor and trading institution choices affect long-run liquidity and asset prices. We show how aggregate outcomes critically depend on financial markets' competitive structure and the linkages between trading venues. In particular, we show that when markets fragment, liquidity and asset prices can move in *opposite* directions.

We consider an economy where a continuum of investors ("farmers") consume a constant stream of dividends ("perishable fruits") from their random endowment of a real asset (a non-stochastic Lucas (1978) "tree"). Investors can differ in their valuation of the dividend over time due to idiosyncratic preference shocks, which creates potential gains from trade. The size of these shocks can differ from investor to investor in a mean-preserving fashion, as they would if some agents were more *leveraged*. Investors cannot buy or sell the asset directly; they can only do so by participating in costly trading venues, generically denoted here as exchanges. Participating investors gain repeated access to a given exchange to trade, but with a certain time delay. Trading speed allows investors to readjust holdings more effectively in response to shocks and thus realize higher gains from trade. Exchanges

[1] Amihud, Mendelson, and Pedersen (2006), for example, argue that "Alleviating frictions is costly ... and the institutions which alleviate frictions may be able to earn rents. For instance, setting up a market requires computers, trading systems, clearing operations, risk and operational controls, legal documentation, marketing, information and communication systems, and so on."

[2] See, for example, the discussion in the SEC's 2010 Concept Release on the U.S. market structure, and the discussions following the recent creation of the Joint Commodity Futures Trading Commission–SEC Advisory Committee on Emerging Regulatory Issues.

therefore have incentives to earn rents by investing in trading technologies that enhance speed.

When all trading is consolidated in a single exchange, speed alleviates market contact frictions, enhances asset liquidity and thus raises the equilibrium asset price. This result is both intuitive and consistent with empirical findings.[3] However, market power gives a monopolist exchange the ability to limit investor participation by raising access fees. Investors with identical mean valuations for the asset but unequal leverage may rationally make different market participation decisions. As access costs increase, only the more highly levered investors (for whom gains from trade are large) join the market. Investor participation frictions thus alter the composition of traders in the market, which distorts the market clearing price. In particular, the equilibrium asset price can increase with participation costs and be higher than its frictionless Walrasian counterpart.[4]

Do frictions generate the same outcomes in economies with fragmented trading? We show in a simple two-exchange economy that speed-driven competition further reduces average trading delays, encourages investor participation, and increases traded volumes. All these factors enhance asset liquidity vis-à-vis the consolidated market. However, the equilibrium effect on asset prices may be surprisingly different from the single-exchange case. Competition between exchanges, while leaving the relation of tradable assets to market participants unaltered, results in lower market access fees and thus increases the fraction of participating investors. As this fraction increases, the market average leverage decreases and the marginal participant (the one that clears the market) sell her endowment to a pool of investors whose average gains from trade become smaller. Provided the asset supply is not "too large,"[5] the marginal investor's valuation decreases and fragmentation can thus lead to lower asset prices.

These results suggest an interesting empirical relation: When an entrant exchange breaks a monopoly and trading fragments, asset liquidity may *increase* and the asset price level may *decrease*. The relative strength of asset liquidity and investor composition effects on the asset price depend on the ability of exchanges to vertically differentiate their liquidity services (Gabszewicz and Thisse (1979)) and thus relax cost-based competition compared to a Bertrand-like outcome.

Our analysis is closely related to recent empirical work that investigates the impact of trading fragmentation. For example, O'Hara and Ye (2011) find evidence that fragmentation has

[3] See, for example, the single-exchange empirical results of Hendershott, Jones, and Menkveld (2011) who study fast (algorithmic) trading.

[4] This effect of costs on asset prices is of different nature from that stated in Vayanos (1998), where trading costs affect trade sizes and holding periods but not the composition of investors in the market.

[5] If asset supply is sufficiently large relative to investor demand, the asset price reflects the minimum investor valuation regardless of the market structure.

a beneficial effect on widely used measures of "market quality" related to liquidity, such as bid-ask spreads and execution delays.[6] These findings seem to support policymakers in the U.S. and other countries who have encouraged fragmentation in recent years.[7] While our theoretical results are consistent with the findings on liquidity, they also suggest that one should be cautious when interpreting such findings as evidence that fragmentation is enhancing financial markets. Fragmentation may have broader consequences for asset prices, and thus have unintended impacts on related economic variables such as the cost of capital for corporations in primary markets.

When an asset trades in different markets, its price can in principle be different in each market, depending on the degree of integration between venues. Arbitrageurs can, of course, work to move prices closer to each other, but their ability to do so is subject to well-recognized frictions. Policymakers have thus designed mechanisms that address this issue directly, motivated by "protecting investors" from unfavorable prices. In the U.S. equity market such a mechanism is applied via the SEC's "trade-through" rule (Rule 611 in Regulation National Market System, hereafter Reg NMS), which essentially mandates the integration of price formation across markets and gives investors access to the "national best" price independently of their trading location[8] (see Appendix A). Although many observers agree that the trade-through rule had a profound impact on U.S. equity markets since its implementation in 2007, its precise effects on financial markets are less clear. We show that a trade-through rule affects competition between exchanges, redistributes investors across them and, importantly, favors the most illiquid market. The resulting distortions in competition affect asset prices. We show that the national best price lies in between the prices that would result in speed-differentiated exchanges were they perfectly segmented. Moreover, when speed investment costs are moderate, the national best price is lower than the volume-weighted average price (VWAP) in segmented markets.

By considering countries with different financial markets, our results provide several international asset pricing predictions. Consider, for example, an asset with an identical payoff structure that trades in different countries. Our results suggest that the trading price should be higher in economies where there are single exchanges (e.g. China, Spain and Brazil) than in economies with fragmented markets (e.g., the U.K, France and the Netherlands), and

[6] Foucault and Menkveld (2008) and Degryse, Jong, and Kervel (2011) find similar results analyzing European markets. The latter study distinguishes between dark and lit fragmentation, and reports that only lit fragmentation enhances liquidity.

[7] For example, the SEC motivates Reg NMS by stating that: "Mandating the consolidation of order flow in a single venue would create a monopoly and thereby lose the important benefits of competition among markets. The benefits of such competition include incentives for trading centers to create new products, provide high quality trading services that meet the needs of investors, and keep fees low."

[8] Following this concept, the Canadian market regulator, the Investment Industry Regulatory Organization of Canada, adopted a similar rule in 2011.

lowest in countries with fragmented markets *and* price protection (the U.S. or Canada).

At the core of the model lies the idea that, everything else being equal, all investors are (at least weakly) better off by trading faster, but not all investors value speed equally. Trading venues can thus make efforts to differentiate their trading platforms to cater to different clienteles. For example, a retail investor or pension fund manager may not be very sensitive to speed changes measured in small sub-second time intervals. On the other hand, for sophisticated investors such as derivatives market makers or equity index arbitrageurs, speed may be central to their business model. This fact motivates the modeling of heterogeneous agents. We capture investor heterogeneity in a parsimonious fashion by making private valuation processes heteroskedastic, which we interpret as originating in investor-specific leverage levels.[9] Using our building idea, we seek to analyze several market designs (summarized in Table I) within an integrated framework. To understand the shared underlying economics, consider the basic three-stage sequence represented in Figure 1. Each stage represents a conceptually different period in the spirit of Alfred Marshall's (1890) theory of production classification[10]:

- Trading period: All determinants of liquidity are taken as given by investors. One or more continuous-time markets for the asset open and participants trade.
- Short-run: Given trading technologies, one or two exchanges set access fees. Investors make participation and trading location decisions, and pay fees accordingly.
- Long-run: Given the number of exchanges, speed investment decisions are made.

For example, the trading period can represent any given trading day, the short-run may represent one or more quarters, and the long-run one or more years. In Section IV.B we add a regulation component to the long-run analysis to study market linkages.

To understand the pricing consequences of market designs and investments, we derive instructive decompositions of equilibrium prices and calibrate the model using U.S. equity data from the 2000s. In our framework, deviations of the market price from a frictionless Walrasian counterpart are captured by two quantities. The first is an illiquidity discount (ILD) that reflects the degree of imperfection in market contact, and is driven by trading

[9] Leverage is a universal feature of financial markets, the banking sector, and real markets (e.g. mortgages in real estate equity). Admittedly, our analysis of leverage is stylized since we do not model the origin of leverage differences. Frazzini and Pedersen (2010) study a setting where leverage differences arise due to institutional constraints.

[10] According to Marshall, some but not all production inputs can change in the short-run. Investor participation but not trading technologies, both inputs of a hypothetical liquidity production function, can adjust in our model's short-run. Our trading period corresponds to what Marshall called the market period, a period during which all input supplies are fixed.

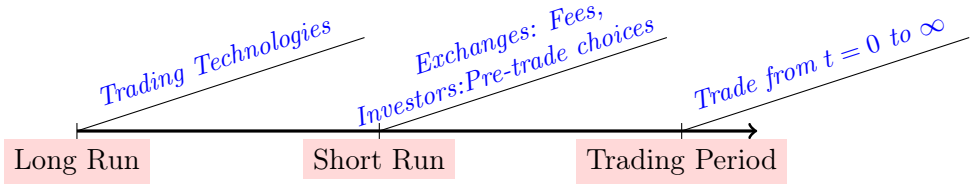

Figure 1: Timing and Structure of the Model

technologies, as well as its shadow valuation, which depends on the marginal investor in the market. The second, which we label limited participation distortion (LPD), corresponds to the difference between the market and Walrasian prices in the absence of market contact frictions. This term arises solely because of the exchange or exchanges' market power, and partly reflects how levered market participants are.

To illustrate the calibration results, let us consider the effect of investments in trading technologies, keeping the number of markets fixed. We find that in the long-run, investments increase the asset price nearly 3.2% in a single exchange economy, but only 1.3% on average in a duopolistic economy. These values reflect the fact that speed is more valuable to highly levered investors, who populate the monopolist market. Consider next an economy where an entrant exchange breaks a monopoly. We find that the equilibrium asset price falls by nearly 12% in the short-run, and 14% in the long-run. The effect in the short-run is dominated by the LPD, which sharply decreases in the transition to a duopoly. In the long-run, investments allow exchanges to improve asset liquidity, reducing the ILD. Moreover, vertical differentiation of their liquidity services allows exchanges to relax Bertrand competition and increase access fees, which partially offsets the initial effect on the LPD. Finally, consider a fragmented market where a trade-through rule is implemented. We find that the national best price is in the long-run 128 basis points (bps) lower than the unregulated average price. Turning to the market environment, we find that in the long-run prices are negatively related to the cost of trading technologies. Interestingly, such costs only affect market participation when markets are fragmented. An increase in investors average leverage raises asset prices in both the short and long-run. The effect is stronger in the long-run due to the fact that exchanges can extract higher rents from investors by investing in faster platforms, rendering the asset more liquid.

Our model also provides novel predictions about trading fragmentation levels. Using a simple Herfindahl-Hirschman Index (HHI) to measure fragmentation, we find that its level (i) decreases as technology becomes cheaper, (ii) increases with positive technology shocks, (iii) decreases with more frequent preference shocks, and (iv) is higher under price protection than in unregulated markets. Turning to trading volume, we find that fragmentation can

Table I: Organization of the Liquidity and Asset Markets

Trading Linkages	Consolidated	Fragmented	
		Segmentation	Investor Protection
Liquidity Markets ('Exchanges')	1	2	2
Asset Markets	1	2	1

have significant equilibrium effects. In the long-run equilibrium, fragmented markets achieve approximately 88% of the Walrasian volume, which is more than twice the volume of a monopolist exchange. The volume effect of investor protection is negative but moderate.

Although we largely concentrate on exchanges' investments, we also analyze the effect of investors acquiring fast trading technologies directly (Section III.B). When this is possible, speed investments naturally depend on investors' characteristics, yielding a market with *heterogeneous frequency* traders. We show that in this case, asset prices may *increase* with asset supply: Unlike in Walrasian analysis where the market structure is given, an increase in supply can incentivize investors to become more technologically sophisticated, rendering assets more liquid, and increasing prices in equilibrium. The economic intuition is related to Acemoglu (1998) in the context of labor economics and directed technological change.

In summary, this paper makes four key points about secondary markets' frictions and asset pricing. First, investors with identical mean valuations for the asset but heterogeneous leverage can rationally make different market participation decisions. When exchange institutions have market power, the resulting distortion in the composition of active participants can drive the asset price higher than its frictionless level. This effect stresses the fact that equilibrium valuations do not directly reflect the stream of cash flows but, rather, the consumption stream that asset ownership generates. Second, the relationship between liquidity and price evolutions crucially depends on the competitive structure in financial markets. Liquidity can increase over time because of technical progress or increased competition between markets. However, when markets fragment, asset prices can evolve in the opposite direction. This result may shed new light on recent empirical results that link fragmentation with enhanced market quality. Third, regulations that protect investor executions distort competition between trading venues and thus have equilibrium consequences on asset prices. The result here should be of interest for policy makers evaluating optimal regulations in this regard. Finally, our results illustrate that differences in leverage can have important asset-pricing effects and can amplify or diminish illiquidity effects. Understanding their effects on financial markets is crucial, given the increasingly important role of institutions in portfolio and investment decisions.

Relation to the Literature Garbade and Silber (1977) provide early work on the issue

of separated markets and the role of speed.[11] Theoretical analyses of fragmentation include those of Mendelson (1987), Pagano (1989), and Madhavan (1995). These early papers focus on the trade-off between liquidity externalities and market power. Our focus on differentiation is in the spirit of Harris (1993), who argues that markets fragment partly because not all traders solve the same problem. Importantly, these papers analyze liquidity and price informativeness. We contribute to this literature by characterizing the effect of fragmentation on equilibrium asset price levels. To the best of our knowledge, we also provide the first formal analysis of the effects of investor protection on asset prices.

This paper centers around liquidity and the pricing implications of institutions' responses to alleviate frictions. A different focus is adopted by Pagnotta and Philippon (2012) in a companion paper. These authors study the welfare implications of speed-driven competition, the entry of exchanges, and optimal market design. Since our focus is on positive pricing and quantity implications, our study examines price formation where investors are also able to acquire technologies to reduce their individual "distance" to a market center. Appendix B provides a version of the model with generalized asset holdings and preferences.

Exchanges can differentiate in areas other than speed.[12] For example, Santos and Scheinkman (2001) study competition in margin requirements, while Foucault and Parlour (2004) analyze competition in listing fees.[13] These papers do not analyze speed differentiation and thus justifiably consider static frameworks. Our focus on technological speed reflects its prominent role in modern asset markets and its direct relation with secondary market liquidity and explains our effort in developing a suitable dynamic model where speed plays an explicit role.

Our paper complements the vast literature that analyzes the asset-pricing effects of illiquidity. Important theoretical work in this area includes that of Amihud and Mendelson (1986), Constantinides (1986), Vayanos (1998), Lo, Wang, and Mamaysky (2004), Eisfeldt (2004), Acharya and Pedersen (2005), among many others.[14] We contribute to this literature by developing links between the origin of liquidity in secondary markets (i.e. competitive structure, investor participation and technology) and asset price levels. Our trading model

[11] Silber and Garbade offer a historical perspective of speed, that surprisingly resembles in spirit many current developments. They analyze the impact of the domestic telegraph system between the New York Stock Exchange (NYSE) and the Philadelphia Stock Exchange during the 1840s and the transatlantic cable between New York and London in 1866.

[12] Models of vertically differentiated oligopolies have been pioneered by Gabszewicz and Thisse (1979) and Shaked and Sutton (1982). Unlike classical models, we endogenize the value of "quality" (trading delays here) through a microfounded trading game. In particular, we enrich the basic setup by having investors "consuming" a differentiated product first (liquidity) and a homogeneous product (asset cash flows) second.

[13] Biais (1993), Glosten (1994), Hendershott and Mendelson (2000) and Parlour and Seppi (2003) study competition between markets with different trading rules. More recently, Colliard and Foucault (2012) study the effect of trading fees in a context where an exchange competes with an over-the-counter dealer.

[14] Amihud, Mendelson, and Pedersen (2006) provide a thorough survey.

approach is closest to the part of this literature that incorporates search-like frictions, which has been fostered by Duffie, Garleanu, and Pedersen (2005).[15] In particular, our trading model is in the spirit of Lagos and Rocheteau (2009), whose model we extend by incorporating heterogeneously levered agents.[16] We complement this literature by endogenizing several important aspects of the market structure. To the best of our knowledge, this paper is the first within this class to analyze liquidity and pricing of a single security across different trading venues.[17]

Key to our results are participation decisions and the composition of investors in the market.[18] Consistent with the model's outcomes, using microdata, Mankiw and Zeldes (1991) and Vissing-Jorgensen (2002) find that the preferences of stock market participants differ greatly from those of non-market participants.

The remainder of the paper is organized as follows. Section I presents the empirical motivation for our model. Section II introduces the theoretical model in the case of a single trading venue and Section III derives the theoretical results for asset prices in this setting. Section IV analyzes competition among a given set of trading venues, the allocation of investors across these venues, and the resulting asset prices. It also analyzes the effects of investor protection on asset prices. Section V presents empirical implications for asset prices, trading volume and fragmentation, as well as a calibration that illustrates the size of the effects. Section VI discusses relationships between our results and previous findings. SectionVII concludes the paper. Appendix A describes different investor protection regulations and Appendix B develops an extension of the trading model with unrestricted asset portfolios. All proofs are provided in the Online Appendix.

I Empirical Trends That Motivate the Theory

This section presents a brief account of the empirical facts that motivate our theory.[19] It also illustrates some of the speed choices that investors face in modern markets. See Appendix

[15] Trejos and Wright (2012) discuss this growing literature, see the references therein.

[16] While the search literature in finance mainly concentrates on over-the-counter markets, Weill (2007) and Biais, Hombert, and Weill (2012) also use a related framework to analyze trading in exchanges. Garleanu, Pedersen, and Poteshman (2009) also uses Poisson contact times to analyze the effect of trading frictions on prices.

[17] Vayanos and Wang (2007) and Weill (2008) use multiple assets but not multiple trading venues.

[18] Contributions to the analysis of participation costs in financial markets include the works of Brennan (1975), Merton (1987), Allen and Gale (1994), Chatterjee and Corbae (1992), and more recently, Huang and Wang (2009). Within the search tradition, Afonso (2011) studies a version of the model of Vayanos and Wang (2007) with flow investor entry in the presence of thick market and congestion externalities.

[19] The Securities and Exchange Commission (2010) presents an informative summary of these trends and a discussion of the challenges they pose to market quality.

A for a discussion of some of the important regulations that affect the evolution of speed and fragmentation.

Trading Speed. Major market centers have made costly investments in fast, computerized trading platforms to reduce order execution and communication latencies. This process has gone beyond stock exchanges to include futures, options, bonds, and currencies. Although these investments were chiefly first observed in the U.S., they have more recently been undertaken on a global scale. This trend accelerated during the second half of the 2000s, as Figure 2 illustrates with the average execution speed in the NYSE. The driving forces underlying this speed frenzy are likely to be different from those of other historical periods. In the human-driven trading era, for example, higher execution speeds helped reduce moral hazard problems between, say, floor brokers and investors. However, this aspect has become less relevant today.

In our model, some investors decide to pay a premium for speed. In real markets this idea relates to several dimensions of the investment process. Some important examples are the following.

- *Colocation.* Colocation is a service offered by trading centers that operate their own data centers and by third parties that host the matching engines of trading centers. The trading center rents rack space to investors, which enables them to place their servers in close physical proximity to a trading center's matching engine, which helps minimize network and other types of delays.

- *Exchanges versus alternative venues.* According to the SEC classification, U.S. investors can opt to direct their trading orders to registered exchanges (such as the NASDAQ), Electronic Communication Networks, dark pools, or Broker-Dealer Internalizers. Although alternative and over-the-counter venues have also made significant technological progress, organized exchanges typically offer investors the fastest communication and trading responses.

- *Markets with speed restrictions.* In foreign exchange markets, for example, many sophisticated traders concentrate on Electronic Broking Services and Reuters inter-dealer brokerage platforms, both of which have minimum quote life or minimum fill ratios. One exchange that does not have such a minimum is Currenex, which is therefore particularly attractive to high-frequency traders.

- *Inter-market connectivity.* The provision of connections between financial centers plays a key role in modern markets. Take Chicago-New York as an example, an essential

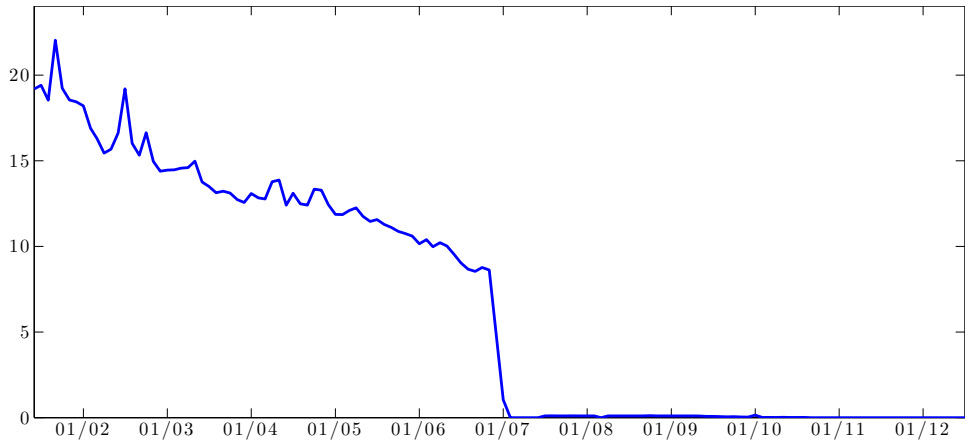

Figure 2: NYSE Average Order Execution Speed (Seconds)
Source: NYSE SEC Rule 605 reports (small orders executed at market; 5% tails excluded from the average).

route for derivative traders. In 2011 a firm named Spread Networks invested approximately $300 million in a new fiber optic cable that links these cities through the straightest possible route, saving about 100 miles over existing routes. This allows the company to shave 6 milliseconds off their delay, for a total delay of 15 milliseconds.[20] Clients of these superior connectivity services are willing to pay premium fees for their use.

Fragmentation. Securities trading, especially in North America and Europe, has become significantly more fragmented over the last decade. Figure 3 illustrates this trend. Traditional markets in the U.S. and Europe have lost market share to (usually faster) entrants such as Direct Edge, BATS and Chi-X. For instance, the fraction of NYSE-listed stocks traded at the NYSE decreased from 80% in 2004 to just over 20% in 2009. Similar patterns have been observed in other asset classes. For example, there are more than 10 options exchanges in the U.S. alone. Overall, fragmentation has increased so dramatically that market participants now keep track of fragmentation indexes across asset classes and countries.[21] Figure 3 also illustrates that the levels of fragmentation in other regions of the world are still well behind those of North America and Europe. Given that several developing countries are currently trying to foster competition between exchanges, the results in this paper may help them anticipate effects on prices, liquidity and trading volumes.

[20]The success of this business model motivated McKay Brothers, a leading provider of low-latency wireless transport equipment, to contract the creation of a $300 million microwave-based network of towers connecting the same cities to Aviat Networks.

[21]See, for example, the Fidessa fragmentation indexes

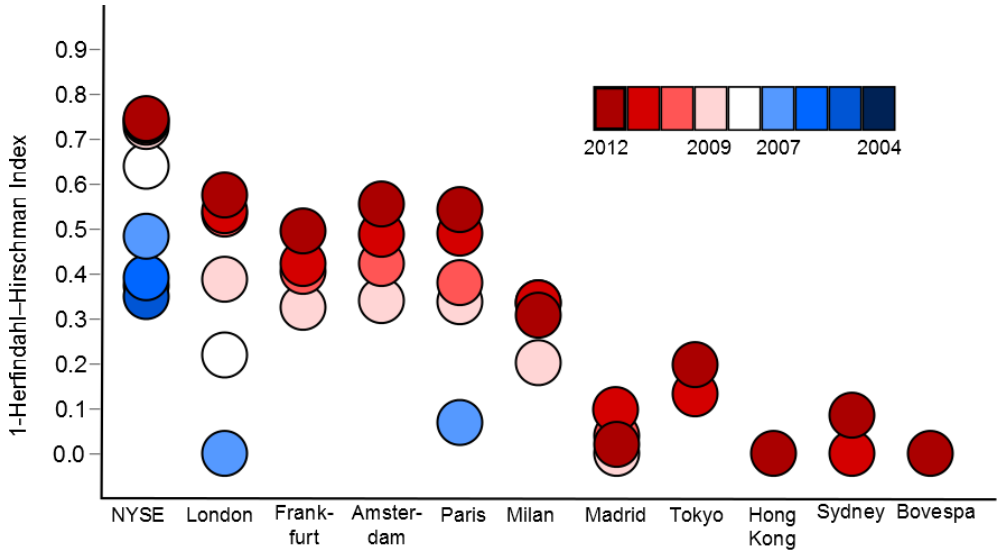

Figure 3: Changes in Fragmentation Levels in Global Equity Markets

The Herfindahl-Hirschman Index of market concentration (the sum of squares of market shares; e.g., for the NYSE in 2012, the sum of squares of market shares in NYSE-listed stocks trading for the NYSE, Nasdaq, BATS, DirectEdge and others) for a cross-section of equity exchanges. Market share is defined as the proportion of exchange-listed shares traded through the exchange itself and excludes dark pools.

Sources: Exchanges websites, RBC Global Asset Management, Bloomberg, and HSBC Research.

II Trading Model

This section provides explicit micro foundations for how investors demand speed in financial markets. We first analyze the problem of an investor trading in a single cost-free market and characterize the ex ante value of speed. We then characterize an equilibrium where investors decide whether to participate in such market by paying an access fee and a single exchange sets the fee to maximize profits.

A Preferences and Technology

We start by describing the main building blocks of our model: investor preferences and trading technology. The preferences need to incorporate heterogeneity to create gains from trade as well as interesting participation decisions among exchanges. The trading technology must capture the role of speed in financial markets.

Time is continuous and runs forever. The model has a continuum of heterogeneous investors, two goods, and one asset. The measure of investors is normalized to one and their preferences are quasilinear. The numéraire good (cash) has a constant marginal utility normalized to one and can be freely invested at a constant rate of return r. We restrict asset holdings to

$a_t \in \{0, 1\}$.[22]The asset is in fixed supply, $\bar{a}$, which is normalized to lie in $[0, 1]$ so that it also represents the expected endowment of each investor. That is, before trading, each investor is endowed with one unit of the asset with probability $\bar{a}$.

One unit of asset pays a constant dividend equal to one of a perishable non-tradable good. The flow utility that an investor derives from holding a_t units of the asset at time t is

$$u_{\sigma,\epsilon_t}(a_t) = (\mu + \sigma\epsilon_t)\, a_t, \tag{1}$$

where (σ, ϵ_t) denotes the type of investor. This type is defined by a fixed component σ and a time-varying (random) component ϵ_t.[23] The fixed component $\sigma \in [0, \bar{\sigma})$ is known at time zero and distributed according to the twice-differentiable cumulative distribution G, with a log-concave density function g that is positive everywhere. The type $\epsilon_t \in \{-1, +1\}$ changes randomly over time. The times when a change can occur are distributed exponentially with parameter γ. Conditional on a change, the temporary shocks are independent and identically distributed, and the value $\epsilon = 1$ occurs with probability $\phi(1) = \phi \in (0, 1)$.

Our paper focuses on the trading technology for the asset. For clarity, we describe here the case where all investors trade at the same speed (later we endogenize speed choices and consider markets with different speeds). The market where investors trade the asset is characterized by a constant contact rate ρ. Conditional on being in contact, the market is Walrasian and clears at a price p. Investors that are not in contact simply keep their holdings constant.

Our assumptions about technology and preferences imply that the value function of a class-σ investor, with current valuation $\epsilon(t)$, and current asset holdings a at time t is

$$V_{\sigma,\epsilon_t}(a, t) = \mathbb{E}_t\left[\int_t^T e^{-r(s-t)} u_{\sigma,\epsilon_s}(a) ds + e^{-r(T-t)}\left(V_{\sigma,\epsilon_T}(a_T, T) - p_T(a_T - a)\right)\right], \tag{2}$$

where the realization of the random type at time $s > t$ is $\epsilon(s)$, T denotes the next time the investor makes contact with the market, and a_T corresponds to optimal time T holdings. Expectations are defined over the random variables T and $\epsilon(s)$ and are conditional on the current type $\epsilon(t)$ and the speed of the market ρ.

[22]See Appendix B for a generalization of the model where the equilibrium price is found relaxing this assumption

[23]All random variables are defined on a probability space $(X, \mathcal{F}, \Pr)$, and all random variables at time t are measurable with respect to the filtration $\{\mathcal{F}_t : t \geq 0\}$ representing the information commonly available to investors.

B Trading Equilibrium and the Value of Speed

We characterize here the trading equilibrium with free market participation where all investors join the market. We show that the asset price remains constant during the trading game, and the value functions are thus time independent and satisfy

$$rV_{\sigma,\epsilon}(a) = u_{\sigma,\epsilon}(a) + \gamma\phi\left(\epsilon' \neq \epsilon\right)\left[V_{\sigma,\epsilon'}(a) - V_{\sigma,\epsilon}(a)\right] + \rho\left[V_{\sigma,\epsilon}\left(a^*_{\sigma,\epsilon}(p)\right) - V_{\sigma,\epsilon}(a) - p\left(a^*_{\sigma,\epsilon}(p) - a\right)\right]. \tag{3}$$

To analyze Equation 3, we then need to characterize the demand functions $a^*_{\sigma,\epsilon}(p)$. Note that, on average, a proportion ϕ of investors are of trading type $\epsilon = +1$ representing natural buyers, who are on the long side of the market when $\phi \geq \overline{a}$. Following Duffie, Garleanu, and Pedersen (2005), we concentrate on the case where supply is short hereafter and treat the complementary case in Appendix B.[24] To simplify the exposition we consider a symmetric shock structure by setting $\phi = 1/2$. Since supply is short, low-σ types always sell their entire holdings when they contact the market. Moreover, there is a marginal type $\hat{\sigma}$ that is indifferent between buying and not buying when $\epsilon = 1$. Thus, we prove that when $\phi > \overline{a}$ the demand system is as follows.

Lemma 1. *The demand functions are* ${a_{\sigma,\epsilon}}^* = 0$ *when* $\epsilon = -1$ *or when* $\sigma < \hat{\sigma}$*, and* ${a_{\sigma,\epsilon}}^* = 1$ *when* $\epsilon = +1$ *and* $\sigma \geq \hat{\sigma}$*, where*

$$\hat{\sigma}(p,\rho) \equiv \left(1 + \frac{\gamma}{r+\rho}\right)(rp - \mu). \tag{4}$$

Clearly, there cannot be an equilibrium where $a_{\sigma,-} = 1$ since in that case total demand would exceed $\overline{a}$ preventing market clearing. Note also that the asset holdings of types $\sigma < \hat{\sigma}$ are non-stationary since they never purchase the asset. A type $\sigma < \hat{\sigma}$ sells its holdings on first contact with the market and never holds the asset again. The value of $\hat{\sigma}$ then naturally classifies investors into two categories, which we label as temporary investors ($\sigma < \hat{\sigma}$) and active ($\sigma \geq \hat{\sigma}$) investors.[25] Over time the assets move from the low-σ types to the high-σ types and then keep circulating among high-σ types in response to ϵ shocks and trading opportunities.

It is easy to see that the price remains constant along the transition path. Given the ex ante supply of the asset $\bar{a}$ and the market contact rate ρ, the gross supply of assets is always

[24] Although binary asset holdings simplify the analysis greatly, this assumption is not without loss of generality. Like in Duffie, Garleanu, and Pedersen (2005), asset indivisibility generates an equilibrium price that decreases in the contact rate ρ when asset supply is large, making such case somewhat less compelling. We know from Lagos and Rocheteau (2009) that this extensive margin considerations are trivial when asset holdings are unrestricted. We provide an extension of our framework with $a \geq 0$ in Appendix B.

[25] An alternative classification involves small- and high-volume investors, which would be natural under a generalization with unrestricted asset holdings.

$\rho\bar{a}$. All negative trading types $\epsilon = -1$ want to hold $a = 0$ and they represent half of the traders. The trading types $\epsilon = +1$ want to hold one unit if $\sigma > \hat{\sigma}$ and nothing if $\sigma < \hat{\sigma}$. The gross demand from high types is then always equal to $\rho\left(1 - G\left(\hat{\sigma}\right)\right)/2$. The market clearing condition, which holds at all times, is therefore

$$\frac{1}{2}\left[1 - G\left(\hat{\sigma}\right)\right] = \bar{a}. \tag{5}$$

Let $\alpha_{\sigma,\epsilon}\left(a\right)$ be the share of class-σ investors with trading type ϵ currently holding a units of asset. Using the demand system in Lemma 1, and the market clearing condition 5, we can now characterize the steady-state distribution among types $\sigma > \hat{\sigma}$ and the single-market equilibrium price as follows.

Proposition 1. *The single market, full-participation trading equilibrium is as follows.*

- *When $\overline{a} < \frac{1}{2}$ there is a unique equilibrium price given by*

$$p = \frac{\mu}{r} + \frac{G^{-1}\left(1 - 2\overline{a}\right)}{r}\left(\frac{r+\rho}{r+\gamma+\rho}\right) \tag{6}$$

- *Transition dynamics: The price remains constant while asset holdings shift from low σ-types to high σ-types. Low types $(\sigma < \hat{\sigma})$ sell their initial holdings and never purchase the asset again. High types $\sigma \geq \hat{\sigma}$ buy when $\epsilon = 1$ and sell when $\epsilon = -1$.*

- *The distribution of holdings among high σ-types converges to the steady-state distribution of well-allocated assets $\alpha_{\sigma,+}\left(1\right) = \alpha_{\sigma,-}\left(0\right) = \frac{1}{4}\frac{2\rho+\gamma}{\gamma+\rho}$ and misallocated assets $\alpha_{\sigma,+}\left(0\right) = \alpha_{\sigma,-}\left(1\right) = \frac{1}{4}\frac{\gamma}{\gamma+\rho}$.*

- *When $\rho \to \infty$, allocations and the equilibrium price converge to the Walrasian outcome $\alpha_{\sigma,+}\left(0\right) = \alpha_{\sigma,-}\left(1\right) = 0$ and $p_W = \frac{1}{r}\left[\mu + G^{-1}\left(1 - 2\overline{a}\right)\right]$.*

With full participation the value of $\hat{\sigma} = G^{-1}\left(1 - 2\overline{a}\right)$ can be derived directly from the market clearing condition, and thus only depends on the asset supply and the distribution of investor types. Consequently, the frictionless Walrasian price depends only on these variables.

Our next task is then to compute the value of speed for investors. We proceed in two steps: we first compute the steady-state value functions for active investors, and later compute the ex ante values, taking into account the transition dynamics. Consider the steady-state value functions for any type $\sigma > \hat{\sigma}$. Given the Bellman equation of Equation 3 and Lemma 1, for

the types holding assets, we have

$$rV_{\sigma,+}(1) = \mu + \sigma + \frac{\gamma}{2}\left[V_{\sigma,-}(1) - V_{\sigma,+}(1)\right] \tag{7}$$

$$rV_{\sigma,-}(1) = \mu - \sigma + \frac{\gamma}{2}\left[V_{\sigma,+}(1) - V_{\sigma,-}(1)\right] + \rho\left(p + V_{\sigma,-}(0) - V_{\sigma,-}(1)\right). \tag{8}$$

For the types not holding the assets, we have

$$rV_{\sigma,-}(0) = \frac{\gamma}{2}\left[V_{\sigma,+}(0) - V_{\sigma,-}(0)\right] \tag{9}$$

$$rV_{\sigma,+}(0) = \frac{\gamma}{2}\left[V_{\sigma,-}(0) - V_{\sigma,+}(0)\right] + \rho\left(V_{\sigma,+}(1) - V_{\sigma,+}(0) - p\right). \tag{10}$$

It is convenient at this point to consider the change of variables given by

$$s(\rho) \equiv \frac{\rho}{r + \gamma + \rho}. \tag{11}$$

The variable $s \in [0,1]$, which we denote effective speed, is economically more informative than ρ alone, since it reflects the market's ability to reallocate the asset between investors for a given preference shock rate γ.

We can find the ex ante participation value, denoted W. For a given investor of type σ joining a market with effective speed s and a marginal investor given by $\hat{\sigma}$, this function is

$$W(\sigma, \hat{\sigma}, s) \equiv \frac{\overline{a}}{2}\sum_{\epsilon} V_{\sigma,\epsilon}(1) + \frac{1-\overline{a}}{2}\sum V_{\sigma,\epsilon}(0). \tag{12}$$

The interpretation of the right-hand side of Equation 12 is straightforward: The investor is endowed with one unit of the asset with probability $\overline{a}$, and the probability of a temporary type ϵ is one-half. The no-trade outside option of any investor is

$$W_{out} = \frac{\mu\overline{a}}{r}. \tag{13}$$

Naturally, given the symmetric shock structure, W_{out} does not depend on σ.

Given the system of Equations 7-10, we have the following.

Lemma 2. *The function W for a class-σ investor is given by*

$$\underbrace{W(\sigma,\hat{\sigma},s)}_{\textit{Ex ante value}} = \underbrace{\frac{\mu\overline{a}}{r}}_{\textit{Outside option}} + \underbrace{\frac{s\overline{a}\hat{\sigma}}{r}}_{\textit{Ownership Surplus}} + \underbrace{\frac{s}{2r}\max(0,\sigma-\hat{\sigma})}_{\textit{Repeated Trading Surplus}} \tag{14}$$

where s is defined by Equation 11, and the marginal type $\hat{\sigma}(p,\rho)$ is as in Equation 4.

The intuition is that the market participation gains, W net of the outside option W_{out}, is composed of two parts. The (transient) ownership value $\frac{\mu\bar{a}+s\bar{a}\hat{\sigma}}{r}$, independent of σ is the value that can be achieved by all types $\sigma < \hat{\sigma}$ with the "sell and leave" strategy. The second part, $\frac{s}{2r}\max(0;\sigma-\hat{\sigma})$, represents the surplus of trading forever, which depends on the type σ. Importantly, the latter part is super-modular in (s,σ). Hence, the value assigned to any effective speed s increases with the investor type.

C Equilibrium with Costly Participation

We can now analyze investors' participation decisions and formally define an equilibrium with costly participation. Let q_i be the cost of accessing venue i, and let $\mathcal{P}(q,s|\sigma)$ be the participation mapping onto $\{0,1,2,...,I\}$, where $\mathcal{P}=0$ means staying out, $\mathcal{P}=1$ means joining venue 1, and so forth. Staying out costs nothing, so we take $q_0=0$. Recall that G was the ex ante distribution of permanent types. Let $\tilde{G}_i(\sigma)$ be the measure of trader types lower than σ in market i. If all potential investors join market i, as before, we simply have $\tilde{G}_i = G$. In the generic case, however, we have $\tilde{G}_i \leq G$ since some investors may not participate. Indeed, we shall see that in the multiple-venue model the distribution $\tilde{G}$ is typically discontinuous. We then have the following.

Definition 1. An equilibrium is a set of market access fees $q^*=(q_1^*,..,q_I^*)$, effective speeds $s=(s_1,...,s_I)$, participation decisions $\{\mathcal{P}(q,s|\sigma),\sigma\in[0,\overline{\sigma})\}$ by investors, asset prices $p_i\left(q,s,\tilde{G}_i\right)$ and marginal types $\hat{\sigma}_i\left(q,s,\tilde{G}_i\right)$ such that for all $i\leq I$

- $q_i^* \in \arg\max_{q_i} q_i \int_0^{\overline{\sigma}} 1\left\{\mathcal{P}\left(q_i,q_{-i}^*,s\middle|\sigma\right)=i\right\} dG(\sigma)$
- $\mathcal{P}(q,s|\sigma) \in \arg\max_{i\in\{0,...,I\}}\left[W(\sigma,\hat{\sigma}_i,s_i)-q_i\right]$
- $p_i\left(q,s,\tilde{G}_i\right)$ and $\hat{\sigma}_i\left(q,s,\tilde{G}_i\right)$ solve $\hat{\sigma}_i=(rp_i-\mu)\left(1+\frac{\gamma}{r+\rho_i}\right)$ and $\frac{1}{2}\left(\tilde{G}_i(\overline{\sigma})-\tilde{G}_i(\hat{\sigma}_i)\right)=\bar{a}\tilde{G}_i(\overline{\sigma})$, where $\tilde{G}_i(\sigma)\equiv\int_0^{\sigma}1\{\mathcal{P}(q,s|\xi)=i\}dG(\xi)$

This equilibrium concept naturally extends to the case in which the speed s is also endogenous. With only one venue, the investor with marginal trading type $\hat{\sigma}$ must simply be indifferent between joining and not joining the market. So we must have $W(\hat{\sigma},\hat{\sigma},s)-W_{out}=q$ and therefore $q=\frac{s\bar{a}\hat{\sigma}}{r}$. Consequently, all types below $\hat{\sigma}$ are indifferent between joining and staying out since they obtain zero net participation gains. Let δ be the mass of investors that join, sell, and leave. Market clearing requires δ to equal $(1/2\bar{a}-1)(1-G(\hat{\sigma}))$. This condition holds at an interior solution as long as $\delta<G(\hat{\sigma})$, or in other words, as long as $\frac{G(\hat{\sigma})}{1-G(\hat{\sigma})}>\frac{1}{2\bar{a}}-1$. In the remainder of the paper we assume that either $\bar{a}$ is close enough to

1/2 or that there is a sufficient mass of low-type investors to ensure the existence of interior solutions.

To derive analytical results, when convenient we assume the following distribution.

Assumption 1 (A.1).

$$G(\sigma) = 1 - e^{-\frac{\sigma}{\nu}},\ \nu > 0 \tag{15}$$

Note that under A.1 the Walrasian price is simply given by $\frac{1}{r}[\mu - \nu \log(2\overline{a})]$.

Using Equation 11, we can express the equilibrium price 6 as

$$p = \frac{\mu}{r} + \frac{\hat{\sigma}}{r}\left(\frac{r+\gamma s}{r+\gamma}\right). \tag{16}$$

With costly participation we generally have $\hat{\sigma} > G^{-1}(1-2\overline{a})$. It is worth noting that the equilibrium price 16 differs from the benchmark liquidity-adjusted price in the literature (e.g., Duffie, Garleanu, and Pedersen (2005)) in two key aspects:

1. Investors participation decisions that affect $\hat{\sigma}$ are determined endogenously. Naturally, these decisions will be interrelated with the exchange(s) optimization problem in equilibrium.

2. Market contact frictions captured by s will also be endogenously determined.

Consequently, we can explicitly analyze how $(\hat{\sigma}, s)$ are jointly determined as a function of: investor characteristics, the state of technology, the market competitive structure, and regulation. This is the main objective of the following two sections.

D Discussion of Assumptions

The ϵ-shocks can capture time-varying liquidity demands, financing costs, hedging demands, or specific investment opportunities (see Duffie, Garleanu, and Pedersen (2007) for a discussion). The important point is that these shocks affect the private value of the asset, not its common value. For instance, a corporate investor may need to sell financial assets to finance a real investment. A household may do the same for the purchase of a durable good or a house. The parameter σ then simply measures the size of these shocks and thus the volatility of the private value process. One can interpret σ as capturing leverage levels for a given investor. For example, as a group, retail investors generally have lower leverage levels than institutional investors. Thus, retail investors would correspond to lower values of σ in the model. Moreover, note that individual preference types do not drive the endowment

process in any way. Random ownership of a tree share may represent, for example, the outcome of an unmodeled labor market where there is some sort of "stock compensation."

The parameter γ measures the mean reversion of the utility flow process and is assumed for simplicity to be the same for all investors. In the context of delegated management, the shock frequency can represent the sum of the shocks affecting all the investors in a given fund or brokerage house. We introduce heterogeneity in $\sigma$and not in γ because the key point in our analysis is the link between gains from trade and speed. It is important to understand that a higher value of γ implies *lower* gains from trade. Investors with a high value of γ are not eager to trade since they can simply wait for their type to mean-revert. In particular, a high value of γ would not capture the idea of fleeting trading opportunities. This idea is better captured by a high value of σ.

The non-Walrasian feature of the market is captured in a parsimonious fashion by a single market contact rate ρ. Real markets are of course more sophisticated, but for tractability reasons we abstract from modeling the explicit connections between the exchange and, say, a population of potential market makers. One could add bargaining with market makers and bid–ask spreads, but this would not likely bring new insights beyond those of Duffie, Garleanu, and Pedersen (2005) and Lagos and Rocheteau (2009). A market mechanism similar to ours is considered in the monetary economy of Rocheteau and Wright (2005), which they label competitive equilibrium. We also abstract from liquidity (i.e., thick market) externalities. While this abstraction is not without loss of generality and liquidity externalities may still be relevant for some exchange-traded assets, technology may arguably help realize them, even when several trading venues coexist.[26]

Because the focus of exchange differentiation is on trading speed, we interpret q as ex ante costs related to a speed decision that affect multiple trading rounds rather than transaction cost related to a given trade volume. Such costs can include, for example, co-location fees at a particular trading center (see Section I for a discussion). The online appendix of Pagnotta and Philippon (2012) analyzes a model with transaction costs.

III Asset Prices in a Consolidated Market

In this section we derive equilibrium asset prices in an environment with a single profit-maximizing exchange. The single exchange seeks to maximize profit by selecting a given access fee and a given trading speed. Creating or adopting a new trading platform takes

[26]The fact that large cap stocks in the U.S. currently trade in nearly 50 different trading venues suggests that liquidity externalities are not as important or prevalent as they were in the pre-electronic "human trade" era.

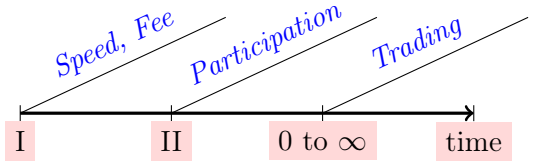

Figure 4: Timing Consolidated Market

After the exchange chooses a trading speed and access fee, investors decide whether to participate. Once these decisions are made, trading starts at time zero.

time, however. We specialize our definition of short and long-run periods in the introduction according to the timing illustrated in Figure 4. After the exchange chooses a trading speed and access fee, investors decide whether to participate. Once these decisions are made, trading starts at time zero. After analyzing the benchmark case, in the remainder of this section we study the case where investors, as opposed to the exchange, have the ability to reduce trading delays.

A Speed Investments and Asset Price

Short Run Exchange Program and Asset Price Consider the case of a market that in the short-run operates under an exogenously given "default" speed $\underline{\rho} > 0$ and seeks to maximize profits by selecting an access fee. In this case the single exchange behaves like a classic monopolist and total profits are given by $\pi = q\left(1 - G\left(\hat{\sigma}\right) + \delta\right)$, which we can write using the market clearing condition 5 as $\pi = \frac{q}{2\bar{a}}\left(1 - G\left(\hat{\sigma}\right)\right)$. Note that if $\bar{a} = 1/2$ we obtain $\delta = 0$, the simplest case to analyze. When $\bar{a}$ is less than $1/2$, we simply need to remember that a mass δ of investors joins, sells, and becomes inactive. The program of the monopolist then has a first-order condition given by

$$\hat{\sigma}_{con} = \frac{\left[1 - G\left(\hat{\sigma}_{con}\right)\right]}{g\left(\hat{\sigma}_{con}\right)}, \tag{17}$$

where the subscript *con* denotes a market with consolidated trading. Importantly, the implied marginal type depends only on the distribution function G. Denoting the price with consolidated trading as p_{con} and using a superscript S to denote the short-run, we obtain

$$p_{con}^{S} = \frac{\mu}{r} + \frac{\hat{\sigma}_{con}^{S}}{r} \times \left(\frac{r + \gamma \underline{s}}{r + \gamma}\right), \tag{18}$$

where as before $\underline{s} = \frac{\underline{\rho}}{r+\gamma+\underline{\rho}}$ and $\hat{\sigma}_{con}^{S}$ is given by Equation 17. Note that under A.1 we have $\hat{\sigma}_{con}^{S} = \nu$.

Long Run Exchange Program and Asset Price In the long-run the exchange can adapt its trading technology s at a cost $C(s)$, affecting the quality of its liquidity service and potentially extracting higher rents from investors. Profit maximization of the exchange requires solving

$$\max_{(q,s)\geq 0} q\left(1 - G\left(\hat{\sigma}\left(q,s\right)\right)\right) - C\left(s\right). \tag{19}$$

Where convenient, we assume the following cost function.[27]

Assumption 2 (A.2). *The cost of the contact rate ρ is given by $c \times \max\left\{\rho - \underline{\rho}; 0\right\}$, where $c > 0$. Equivalently, the cost the cost of the effective speed s is*

$$C\left(s\right) = c \times \max\left\{\left(r+\gamma\right)\frac{s}{1-s} - \underline{\rho}; 0\right\}. \tag{20}$$

In words, under A.2 investment costs are linear in speed improvements $(\rho - \underline{\rho})$ and thus, by Equation 11, a convex function of s.

The analysis of the solution to the program 19 yields the characterization of the equilibrium price p_{con} in the long-run. To gain intuition on the differences between the equilibrium and the benchmark frictionless price p_W, we define the following quantities.

Definition 2. The limited participation distortion (LPD) of the market price in exchange i is given by

$$LPD \equiv \lim_{c\to 0}\left[p_i - p_W\right].$$

Let $\lambda_i \equiv \frac{r+\gamma s_i}{r+\gamma}$. The illiquidity discount (ILD) of price p_i is given by

$$ILD \equiv \hat{\sigma}_i\left(1-\lambda_i\right).$$

The LPD represents the difference between the equilibrium and Walrasian prices in the absence of market contact frictions. That is, it captures the distortion in the value of the marginal investor due to the exchange market power. Naturally, this quantity vanishes when the market access cost approaches zero. The ILD captures the value of the losses that the market marginal investor faces due to her temporary inability to rebalance her portfolio after

[27] Since we are focusing on long-term investment in trading infrastructures, we concentrate on fixed costs of speed enhancements which represent items such as hardware, the development of matching algorithms, integration with clients systems, among others, as opposed to variable costs such as energy or maintenance. The cost of a given technology also does not depend on the number of traders that participate in the exchange. This is, of course, a simplification but, arguably, is does not involve much loss of generality in electronic markets, as opposed to the era of trading floors.

a preference shock. Note that $1-\lambda_i$ can be seen as a measure of market illiquidity, and has a maximum value of zero when the effective speed s_i equals one.[28] Thus, in the remainder of this section we assume A.1 and A.2, as given by Equations 15 and 20, to obtain explicit expressions.

Proposition 2. *The long-run equilibrium price in a consolidated market is given by*

$$p^L_{con} = \underbrace{\frac{1}{r}\left[\mu - \nu \log\left(2\overline{a}\right)\right]}_{\substack{\textit{Walrasian} \\ \textit{Price } (p_W)}} + \underbrace{\frac{\nu}{r}\left[1 + \log\left(2\overline{a}\right)\right]}_{\substack{\textit{Limited Participation} \\ \textit{Distortion}}} - \underbrace{\frac{\nu}{r}\left(\gamma\sqrt{\frac{2rc}{(r+\gamma)}\frac{e}{\nu}}\right)}_{\substack{\textit{Illiquidity} \\ \textit{Discount}}} \tag{21}$$

When market contact frictions are small, the market consolidated price p_{con} is higher than the frictionless price p_W.

The equilibrium long-run price has three components. The first is the Walrasian price p_W, which corresponds to the limit price where market contact frictions vanish and where market participation is costless for investors. The second and third terms correspond to the limited participation and illiquidity adjustments. With a single exchange, the LPD depends on the distribution of investors, as implied by Equation 17, both in the short and long-run, so we have $\hat{\sigma}^L_{con} = \hat{\sigma}^S_{con}$ ($= \nu$ under A.1). The relative value of the distortion also depends on the asset supply and is largest when $\overline{a} = 1/2$. This is because the single exchange chooses $\hat{\sigma}$ irrespective of $\overline{a}$, while $\hat{\sigma}_W$ decreases with asset supply.[29] The size of the ILD depends on the extent of illiquidity, captured by $1-\lambda_{con}$, and its shadow price $\hat{\sigma}_{con} = \nu$ corresponding to the marginal investor type. Note that the ILD approaches zero as market contact frictions become small. This occurs, for example, when c approaches zero. In such cases we have $p_{con} > p_W$, which is a reflection of the levered composition of market participants in the market equilibrium.

B Heterogeneous Frequency Trading and Asset Demand Curves

We characterized in Section III.A the consolidated market equilibrium price in the long-run, where the key investment role is played by the exchange. In modern financial markets, institutional investors also undertake costly investments to ensure fast responses to trading signals and fast communication. For example, a given investment desk may acquire better

[28] Since $s \in (0,1)$, whenever $\gamma \gg r$, λ also lies in the interval $(0,1)$

[29] This holds provided $\overline{a}$ is sufficiently large, that is $\hat{\sigma}_{con} > G^{-1}(1-2\overline{a})$. Otherwise, we must have $\hat{\sigma} = G^{-1}(1-2\overline{a})$

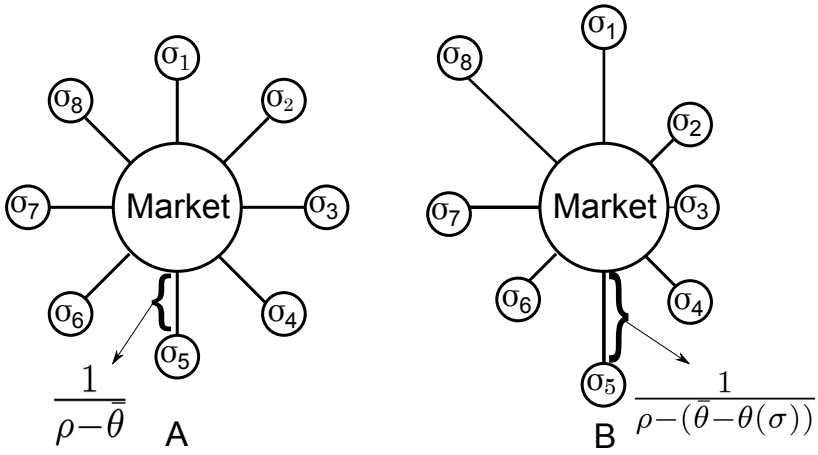

Figure 5: Heterogenous Investor Speeds

Panel A shows that investors that participate in the market are ex ante identical in their "distance" from the market, $\frac{1}{\rho-\theta}$. Panel B shows that endogenizing technology choice leads to σ-dependent heterogeneity in contact rates.

computer equipment, or invest in developing time-efficient trading algorithms. But does the equilibrium asset price depend on who invests in speed? In this section we address this issue by allowing investors to influence their effective rate of market contact before trading starts.

Consider again a market with contact speed ρ, which investors take as given. We now assume that all investors experience an additional time delay, possibly due to equal geographical distance from the market center. In particular, the effective market contact rate for any investor is given by $\rho - \overline{\theta}$, with $\overline{\theta} < \underline{\rho}$. Before joining the market, investors can increase their effective contact rate by investing in a latency reduction technology $\theta \in \left[0, \overline{\theta}\right]$, at a type-independent cost $c_I\theta$ with $c_I > 0$. Whenever investors select a technology θ, their agent-specific effective speed then becomes

$$s\left(\rho, \theta\right) = \frac{\rho - \left(\overline{\theta} - \theta\right)}{r + \gamma + \rho - \left(\overline{\theta} - \theta\right)}. \tag{22}$$

In this environment investors' pre-trade decisions include both a participation and a technology acquisition component:

$$\mathcal{P} : [0, \bar{\sigma}] \longrightarrow \{0, 1\} \times \left[0, \overline{\theta}\right].$$

Figure 5 illustrates that, before trading starts, the effective distance from the market is investor specific. Anticipating investor behavior, the exchange then maximizes profits by

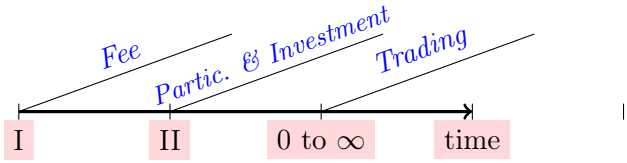

Figure 6: Timing Heterogenenous Frequency Traders

After the exchange chooses an access fee, investors decide whether to participate, and if so, how much to invest in speed. Once these decisions are made, trading starts at time zero.

selecting an optimal market access fee q, as depicted in Figure 6.[30]

In a symmetric equilibrium, all type-σ investors choose the same technology $\theta(\sigma)$, so we can write $s(\sigma) \equiv s(\rho, \theta(\sigma))$. Note that once a particular effective rate $s(\sigma)$ has been determined, trading commences and the equilibrium analysis in Section II remains unchanged. Thus, we can express the expected participation value for a type-σ investor selecting a speed $s(\sigma)$ as

$$W(\sigma, \hat{\sigma}, s(\sigma)) - W_{out} = \frac{\bar{a}}{r} s(\sigma) \hat{\sigma} + \frac{s(\sigma)}{2r} \max(0; \sigma - \hat{\sigma}), \tag{23}$$

which is super-modular in (σ, θ). Because investors will optimally select trading technologies, we can thus express the ex ante expected participation value as

$$\widetilde{W}(\sigma, \hat{\sigma}, s(\sigma)) = \max_{\theta \in [0, \bar{\theta}]} \{W(\sigma, \hat{\sigma}, s(\theta(\sigma))) - C_I(\theta)\}. \tag{24}$$

Consequently, the marginal investor $\hat{\sigma}$ satisfies $\widetilde{W}(\hat{\sigma}, \hat{\sigma}, s(\hat{\sigma})) - q = W_{out}$. Finally, note that the market clearing condition is not affected due to the independence of preference shocks across investor types.

Let p_{hft} denote the equilibrium price in the market with "heterogeneous frequency traders." By jointly solving the investor speed selection problem and the exchange optimal fee problem, we can show the following.

Proposition 3. *When investors select trading technologies, the equilibrium price is*

$$p_{hft} = \underbrace{\frac{1}{r}[\mu - \nu \log(2\bar{a})]}_{\substack{\textit{Walrasian} \\ \textit{Price } (p_W)}} + \underbrace{\frac{\hat{\sigma}_{hft}}{r}[1 + \nu \log(2\bar{a})]}_{\substack{\textit{Limited Participation} \\ \textit{Distortion}}} - \underbrace{\frac{\hat{\sigma}_{hft}}{r}(1 - \lambda_{hft})}_{\substack{\textit{Illiquidity} \\ \textit{Discount}}}, \tag{25}$$

where $\hat{\sigma}_{hft}$ and λ_{hft} are expressions given in the Online Appendix. When speed frictions are

[30] Although one could analyze joint speed choices, to keep the analysis from being overly complicated, we take the market speed as fixed and concentrate on investor choice in this section.

small, investments in speed increase with the asset supply and asset demand curves can slope upward.

According to Proposition 3, investors' choices affect both the LPD and ILD relative to p_{con} in Equation 21. In this case, the marginal investor type does not depend solely on the investor type distribution but also on parameters such as the cost of latency reductions. Consequently, we generally have $\hat{\sigma}_{hft} \neq \hat{\sigma}_{con} = \nu$. The ILD is naturally affected by the change in the value of the marginal investor type, and by the marginal investor's distance to the market (embedded in λ_{hft}). When market contact frictions are relatively small, the value of the ILD will be generally lower than in the previous section due to the fact that, for our calibrated parameters, we have $\hat{\sigma}_{hft} < \hat{\sigma}_{con}$.

The equilibrium with heterogeneous frequency trading also has interesting implications for the shape of demand curves. Consider an increase in the asset supply parameter $\overline{a}$. This change has a direct negative impact on the marginal investor type which tends to lower the asset price 25. However, such an increase has a positive effect on the optimal speed choice s_{hft}, which in turn raises $\hat{\sigma}_{hft}$ indirectly. The intuition behind this effect is clear when comparing the investor ex ante values given by Equations 14 and 23. In the latter, the value of ownership increases with the investor type σ through the interaction between $\overline{a}$ and $s(\sigma)$. In an economy where the asset supply incentive effect on s_{hft} is strong, one can observe, at least at low frequencies, that demand curves slope upward. An implication of this latter effect is that, in an environment with multiple assets, we would expect to see traders that heavily invest in speed concentrate in large assets (e.g., large market capitalization stocks), potentially widening the relative illiquidity of small assets.

IV Asset Prices in Fragmented Markets

In this section we analyze competition between a given set of trading venues, the allocation of investors across these venues, and the resulting asset prices.

Consider two venues, 1 and 2, with speeds ρ_1 and ρ_2 (i.e effective speeds $s_i = \frac{\rho_i}{r+\gamma+\rho_i}$) and participation fees q_1 and q_2, respectively. If both speeds were equal, the exchanges would have to compete à la Bertrand in fees, leaving each venue with zero profits. Exchanges therefore have an incentive to differentiate their intermediation services by offering different speeds and thus relaxing competition in fees. We concentrate on this case and, without loss of generality, take venue 2 as the fast market, so that $\rho_1 < \rho_2$. Since fees can be adjusted more easily than trading platforms, it is natural to model exchange competition as

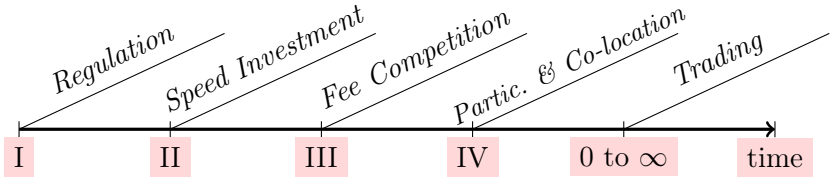

Figure 7: Timing Fragmented Markets

For a given trading regulation, the fast exchange chooses a trading speed and then exchanges set access fees. Investors decide whether to participate and, if so, which market to join. Once these decisions are made, trading starts at time zero.

a sequential game where markets first select a trading technology and subsequently compete on fees. In the remainder of this paper we focus solely on exchanges' speed choices.

With more than one trading venue, investors can find different prices to buy or sell an asset in each of them. Policymakers are often concerned about "protecting" investors from bad execution and have designed a number of rules that regulate how market prices relate to each other (see the discussion in Appendix A). We consider two types of trading regimes, which capture stylized extreme cases.

Definition 3. Trading regimes $\mathsf{T} \in \{seg, prot\}$ are as follows.

- **Market segmentation** (*seg*): When a venue refuses to execute trades coming from investors of another venue. There are then two asset markets and two liquidity markets.
- **Price protection** (*prot*): The market authority enforces a single asset price. There are then one asset market and two liquidity markets.

Under segmentation, an investor joins a market and cannot trade with an investor in another market. Once investors have been allocated, the markets are effectively segmented and equilibrium asset prices can be different. Under price protection, asset prices must be the same in both venues.[31] Note that a single price would also prevail if there were limitless arbitrage opportunities for a mass of potential arbitrageurs. In this regard, price protection is equivalent to perfect arbitrage. A more general framework with costly arbitrage (e.g., Shleifer and Vishny (1997)) would deliver equilibrium asset prices that lie somewhere between the low and high segmented prices we derive.

For a given trading regime, a market structure is characterized by a set of speed and fees (s, q). Investors observe these outcomes and decide whether to participate in financial markets, and which trading venue is more attractive. The pre-trade decisions of type-σ investors

[31] This is our simple way to capture access and trade-through rules in the SEC's Reg NMS. The distinction between the U.S.'s top-of-the-book and Canada's full-depth protection becomes trivial here since we only consider unitary orders. See Appendix A for a discussion of investor protection and more details.

are formally described as

$$\mathcal{P} : (s, q) \times [0, \bar{\sigma}] \longrightarrow \{0, 1, 2\}.$$

The timing of decisions is illustrated in Figure 7. We will see that investors choosing to trade in the fast market do so by paying a higher access fee. In light of the discussion in Section I, investors decision to pay a speed premium can be interpreted as a decision to colocate with an exchange's servers.

A Segmented Markets

Let us start studying investors' pre-trade decisions given (s, q). When markets are effectively segmented, we need to consider three indifference conditions. First, there is a type $\hat{\sigma}_1$ who is indifferent between joining market 1 and staying out. This type must satisfy

$$W(\hat{\sigma}_1, \hat{\sigma}_1, s_1) - W_{out} = q_1,$$

which, using Equation 14, implies

$$q_1 = \frac{\bar{a} s_1 \hat{\sigma}_1}{r}. \tag{26}$$

The second indifference condition defines the marginal type $\hat{\sigma}_{12}$, who is indifferent between joining market 1 and market 2. By definition, this type must be such that

$$W(\hat{\sigma}_{12}, \hat{\sigma}_2, s_2) - q_2 = W(\hat{\sigma}_{12}, \hat{\sigma}_1, s_1) - q_1. \tag{27}$$

The third indifference condition is that the temporary investors are indifferent between joining markets 1 and 2. Therefore we must have

$$W(\hat{\sigma}_2, \hat{\sigma}_2, s_2) - W_{out} - q_2 = W(\hat{\sigma}_1, \hat{\sigma}_1, s_1) - W_{out} - q_1$$

Given the indifference condition for $\hat{\sigma}_1$ this implies $W(\hat{\sigma}_2, \hat{\sigma}_2, s_2) - W_{out} - q_2 = 0$. Combining these conditions, we obtain $\frac{s_1}{2r}(\hat{\sigma}_{12} - \hat{\sigma}_1) = \frac{s_2}{2r}(\hat{\sigma}_{12} - \hat{\sigma}_2)$ and $q_2 = \frac{\bar{a} s_2 \hat{\sigma}_2}{r}$, and therefore

$$\hat{\sigma}_{12} = \frac{r}{\bar{a}} \frac{q_2 - q_1}{s_2 - s_1}. \tag{28}$$

Note that $\hat{\sigma}_1 < \hat{\sigma}_2 < \hat{\sigma}_{12}$. The set of types that join market 2 cannot be an interval. It is composed of all the types above $\hat{\sigma}_{12}$ and some types below $\hat{\sigma}_1$.

Let us consider exchange optimization. The total revenue for slow and fast exchanges is given by $q_1 (G(\hat{\sigma}_{12}) - G(\hat{\sigma}_1) + \delta_1)$ and $q_2 (1 - G(\hat{\sigma}_{12}) + \delta_2)$, where, as in Section II, δ_i represents the mass of temporary investors joining market i. The mass of temporary traders in each

market is determined by market clearing. Due to market segmentation, we need to consider two different market clearing conditions:

$$\left(1-G\left(\hat{\sigma}_{12}\right)+\delta_2\right)\bar{a}=\frac{1-G\left(\hat{\sigma}_{12}\right)}{2}$$

for market 2 and

$$\left(G\left(\hat{\sigma}_{12}\right)-G\left(\hat{\sigma}_1\right)+\delta_1\right)\bar{a}=\frac{G\left(\hat{\sigma}_{12}\right)-G\left(\hat{\sigma}_1\right)}{2}$$

for market 1. Using these conditions, we can express the slow and fast exchange gross profit functions as $\pi_2^{seg}=q_2\frac{1-G(\hat{\sigma}_{12})}{2\bar{a}}$ and $\pi_1^{seg}=q_1\frac{G(\hat{\sigma}_{12})-G(\hat{\sigma}_1)}{2\bar{a}}$.

To simplify the analysis we take market 1's speed as exogenously given by the economy's default speed, that is $s_1=\underline{s}$, while market 2 chooses an effective speed s_2 at a cost $C\left(s_2\right)$. In the speed choice stage, market 2 thus solves

$$\max_{s_2}\frac{q_2\left(s_2\right)}{2\overline{a}}\left[1-G\left(\hat{\sigma}_{12}\left(s_2\right)\right)\right]-C\left(s_2\right) \tag{29}$$

We can naturally extend the equilibrium in Definition 1 to this environment by adding an investment stage. To ensure sub-game perfection we must first find the access fee q_i^* that maximizes π_i^{seg} given $q_{j\neq i}$, for $i,j=1,2$; and where the marginal types are given by Equations 26 and 28. Second, given q_1^* and q_2^*, we must find the speed level s_2^* that solves program 29. We shall see in the next section that the equilibrium solution to this problem implies a level of participation and speed in the fast market that exceeds the one in the consolidated market.

Once these objects are determined, we can characterize the equilibrium prices.

Proposition 4. *In the segmented markets equilibrium, asset prices are given by*

$$p_1=\underbrace{\frac{1}{r}\left[\mu-G^{-1}\left(1-2\overline{a}\right)\right]}_{\substack{Walrasian\\Price}}+\underbrace{\frac{1}{r}\left[\hat{\sigma}_1+G^{-1}\left(1-2\overline{a}\right)\right]}_{\substack{Limited\ Participation\\Distortion\ market\ 1}}-\underbrace{\frac{1}{r}\hat{\sigma}_1(1-\lambda_1)}_{\substack{Illiquidity\\Discount\ market\ 1}}\quad, \tag{30}$$

and

$$p_2=\underbrace{\frac{1}{r}\left[\mu-G^{-1}\left(1-2\overline{a}\right)\right]}_{\substack{Walrasian\\Price}}+\underbrace{\frac{1}{r}\left[\hat{\sigma}_2+G^{-1}\left(1-2\overline{a}\right)\right]}_{\substack{Limited\ Participation\\Distortion\ market\ 2}}-\underbrace{\frac{1}{r}\hat{\sigma}_2(1-\lambda_2(s_2))}_{\substack{Illiquidity\\Discount\ market\ 2}}\quad, \tag{31}$$

where marginal types $\hat{\sigma}_1$ and $\hat{\sigma}_2$ satisfy Equations 59-61 given in the Appendix. The limited participation distortion (LPD) is higher in the fast market. The ILD in the fast market can be lower or higher than in the slow market.

As in a consolidated market, we can see the prices in the markets of Proposition 4 as the sum of three terms: the Walrasian price, an LPD, and an ILD. The LPD here reflects the change in the value of the marginal investor relative to the Bertrand outcome. Intuitively, when speed differentiation increases, markets are able to increasingly relax competition in access fees. Given that investors with high types choose to trade in the fast market, this distortion is naturally greater in market 2. However, it is interesting to note that when markets are segmented the ILD is not necessarily lower in the faster market. Although the amount of illiquidity is lower in market 2 (i.e. $1-\lambda_2 < 1-\lambda_1$), its marginal valuation in that market is higher. In fact, when market contact frictions are small, the ILD is likely to be higher in the fast market.

B Protected Markets

In this section we analyze how price protection in the trading period affects competition between exchanges and equilibrium asset prices.

With price protection asset markets operate under a single market clearing condition. Investors then self-select into trading venues that can be seen as different entry points into a single integrated asset market. Following widely used terminology in U.S. stock markets, we refer to the single clearing price as the national best price, denoted p_{nb}. Active participants in markets 1 and 2 are still characterized by the indifference condition 26 for the marginal type $\hat{\sigma}_1$, and the marginal type $\hat{\sigma}_{12}$ is still characterized by Equation 27. However, with a single asset price temporary traders are not indifferent between joining market 1 and joining market 2. Intuitively, temporary traders now have a stronger incentive to join market 1, where they pay a lower access fee, since they can sell their endowment at the same price in either market. We then have $\delta_2 = 0$ and the market clearing condition becomes

$$\underbrace{\left(1-G\left(\hat{\sigma}_1\right)+\delta_1\right)\bar{a}}_{\text{Total Asset Supply}} = \underbrace{\frac{\left(1-G\left(\hat{\sigma}_1\right)\right)}{2}}_{\text{Stationary Asset Demand}} \tag{32}$$

The redistribution of investors across venues impacts market revenue. In particular, the Online Appendix shows that the gross revenue functions π_1^{prot} and π_2^{prot} are now $\frac{q_1}{2\bar{a}}[1-G(\hat{\sigma}_1)+2\bar{a}(G(\hat{\sigma}_{12})-1)]$ and $q_2\left(1-G\left(\hat{\sigma}_{12}\right)\right)$, respectively. By solving the new two-stage competition game, and imposing market clearing condition 32, we can characterize the national best price.

Proposition 5. *The single equilibrium price under price protection is*

$$p_{nb} = \underbrace{\frac{1}{r}\left[\mu - G^{-1}\left(1-2\overline{a}\right)\right]}_{\substack{\textit{Walrasian}\\ \textit{Price}}} + \underbrace{\frac{1}{r}\left[\sigma_1^{seg} + G^{-1}\left(1-2\overline{a}\right)\right]}_{\substack{\textit{Limited Participation}\\ \textit{Distortion (reg. free)}}} - \underbrace{\frac{1}{r}\sigma_1^{seg}(1-\lambda_1)}_{\substack{\textit{Illiquidity}\\ \textit{Discount (reg. free)}}} + \underbrace{\frac{1}{r}(\sigma_1^{prot} - \sigma_1^{seg})\lambda_1}_{\substack{\textit{Price Protection}\\ \textit{Distortion (PPD)}}} . \tag{33}$$

Under A.1 (Equation 15) the price protection distortion (PPD) is positive.

The first three terms on the right-hand side of Equation 33 in Proposition 5 coincide with the regulation-free price in the slow market, as given by Equation 30. There is an additional term in this case that captures the price distortion of the trading regulation in market 1:

$$PPD \equiv p_{nb} - p_1^{seg}.$$

Under general conditions, such a distortion is positive, and the national best price p_{nb} is greater than p_1^{seg}. The intuition is as follows. Price protection acts as a subsidy to the slow market because its investors are (effectively) allowed to sell their assets to investors in the fast market. This creates, everything else being constant, a larger demand for the slow market, which encourages the slow market to increase its access fee q_1, ultimately raising the value of the marginal type $\hat{\sigma}_1$. This is why we have $\hat{\sigma}_1^{prot} > \hat{\sigma}_1^{seg}$. Protection also softens the price elasticity of the marginal type $\hat{\sigma}_{12}$, which again is good for the slow venue. Thus the slow venue's profits increase under protection for two reasons: more demand and less price elasticity.

Note that Proposition 5 does not characterize the impact of price protection on p_2^{seg}. We analyze this issue further in the following section.

V Empirical Implications

In this section we study the relation between the different asset prices obtained in Sections III and IV, as well as implications for trading volume and fragmentation levels.

Our first task is to compare price outcomes in a consolidated market and in unregulated fragmented markets. To do so, we would like to have a notion of the average price in the segmented markets. A natural choice is to compute the volume-weighted average price (VWAP). Let τ_i represent the steady-state expected number of trades in market i per unit of time, a natural notion of volume. We then define the VWAP as follows.

Definition 4. The VWAP p_{vw} is given by

$$p_{vw} \equiv \left(\frac{\tau_1}{\tau_1+\tau_2}\right) p_1 + \left(\frac{\tau_2}{\tau_1+\tau_2}\right) p_2. \tag{34}$$

Note that we can easily derive the expressions for τ_1 and τ_2. First, the masses of active traders in markets 1 and 2 are given by $G(\hat{\sigma}_{12}) - G(\hat{\sigma}_1)$ and $1 - G(\hat{\sigma}_{12})$, respectively. From Proposition 1, we know that in the steady state the proportion of asset owners wanting to sell their asset and the proportion of agents wanting to buy it is given by $\frac{1}{4}\frac{\gamma}{\gamma+\rho}$. With the contact rates for markets 1 and 2 equal to ρ_1 and ρ_2, respectively, the average number of trades in each market at any moment is then

$$\tau_1 = \rho_1 \frac{\gamma}{4(\gamma+\rho_1)} \left[G(\hat{\sigma}_{12}) - G(\hat{\sigma}_1)\right] \overline{a} \tag{35}$$

$$\tau_2 = \rho_2 \frac{\gamma}{4(\gamma+\rho_2)} \left[1 - G(\hat{\sigma}_{12})\right] \overline{a}. \tag{36}$$

A The Pricing Effect of Trading Fragmentation

Does exchange competition affect asset prices in the long-run? The relationship between p_{con} and p_{vw} is not obvious because there are two opposite effects in force. On the one hand, the monopoly restricts entry and induces a high marginal type $\hat{\sigma}_{con}$, raising the LPD and thereby increasing the equilibrium price. On the other hand, speed increases prices and competition can increase the market average speed. The following result characterizes the relationship.

Proposition 6. (Competition and Asset Prices)

SR. *The limited participation distortion (LPD) is always higher in the consolidated market. Thus, for any given market speed, the consolidated price is higher than the VWAP in the short-run.*

LR. *In the long-run, the VWAP can be higher or lower than the consolidated price. Provided that market contact frictions are moderate, the consolidated price is higher. Under A.1 (Eq. 15), the illiquidity discount (ILD) of the fast market price is always lower than in the consolidated market.*

If both the monopolist exchange and the fast market have access to the same speed in the short-run, according to Equations 21, 30, and 31, price differences between p_{con} and p_{vw} will be due to the LPD only. Since the monopolist exchange restricts participation to a greater extent, the price in the consolidated market will be higher, regardless of the volume

distribution across venues 1 and 2. Figure 8 illustrates this fact by comparing the marginal investor $\hat{\sigma}_{con}$ and a volume-weighted marginal type $\hat{\sigma}_{vw}$. In the long-run, markets can invest in speed technologies and the sign of $p_{con} - p_{vw}$ thus depends on the relative strengths of the effects on the LPD and ILD. We have seen in Proposition 2 that the LPD remains unchanged in a consolidated market but the ILD decreases. The LPD in fragmented markets will increase in the long-run because of speed differentiation allowing markets to relax fee competition, but it remains lower than in the consolidated market. The extent of frictions in the fragmented market relative to the consolidated case is likely to decrease and thus reduce the ILD. In particular, under A.1, we have $s_2^{seg} > s_{con}$ in the long-run and therefore the ILD is lower in the fast market.[32] Consequently, the sign of the total effect of competition on asset prices depends on parameter values. However, when market contact frictions are small (e.g., stocks, exchange-traded derivatives), the LPD effect dominates, implying $p_{con} > p_{vw}$.

B The Pricing Effect of Price Protection

Consider an economy where asset markets are fragmented but unregulated. How would adoption of a trade-through-like rule affect prices? A related question is what would the effect on prices be like when the transition is from a consolidated to a regulated fragmented market? Section IV shows that the protected price is higher than the unregulated price in the slow market. We now want to compare the protected market with the consolidated market and VWAP prices. The following result characterizes these relations.

Proposition 7. (Investor Protection and Asset Prices)

The consolidated market price is always higher than the national best price. Under A.1 (Eq. 15), the protected price p_{nb} lies between the prices of the segmented venues, and is lower than the VWAP, provided the speed cost parameters c and $\underline{s}$ are sufficiently low.

A single exchange will distort participation to a greater extent than in a duopoly, increasing the LPD with respect to the protected market regardless of the time period. This fact is illustrated by the left-pointing arrows in Figure 8: The value of the marginal investor in the consolidated market lies to right of the ones in fragmented markets. Moreover, the speed investments in the long-run will reduce the amount of illiquidity in the consolidated market, increasing p_{con} relative to p_{nb}. Speed investments indirectly affect p_{nb} in the long-run, by increasing the value of the marginal investors as the exchanges differentiate themselves from each other and relax fee competition. Note that the marginal investor in the duopolistic economy resides in the slow market and thus its valuation of the asset depends on the default

[32] It is easy to show that $\hat{\sigma}_{con} > \hat{\sigma}_2^{seg}$ does not depend on A.1.

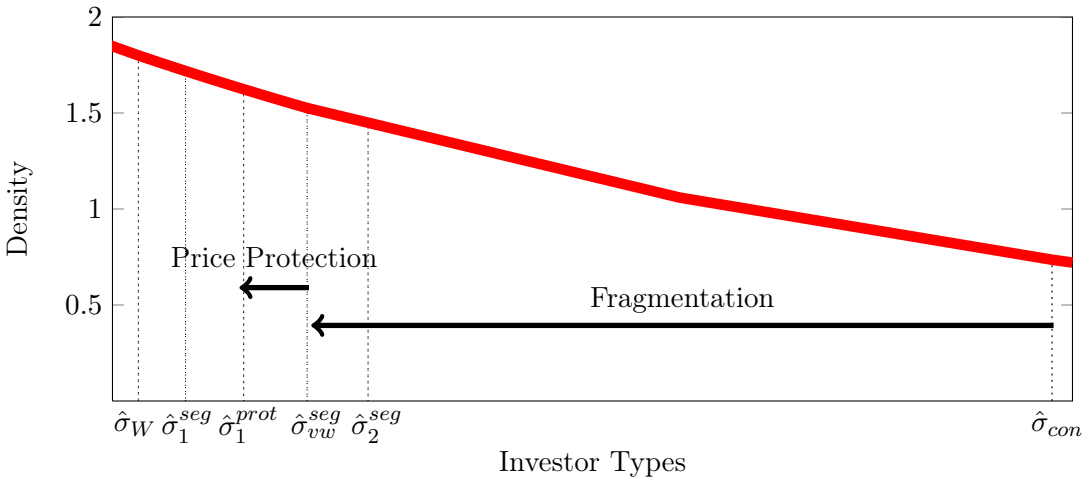

Figure 8: Distortions in Marginal Investor Types

This figure shows the location of marginal investor types in the different cases of the analysis. The parameter values correspond to the baseline calibration (Table II). From left to right, the marginal types correspond to a frictionless market ($\hat{\sigma}_W$), a slow segmented market ($\hat{\sigma}_1^{seg}$), a slow protected market ($\hat{\sigma}_1^{prot}$), a fast segmented market ($\hat{\sigma}_2^{seg}$), and a consolidated market ($\hat{\sigma}_{con}$). The type $\hat{\sigma}_{vw}^{seg}$ represents the volume-weighted average of the marginal investor types in each of the freely segmented markets.

technology. Consequently, there is no direct effect of investments on the ILD in the protected market, and thus the ILD is lower in the consolidated market, regardless of the time period.

Whether the VWAP is higher than the protected price depends on the relative strength of two effects. On the one hand, we know from Proposition 5 that price protection introduces a positive distortion in the asset price of market 1, implying $p_1^{seg} < p_{nb}$. On the other hand, both the LPD and the level of asset liquidity are higher in the fast market than in the protected market, implying $p_2^{seg} > p_{nb}$. The volume distribution then becomes important in comparing p_{nb} and p_{vw}. When the ability of the fast market to differentiate its liquidity service is substantial, this market can attract greater trading flows and thus $p_{vw} > p_{nb}$. This occurs, for example, when technology costs and the default speed level are relatively low. Figure 8 shows that this is the case in our baseline calibration.

C Calibration and Analysis of Price Distortions

In this section we consider a calibration of our model to illustrate the impact of long-run investments and market organization on prices. Assuming A.1–A.2 (Equations 15 and 20), our model involves the nine exogenous parameters listed in Table II. Some of these parameters, we argue, can be calibrated using secondary markets data. In cases where parameters are more difficult to calibrate, we explain our choices and their impact on the

calibration results.

The illiquid asset in the model can represent one of the several asset classes that trade in exchanges, such as stocks, futures, or equity options. In order to be specific, and motivated by the empirical observations in Section I, we concentrate on U.S. equity markets during 2001-2007. This is an important time interval since it corresponds to the period in between the SEC's decimalization mandate and the final implementation of Reg NMS. Further, 2007 is the last year for which accurate data from the SEC Rule 605 is available, since in later years speeds are rounded down to zero. Stock prices obviously reflect market risk exposure, which is not the focus of this paper. Hence, when interpreting the results it is important to take into account that our main goal is instead quantifying the relative contribution of each of the frictions we identify to distortions to the Walrasian price. For simplicity, we also calibrate our parameters taking the NYSE as being the single trading venue. Although other markets, chiefly the NASDAQ, may have competed with the NYSE during the period, this approach allows us to calibrate parameters applying the simple formulas of Section III to market data without much loss of generality.[33]

The asset characteristics are as follows. The interest rate is set equal to 2.5% annually, just below the one-month T-bill rate average value for the period 2001–2007. The annual holding cash flow μ is set to 2.4 units of the consumption good, which implies a dividend yield close to that of S&P500 stocks for this period (relative to the baseline Walrasian price). The asset supply $\overline{a}$ is normalized to 0.47 so that supply is short. This is almost without loss of generality since the market price does not depend on $\overline{a}$ when this parameter is sufficiently close to one half.[34]

To calibrate the speed contact parameters, we consider SEC Rule 605 data for the NYSE for the years 2001 and 2007. We interpret the 2001 value as corresponding to our "slow" (default) speed $\underline{\rho}$, which is relatively intensive in human trading, and the 2007 value as our "fast" (long-run) speed. The default speed then corresponds to a daily Poisson rate of $1,170$, which translates to an average execution speed of approximately 20 seconds. To match an average execution speed of one second in 2007 for the NYSE, we consider a daily contact rate equal to 23,400, the number of seconds in a trading day. In order to annualize these values, as shown in Table II, we multiply daily rates by 252 trading days.

To calibrate the preference switching rate γ we proceed as follows. We normalize the steady state fraction of agents with misallocated assets, equal to $\frac{1}{4}\frac{\gamma}{\gamma+\rho}$ as given in Proposition 1, to be 5%. Using the value of $\underline{\rho}$, we compute the implicit value for γ, which yields a daily rate

[33] According to the Securities and Exchange Commission (2010), the NYSE executed as much as four-fifths of the volume of NYSE-listed stocks in 2005.

[34] From Section II we know that the price is insensitive to asset supply when the marginal investor type in market i, $\hat{\sigma}_i$, is higher than $G^{-1}(1-2\overline{a})$.

equal to 292.5. We interpret this value as representing the trading needs of a given broker, as opposed to a single individual investor, that represents a large number of customers with the same σ-type value. Such a broker, depending on its customers order flow, may find it optimal to switch holdings between 0 and 1 positions multiple times during a trading day.[35]

Since we assume an exponential distribution of investor permanent types (assumption A.1), we only need to calibrate the value of the average investor type ν. This parameter, however, is not easy to compute based on market data. Since our calibration results are sensitive to the choice of ν, we consider a range of different values and analyze the corresponding pricing responses. In particular we take ν to lie in the set $\left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right\}$, and consider $\nu = \frac{1}{2}$ as the baseline value. To illustrate the economic interpretation of these values, consider the median investor type, $m(\sigma)$, when $\nu = \frac{1}{4}$. The annual holding flow utility under a temporary shock ϵ (see equation 1) is $u_{m(\sigma),\epsilon}(1) = 2.44 + \frac{\epsilon}{4}\ln(2)$. Consequently, the private component of the annual utility flow for this investor is approximately 17.5 bps (relative to p_W). This implies that, when facing a negative or positive temporary shock, the annual flow utility equals 2.27 and 2.61 units of consumption, respectively. Similarly, for the baseline value $v = \frac{1}{2}$, the annual utility flow lies in $\{2.14; 2.78\}$. Given the lack of direct evidence on the cumulative distribution function G it is difficult to ascertain whether $\nu = \frac{1}{2}$ is the most plausible value, so the success of our calibration analysis should be qualified.

The marginal cost of speed investment c is chosen so that in the long-run, and given the other parameter values, the optimal single-market contact speed implies an average execution speed of approximately one second.[36] This value is close to the average trading speed for the NYSE in 2007 according to the SEC Rule 605 data.

[35] To illustrate this interpretation, consider a discretization of the steady-state trade volume formula 35: $\tau = \frac{1}{4}\frac{\gamma}{\gamma+\rho} \times \rho \times$ Number of brokers. According to data available at www.nyse.nyx.com, the average number of daily trades for all 3,025 NYSE-listed stocks in 2001 was approximately equal to 452. Taking $\tau = 452$, and considering our calibrated parameters, this formula implies that there were, on average, 8 active trading brokers for the average NYSE-listed stock on any given day.

[36] Inverting equation 47 in the Appendix, one finds that

$$c = \frac{\nu(r+\gamma)}{2er(r+\gamma+\rho_{con})^2}.$$

The value of c is then recovered considering $\rho_{con} = 23,400 \times 252$.

Table II: Parameter values in Calibration

Parameter	Notation	Baseline Value*
Interest rate	r	2.5%
Holding cash flow	μ	2.44
Default contact rate	$\underline{\rho}$	2.95×10^5
Short-run contact rate market 2	$\underline{\rho}_2$	1.18×10^6
Long-run contact rate consolidated market	ρ_{con}	5.90×10^6
Switching intensity temporary types	γ	73,710
Marginal cost of speed investments	c	7.6×10^{-9}
Asset supply	$\bar{a}$	0.47
Average investor type (baseline value)	ν	0.5

*The values of parameters $\{r, \mu, \underline{\rho}, \underline{\rho}_2, \rho_{con}, \gamma\}$ correspond to annual rates.

We start considering a short-run equilibrium where fees and market participation are endogenously determined, but trading platforms' speed are a given. To analyze the case of fragmented markets, we need to include in the parameter space a trading speed for the fast market, denoted $\underline{\rho}_2$. We consider an exchange that has an average execution speed of five seconds, that is, $\underline{\rho}_2 = 4\underline{\rho}$. This value is close to that reported by Angel, Harris, and Spatt (2011) for the NASDAQ in late 2001, and can be seen as technologically feasible for other markets during the early 2000s. It results, of course, in a slower average execution speed than the one reported for the NYSE around 2007 (around one second).

Table III reports the model-implied value of the price decompositions derived in Sections III and IV. A number of interesting observations are in order. First, we observe that consistent with Propositions 6 and 7, for an asset class such as stocks where market contact frictions are small, we have $p_{con} > p_{vw}$ and $p_{vw} > p_{nb}$. Note that in both inequalities, the LPD plays a key role. In fact, without this quantity being endogenized, all the computed prices should be below the frictionless Walrasian value (e.g., Duffie, Garleanu, and Pedersen (2005)), while only p_1 and p_{nb} are below p_W. For example, considering the baseline $\nu = 0.5$, the absolute value of the ILD in the slow market (33 bps) is higher than the LPD (5 bps). It is also interesting to note that an observer comparing the national best and Walrasian price could conclude that the market is essentially frictionless, while in fact the small aggregate distortion is due to the sum of the values of the three individual distortions we identified. Although the absolute value of the PPD is fairly small in our parametrization, note that it represents nearly half the size of the ILD, and about 42% of the LPD.

Let us now consider the long-run equilibrium of asset prices under the baseline parametrization, depicted in Table IV. Consistent with Proposition 2, the value of the LPD in a consolidated market (driven by the ex ante distribution of types G) is unchanged. However,

Table III: Short-Run Price Decomposition (Walrasian Price=100)

	Limited Participation Distortion	Illiquidity Discount	Price Protection Distortion	Price
		$\nu = 0.25$		
Consolidated	9.55	-2.04		107.52
Slow Venue	0.19	-0.16		100.03
Fast Venue	0.90	-0.09		100.81
VWAP	0.66	-0.14		100.55
National Best	0.19	-0.16	0.08	100.11
		$\nu = 0.5$		
Consolidated	18.98	-4.05		114.94
Slow Venue	0.38	-0.33		100.06
Fast Venue	1.78	-0.18		101.60
VWAP	1.32	-0.28		101.09
National Best	0.38	-0.33	0.16	100.21
		$\nu = 0.75$		
Consolidated	28.30	-6.03		122.26
Slow Venue	0.57	-0.49		100.08
Fast Venue	2.66	-0.27		102.39
VWAP	1.96	-0.41		101.62
National Best	0.57	-0.49	0.24	100.32
		$\nu = 1$		
Consolidated	37.50	-7.99		129.50
Slow Venue	0.76	-0.65		100.11
Fast Venue	3.52	-0.35		103.17
VWAP	2.60	-0.55		102.15
National Best	0.76	-0.65	0.31	100.42

This table presents the values of the price distortions and of the equilibrium asset price relative to the frictionless Walrasian price, which is normalized to 100, in the short-run, for four values of the average investor type ν. Parameter values (except ν) correspond to the baseline calibration (Table II). In the short-run, investor participation is variable, but trading technologies are fixed. The Walrasian price is the price in the absence of market contact and investor participation frictions. The VWAP corresponds to that between the slow and fast venues. The national best price correspond to the unique price in fragmented markets under price protection. The LPD and ILD are as in Definition 2.

the ILD decreases significantly: It decreases to 25 bps from 4.05%. This change reflects the effect of speed investments on allocative efficiency. In the case of fragmented trading, note that the LPD increases in all cases, which reflects the enhanced ability of exchanges to increase their access fees due to the relaxation of Bertrand competition through speed differentiation (the average execution speed is now 20 times faster in the fast exchange). This increase reflects the interaction between the time-varying competitive environment due to technological progress and trading regulations.

Finally, note that both in the short and long-run the LPD terms increase monotonically in the value of the average investor type ν. While the effect is large for the consolidated market, since the marginal investor type equals ν, in the other market structures analyzed the effect of changes in ν are quantitatively small.

D Comparative Statics of Asset Prices

Consolidated Market How is the equilibrium price in a consolidated market affected by changes in the environment? Are the relations affected in the long-run? By differentiating Equations 18 and 21, we can characterize the behavior of the market clearing asset price as follows.

Proposition 8. *In a consolidated market,*

(i) The equilibrium asset price increases in the volatility of investors' private utility process. The effect is stronger in the long-run.

(ii) The equilibrium asset price decreases in the frequency of preference shocks. The effect is weaker in the long-run, provided the cost of speed is not "too high."

(iii) The equilibrium asset price decreases in the cost of speed.

Figure 9 graphically displays these relationships. An increase in the average investor type (ν) increases the value of the marginal investor type. Everything else being constant, this increases the LPD in the short-run. In the long-run, an increase in ν also makes speed investments more profitable for the exchange, resulting in a lower-ILD equilibrium and further raising the asset price.

In the short-run, the equilibrium asset price decreases with the frequency of the temporary shocks γ. This effect is intuitive: For a given installed speed capacity, an increase in γ increases the proportion of agents with misallocated assets at any given time, rendering the asset less valuable. In the long-run, speed investments facilitate more efficient asset

Table IV: Long Run Price Decomposition (Walrasian Price=100)

	Limited Participation Distortion	Illiquidity Discount	Price Protection Distortion	Price
		$\nu = 0.25$		
Consolidated	9.55	-0.18		109.37
Slow Venue	0.39	-0.20		100.19
Fast Venue	1.28	-0.03		101.26
VWAP	0.99	-0.15		100.91
National Best	0.39	-0.20	0.10	100.28
		$\nu = 0.5$		
Consolidated	18.98	-0.25		118.73
Slow Venue	0.82	-0.41		100.40
Fast Venue	2.62	-0.04		102.58
VWAP	2.04	-0.29		101.88
National Best	0.82	-0.41	0.19	100.60
		$\nu = 0.75$		
Consolidated	28.30	-0.30		127.99
Slow Venue	1.24	-0.62		100.62
Fast Venue	3.96	-0.05		103.91
VWAP	3.08	-0.44		102.84
National Best	1.24	-0.62	0.29	100.91
		$\nu = 1$		
Consolidated	37.50	-0.35		137.15
Slow Venue	1.66	-0.83		100.84
Fast Venue	5.28	-0.06		105.22
VWAP	4.11	-0.58		103.80
National Best	1.66	-0.83	0.39	101.22

This table presents the values of the price distortions and of the equilibrium asset price relative to the frictionless Walrasian price, which is normalized to 100, in the long-run, for four values of the average investor type ν. Parameter values (except ν) correspond to the baseline calibration (Table II). In the long-run both investor participation and trading technologies are variable. The Walrasian price is the price in the absence of market contact and investor participation frictions. The VWAP corresponds to that between the slow and fast venues. The national best price correspond to the unique price in fragmented markets under price protection. The limited LPD and hte ILD are as in Definition 2.

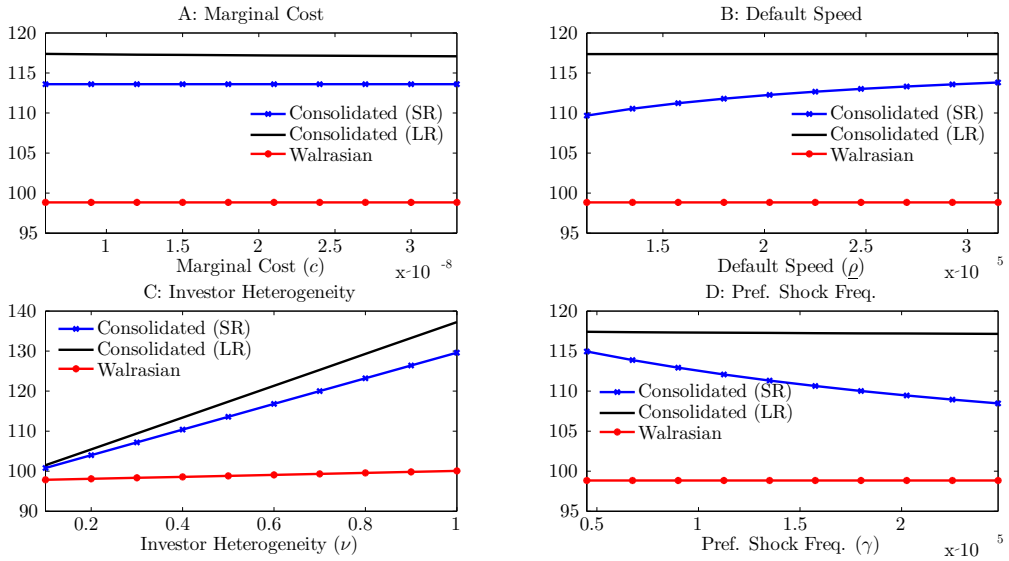

Figure 9: Comparative Statics of Prices in a Consolidated Market

This figure plots the effect of parameter changes on consolidated market price in the short-run (SR), long-run (LR), and in the frictionless Walrasian case. The parameter changes are relative to the baseline calibration (Table II). Panel A shows the marginal cost (c); Panel B shows the default speed ($\underline{\rho}$); Panel C shows the average investor type (ν), and Panel D the preference switching frequency (γ).

re allocation and, as the ILD approaches zero, the price becomes relatively insensitive to changes in γ.

An increase in the cost of speed decreases the asset price since, all else being equal, the exchange invests less in speed rendering the asset more illiquid. Interestingly, p_{con} is not affected in the long-run by the default speed level. This fact reflects the lack of competitive forces in a single exchange economy, and contrasts with the results for fragmented markets in the following section.

Fragmented Markets Figure 10 displays the effect of changes in parameter values on the VWAP and the national best price. As in the consolidated case, an increase in the marginal cost of speed decreases asset prices, both with and without price protection. We can observe in Panel A that the VWAP is more sensitive to this change. This is due to the fact that an increase in c changes the ILD in market 2 as well as the marginal investor types $\hat{\sigma}_1$ and $\hat{\sigma}_2$ due to easier speed differentiation. Only the latter effect impacts the equilibrium price in the protected market.

Interestingly, an increase in the level of default speed $\underline{\rho}$ decreases the VWAP. This is due

Table V: Trading Volume (Volume in Walrasian market=100)

	Short Run	Short Run (Protected)	Long Run	Long Run (Protected)
Consolidated	31.31		38.65	
Slow Venue	29.01	28.66	28.64	28.37
Fast Venue	58.23	57.74	59.61	58.82
Slow + Fast	87.23	86.40	88.26	87.19

This table presents the trading volumes in the steady-state equilibrium relative to the frictionless Walrasian case, which is normalized to 100. The parameter values correspond to the baseline calibration (Table II). In the short-run investor participation is variable, but trading technologies are fixed. The Walrasian case correspond to a market without contact or investor participation frictions. The volume is given by the instantaneous expected transaction rate as in Equations 35 and 36.

to the fact that, as market 1 becomes faster, speed differentiation becomes less effective in relaxing fee competition. Note that we interpret the default speed as a minimum level of speed that is acceptable to investors. Naturally, this effect could be mitigated if the slow market were able to successfully market a service with a speed below the default level. The effect on the national best price, however, can be very different. This is due to the fact that in the protected market, not only the marginal investor $\hat{\sigma}_1^{prot}$ (and thus the LPD) but also the ILD depends directly on the default speed level. An increase in $\underline{\rho}$ exerts opposite effects on the asset price: It decreases the ILD, raising the asset price, but it may reduce the LPD, by reducing the scope of differentiation among exchanges.

An increase in the average investor type increases the equilibrium price in all cases, by raising the value of the marginal investor type, but the effect is strongest in the absence of price protection. This is because an increase in ν provides the fast market with stronger incentives to invest in speed, reducing the ILD in market 2, with a direct impact on p_{vw}, but only affecting p_{nb} indirectly by further raising the value of the marginal investor in market 1. Finally, an increase in the preference switching rate γ has a twofold effect on the asset price. On the one hand, it increases the demand for trading and speed. On the other hand, it negatively impacts the market allocative efficiency and investors' value functions. The negative effect is relatively stronger in market 1, where speed remains fixed. Thus, the national best price increases at a slower rate than the VWAP following an increase in this rate.

E Trading Volume

Table V presents model-implied trading volumes relative to a Walrasian market. The effect of changes in the average investor type ν are moderate and thus, for brevity, we report values for the baseline calibration only. Our measure of volume is given by the steady state value

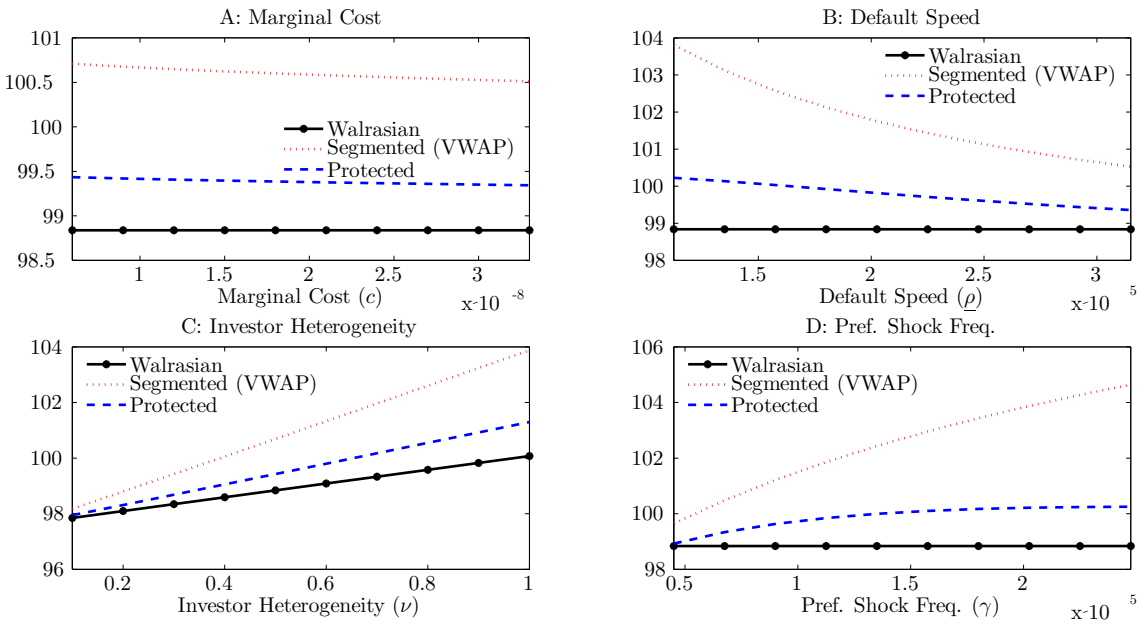

Figure 10: Comparative Statics of Prices in Fragmented Markets

This figure plots the effect of parameter changes on the VWAP in segmented markets, the single national best price with price protection, and the frictionless Walrasian price. The parameter changes are relative to the baseline calibration (Table II). Panel A shows the marginal cost (c); Panel B shows the default speed ($\underline{\rho}$); Panel C shows the average investor type (ν), and Panel D the preference switching frequency (γ).

of transaction rates per unit of time, as in Equations 35 and 36. We can observe that the lack of competition between the exchanges has a first order negative effect on volume: The total volume in a consolidated market is nearly half that in the fragmented market cases in both the short and long-run. However, note that the ability of the single exchange to invest in trading technologies has a large impact on traded values. Everything else being constant, when the average execution speed decreases from 20 seconds to 1 second, the relative transaction rate increases from 31.31 to 38.65.

In fragmented markets, the fast venue displays significantly higher volume than the slow venue. This volume difference increases in the long-run as the degree of speed differentiation increases. Furthermore, consistent with our results in Proposition 5, price protection decreases total investor participation in fragmented markets, generating a fall in traded volumes of approximately 1% of the Walrasian values in both the short and long-run.

F Trading Fragmentation

How do parameter changes affect the volume distribution across trading venues in fragmented markets? To answer this question we need a formal notion of trading fragmentation. A simple metric is given by $1 - HHI$, where HHI is the standard Herfindahl–Hirschman Index. Using the stationary equilibrium of the model, we can compute the HHI simply as the sum of the squared terms $(\frac{\tau_i}{\tau_i+\tau_j})^2$, where the trading rates τ_1 and τ_2 are given by Equations 35 and 36, respectively.

Figure 11 displays how fragmentation is affected by the trading environment in both segmented and protected markets. Note that a general pattern arises across panels: Everything else being equal, the level of fragmentation is higher under price protection. This makes sense in light of our discussion in Section IV. Price protection increases the demand for the slow market, reducing the amount of ex post competition between venues.

Lets analyze the effect of the cost parameters. We observe in Panels A and B of Figure 11 that, regardless of the trading regime, both an increase in the marginal speed cost and an increase in the default speed level induce higher trading fragmentation. The intuition behind this effect is that these cost parameter changes reduce markets' ability to differentiate, and thus exchanges are forced to compete more intensely in fees. As trading fees decrease, the relative volume of the fast market increases, reducing fragmentation.

Panels C and D show that an increase in preference parameters ν and γ induce the opposite effect on trading fragmentation. Everything else equal, an increase in the investor average type results in more speed-sensitive investors, which helps the fast venue attract a higher mass of the population and raises equilibrium trading volumes. Similarly, an increase in the

frequency of preference shocks increases the relative demand of the fast market, which is able to realize a greater fraction of the total gains from trade.

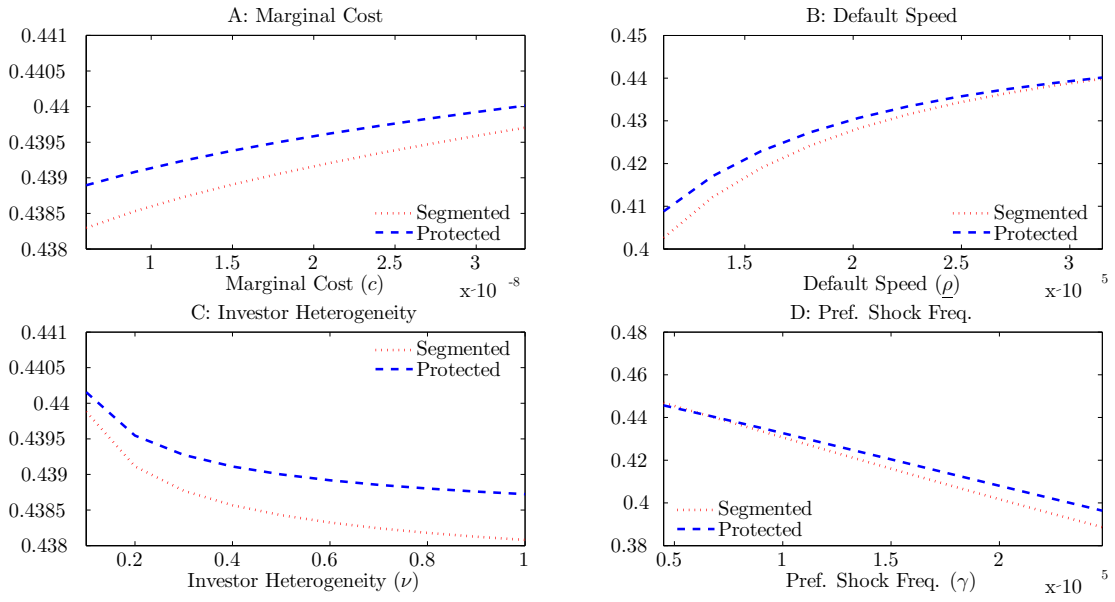

Figure 11: Comparative Statics of Trading Fragmentation

This figure plots the effect of parameter changes on fragmentation levels in the segmented and protected markets. Trading fragmentation is measured as 1-HHI. The parameter changes are relative to the baseline calibration (Table II). Panel A shows the marginal cost (c); Panel B shows the default speed ($\underline{\rho}$); Panel C shows the average investor type (ν), and Panel D the preference switching frequency (γ).

VI Discussion

Liquidity and Participation Frictions. The quantitative analysis suggests that distortions in participation incentives are important to understand the effects of market structures on liquidity and asset prices. Our analytical and quantitative analysis of participation distortions complements the work of Huang and Wang (2009). Note, however, an important difference: While participation costs that lead to lower market liquidity always decrease prices in their framework, they can increase prices in ours by changing both the mass and composition of investors.

Although speed plays a key role as a differentiating factor for exchanges, the frequency of market contact per se has a smaller effect on prices vis-a-vis participation distortions. For our baseline calibration, the long-run value of the LPD is two times larger than the ILD in the protected market, and seven times larger in the segmented market. The relatively small

value of the ILD is natural here given that our calibration is based on public equity data. It is also consistent with the calibration results of Gârleanu (2009). In contrast, Longstaff (2009) finds large pricing effects of illiquidity by calibrating an asset pricing model using U.S. private equity data.

Does the Empirical Evidence Support the Model's Results? The fact that fast trading enhances liquidity is documented by Hendershott, Jones, and Menkveld (2011) and Hasbrouck and Saar (2010), among others. Boehmer (2005) documents the trade-off between execution speed and costs in U.S. markets and finds that, analogously to market 2 in our model, the NASDAQ is more expensive than the NYSE, but also faster.

Recent studies by O'Hara and Ye (2011) on U.S. markets and by Foucault and Menkveld (2008) and Degryse, Jong, and Kervel (2011) on European markets support the model's prediction that liquidity increases with exchange competition. Importantly, our results on overall *asset price levels* suggest caution in interpreting such findings as definitive proof of the "benign" effects of fragmentation on market quality.

To the best of our knowledge, Amihud, Lauterbach, and Mendelson (2003) provide the only direct evidence of the effect of trading consolidation on asset prices. The authors study consolidation of two almost identical equity claims: A stock and a warrant on the stock that is deep in the money at expiration. They find that stock prices appreciate 1.27%, on average, on the warrant expiration. This empirical finding is consistent with Proposition 6. Arguably, their empirical analysis does not provide a direct test of our model's prediction, which is based on exchange competition. However, since the Tel Aviv Stock Exchange market power is not altered in their study, one could interpret such price appreciation as representing a reduction of the ILD, with the LPD part of the price remaining constant.

An International Asset Pricing Perspective. In Section V we compare the pricing outcomes of three stylized economies: (1) a consolidated market, (2) an unregulated fragmented market, and (3) a price-protected fragmented market. We argue that when market contact frictions are small (e.g., for stocks), the model offers a pricing ordering $p_{con} \geq p_{vw} \geq p_{nb}$. In our long-run calibration, these prices are equal to 118.73, 101.88, and 100.6, relative to the Walrasian price. How could an econometrician test this hypothesis? The first step would consist in identifying economic areas with market organizations that can be represented by our stylized cases. Incumbent exchanges in Hong Kong, China, Brazil and Spain, for example, experience little or no competition, and can thus be represented by (1) (see Figure 3). The U.S. and Canada are chief examples of financial markets with trade-through rules. Most of continental Europe and the U.K. can be seen as having unprotected

fragmented markets (see Appendix A) and thus are closest to (2). The second step would be to match assets across such areas with almost identical fundamentals (e.g., loadings on risk factors). An econometrician could then test whether within-groups prices are, say, highest in Hong Kong, intermediate in Europe, and lowest in the U.S.

Is Price Protection Important? Segmented equity markets are obviously an analytical abstraction here.[37] In real markets arbitrageurs and smart routing technologies work to (at least partially) undo price differentials between markets. Does this fact make a trade-through rule redundant? Interestingly, empirical evidence by Foucault and Menkveld (2008) suggest that the answer is no. These authors study the competition between a London Stock Exchange order book (EuroSETS) and Euronext Amsterdam for Dutch firms and find that, even when there is formal entry barrier to arbitrageurs, the trade-through rate in their sample equals 73%.

Our calibration results suggest that price protection distorts the asset price in the slow venue by approximately 20 bps in the baseline calibration. Although this value may seem low, note that we define the PPD conceptually as $p_{nb} - p_1^{seg}$. This difference may not represent the full effect of the introduction of such regulation. If a given market, say, the U.K., adopts a trade-through rule, the total pricing effect would instead be given by $p_{nb} - p_{vw}$, which equals 128 bps in our calibration. As mentioned, our VWAP may not accurately represent prices in partially integrated markets, but as an approximation it suggests that the total effect of regulation may in fact be larger than our PPD. Further research on this topic is needed to assess this issue quantitatively.

Beyond its effect on asset prices, price protection affects the nature of competition between exchanges in our model. This is consistent with an earlier discussion by Stoll (2006), who argues the following:

> "The casual observer of the heated debate that has surrounded the order protection rule may well wonder what the fuss is all about. After all, we are just talking about pennies. But for the exchanges, it may be a matter of business survival. Pennies matter, but more important, the rule requires the linkage of markets, which threatens established markets and benefits new markets. The battle appears to be over pennies, but in fact, it is over the ability of markets to separate themselves from the pack."

[37] If one reinterprets the model market choice from an international perspective, say, a choice between two European countries, stock market segmentation is far from negligible (Bekaert, Harvey, Lundblad, and Siegel (2011)).

Furthermore, by endogenizing the entry decisions of exchanges, Pagnotta and Philippon (2012) show that price protection can expand the ex ante number of trading venues.

Fast and Slow Traders. The result that equilibrium prices can increase in the asset supply through investors' speed investments is, to the best of our knowledge, new to the literature. Our result suggests that traders that invest heavily in speed concentrate on big cap assets, such as large S&P500 stocks or the E-mini futures. Industry reports and the evidence of Brogaard (2011) and Kirilenko, Kyle, Mehrmad, and Tuzun (2010) for high-frequency traders support this intuition.

VII Concluding Remarks

We study a tractable model that links the organization of financial markets with asset liquidity and prices. The model highlights the importance of competition between exchanges, their incentives to innovate to differentiate their services, and the linkages with investors' choices. We show that trading fragmentation can improve market quality, as measured by traditional measures of liquidity, while having negative effects on asset prices in the long-run. The model also provides, to the best of our knowledge, the first analysis of the impact of order protection regulations on asset prices.

Our model suffers from several limitations and suggests interesting avenues for further research. First, it would be natural to introduce heterogeneity in investors' information on the asset common value component. This would provide additional incentives for investors to demand speed. Promisingly, recent theory developments (e.g., Guerrieri, Shimer, and Wright (2010)) have made progress integrating search-like frictions and asymmetric information. Another interesting extension is to consider more sophisticated execution mechanisms. For example, one could exploit the modeling technology of Biais, Hombert, and Weill (2012) to include limit orders.

Empirically, it would be interesting to test the effects of fragmentation on prices more directly. In this regard, recent national regulation reforms that allow the introduction of new trading venues (e.g., South Korea, Australia, Brazil) can provide useful natural experiments for event studies. Such empirical work is important to guide policymakers and to further our understanding of the connections between secondary and primary markets and their effects on the cost of capital and other important variables for corporations.

The economics of the model are also relevant for large classes of derivatives that will start trading in exchanges in the near future. These market will see a growth in the number of

electronic platforms and liquidity will spread between Operating Trading Facilities, Swap Execution Facilities, and other institutions as mandated by regulations such as Dodd-Frank and EMIR. As derivatives and equity market structures continue to converge, understanding liquidity and pricing in multiple derivative markets will require some of the same techniques developed here.

References

Acemoglu, Daron, 1998, Why do new technologies complement skills? Directed Technical Change and Wage Inequality, *The Quarterly Journal of Economics* pp. 1055–1089. 7

Acharya, Viral V., and Lasse H. Pedersen, 2005, Asset pricing with liquidity risk, *Journal of Financial Economics* 77, 375–410. 8

Afonso, Gara M., 2011, Liquidity and congestion, *Journal of Financial Intermediation* 20, 324–360. 9

Allen, Franklin, and Douglas M. Gale, 1994, Limited Market Participation and Volatility of Asset Prices, *The American Economic Review* 84, 933–955. 9

Amihud, Yakov, Beni Lauterbach, and Haim Mendelson, 2003, The Value of Trading Consolidation : Evidence from the Exercise of Warrants, *Journal of Financial and Quantitative Analysis* 38, 829 –846. 45

Amihud, Yakov, and Haim Mendelson, 1986, Asset pricing and the bid-ask spread, *Journal of Financial Economics* 17, 223–249. 8

——— , and Lasse H. Pedersen, 2006, Liquidity and asset prices, *Foundations and Trends in finance* 1, 4086–99. 2, 8

Angel, James J., Lawrence E. Harris, and Chester S. Spatt, 2011, Equity Trading in the 21st Century, *The Quarterly Journal of Finance* 01, 1. 36

Bekaert, Geert, Campbell R. Harvey, Christian T. Lundblad, and Stephan Siegel, 2011, What Segments Equity Markets?, *Review of Financial Studies* 24, 3841–3890. 46

Biais, Bruno, 1993, Price Formation and Equilibrium Liquidity in Fragmented and Centralized Markets, *Journal of Finance* 48, 157–185. 8

——— , Johan Hombert, and Pierre olivier Weill, 2012, Pricing and Liquidity with Sticky Trading Plans, *Working Paper, UCLA*. 9, 47

Boehmer, Ekkehart, 2005, Dimensions of execution quality : Recent evidence for US equity markets, *Journal of Financial Economics* 78, 553–582. 45

Brennan, Michael J., 1975, The Optimal Number of Securities in a Risky Asset Portfolio When There are Fixed Costs of Transacting: Theory and Some Empirical Results, *Journal of Financial and Quantitative Analysis* 10, 483–496. 9

Brogaard, Jonathan A., 2011, High Frequency Trading and its Impact on Market Quality, *Working Paper, University of Washington*. 47

Chatterjee, Satyajit, and Dean Corbae, 1992, Endogenous Market Participation and the General Equilibrium Value of Money, *Journal of Political Economy* 100, 615–646. 9

Colliard, Jean-Edouard, and Thierry Foucault, 2012, Trading Fees and Efficiency in Limit Order Markets ., *Working Paper, HEC Paris*. 8

Constantinides, George M., 1986, Capital Market Equilibrium with Transaction Costs, *Journal of Political Economy* 94, 842–862. 8

Degryse, Hans, Frank De Jong, and Vincent Van Kervel, 2011, The impact of dark trading and visible fragmentation on market quality, . 4, 45

Duffie, Darrell, Nicolae Garleanu, and Lasse H. Pedersen, 2005, Over-the-Counter Markets, *Econometrica* 73, 1815–1847. 9, 14, 18, 19, 36

———— , 2007, Valuation in Over-the-Counter Markets, *Review of Financial Studies* 20, 1865–1900. 18

Eisfeldt, Andrea L., 2004, Endogenous Liquidity in Asset Markets, *Journal of Finance* LIX, 1–30. 8

Foucault, Thierry, and Albert J. Menkveld, 2008, Competition for Order Flow and Smart Order Routing Systems, *Journal of Finance* LXIII, 119–158. 4, 45, 46

Foucault, Thierry, and Christine A Parlour, 2004, Competition for Listings, *Rand Journal of Economics* 35, 329–355. 8

Frazzini, Andrea, and Lasse H. Pedersen, 2010, Betting Against Beta, . 5

Gabszewicz, Jaskold, and J.-F. Thisse, 1979, Price Competition, Quality and Income Disparities, *Journal of Economic Theory* 20, 340–359. 3, 8

Garbade, Kenneth D., and William L. Silber, 1977, Technology, Communication and the Performance of Financial Markets: 1840-1975, *Journal of Finance* 33, 819–832. 7

Gârleanu, Nicolae, 2009, Portfolio choice and pricing in illiquid markets, *Journal of Economic Theory* 144, 532–564. 45

Garleanu, Nicolae, Lasse H. Pedersen, and Allen M. Poteshman, 2009, Demand-Based Option Pricing, *Review of Financial Studies* 22, 4259–4299. 9

Glosten, Lawrence R., 1994, Is the electronic open limit order book inevitable?, *Journal of Finance* 49, 1127–1161. 8

Guerrieri, Veronica, Robert Shimer, and Randall Wright, 2010, Adverse Selection in Competitive Search Equilibrium, *Econometrica*. 47

Harris, Lawrence E., 1993, Consolidation, fragmentation, segmentation and regulation., *Financial Markets, Institutions & Instruments* 2, 1–28. 8

Hasbrouck, Joel, and Gideon Saar, 2010, Low-Latency Trading, . 45

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, 2011, Does Algorithmic Trading Improve Liquidity ?, *Journal of Finance* LXVI, 1–33. 3, 45

Hendershott, Terrence, and Haim Mendelson, 2000, Crossing Networks and Dealer Markets: Competition and Performance, *Journal of Finance* 55, 2071–2115. 8

Huang, Jennifer, and Jiang Wang, 2009, Liquidity and Market Crashes, *Review of Financial Studies* 22, 2607–2643. 9, 44

Kirilenko, Andrei, Albert S. Kyle, Samadi Mehrmad, and Tugkan Tuzun, 2010, The Flash Crash : The Impact of High Frequency Trading on an Electronic Market, . 47

Lagos, Ricardo, and Guillaume Rocheteau, 2009, Liquidity in Asset Markets With Search Frictions, *Econometrica* 77, 403–426. 9, 14, 19, 54, 1

Lo, Andrew W., Jiang Wang, and Harry Mamaysky, 2004, Asset Prices and Trading Volume under Fixed Transactions Costs, *Journal of Political Economy* 112, 1054–1090. 8

Longstaff, Francis A., 2009, Portfolio Claustrophobia : Asset Pricing in Markets with Illiquid Assets, *The American Economic Review* 99, 1119–1144. 45

Lucas, Robert E. Jr, 1978, Asset Prices in an Exchange Economy, *Econometrica* 46, 1429–1445. 2

Madhavan, Ananth, 1995, Consolidation, Fragmentation, and the Disclosure of Trading Information, *Review of Financial Studies* 8, 579–603. 8

Mankiw, N. Gregory, and Stephen P. Zeldes, 1991, The consumption of stockholders and nonstockholders, *Journal of Financial Economics* 29, 97–112. 9

Marshall, Alfred, 1890, *Principles of Economics* (Macmillan). 5

Mendelson, Haim, 1987, Consolidation, fragmentation, and market performance, *Journal of Financial and Quantitative Analysis* 22, 187–207. 8

Merton, Robert C., 1987, A Simple model of Capital Market Equilibrium with Incomplete Information, *Journal of Finance* 42, 483–510. 9

O'Hara, Maureen, and Mao Ye, 2011, Is market fragmentation harming market quality?, *Journal of Financial Economics* 100, 459–474. 3, 45

Pagano, Marco, 1989, Trading volume and Asset Liquidity, *Quarterly Journal of Economics* 104, 255–274. 8

Pagnotta, Emiliano S., and Thomas Philippon, 2012, Competing on Speed, *Working Paper, NYU Stern*. 8, 19, 47, 3, 11

Parlour, Christine A., and Duane J. Seppi, 2003, Liquidity-based competition for order flow, *Review of Financial Studies* 16, 329–355. 8

Rocheteau, Guillaume, and Randall Wright, 2005, Money in search equilibrium, in competitive equilibrium, and in competitive search equilibrium, *Econometrica* 73, 175–202. 19

Santos, Tano, and Jose A. Scheinkman, 2001, Competition Among Exchanges, *Quarterly Journal of Economics* 116, 225–1061. 8

Securities and Exchange Commission, 2010, *Concept Release on Equity Market Structure* No. 34. 2, 9, 34

Shaked, Avner, and John Sutton, 1982, Relaxing Price Competition Through Product Differentiation, *Review of Economic Studies* 49, 3–13. 8

Shleifer, Andrei, and Robert W. Vishny, 1997, The Limits of Arbitrage, *Journal of Finance* 52, 35. 26

Stoll, Hans R., 2006, Electronic Trading in Stock Markets, *Journal of Economic Perspectives* 20, 153–174. 46

Trejos, Alberto, and Randall Wright, 2012, Money and Finance : An Integrated Approach, . 9

Vayanos, Dimitri, 1998, Transaction Costs and Asset Prices: A Dynamic Equilibrium Model, *Review of Financial Studies* 11, 1–58. 3, 8

——— , and Tan Wang, 2007, Search and endogenous concentration of liquidity in asset markets, *Journal of Economic Theory* 136, 66–104. 9

Vissing-Jorgensen, Annette, 2002, Limited Asset Market Participation and the Elasticity of Intertemporal Substitution, *Journal of Political Economy* 110, 825–853. 9

Weill, Pierre-Olivier, 2007, Leaning Against the Wind, *Review of Economic Studies* 74, 1329–1354. 9

Weill, Pierre-olivier, 2008, Liquidity premia in dynamic bargaining markets, *Journal of Economic Theory* 140, 66–96. 9

Appendices

A Regulatory Frameworks

In this section we briefly discuss some regulations that are closely related to the model in the main body of the paper.

Entry of New Exchanges. The SEC introduced the Regulation of Exchanges and Alternative Trading Systems (Reg ATS) in 1998. It was designed with the aim of protecting investors and resolving concerns arising from alternative trading systems. A second objective of this regulation was to foster innovation in the space of trading systems and matching technologies by facilitating the entry of new participants. This intention is reflected in the "lax" control of trading venues that represent less than 5% of the trading volume for any given security.

In Europe, the Markets in Financial Instruments Directive (MiFID) implemented during the last half of the 2000s played a similar role in fostering competition between trading venues.

Investor Protection. There are essentially two approaches to investor protection: the trade-through model and the principles-based model.

Trade-Through Model. Under this approach price is the primary criterion for best execution. Market centers must be connected to one another and prevent trading through better prices available elsewhere, which requires complex connections as well as strong monitoring activity by regulators. In the U.S. the modern form of investor protection was introduced by Reg NMS. In particular, Rule 611 (trade-through) states that prices are quoted gross of trading fees (the SEC places a cap on fees) and only the top of the book is protected: When a big trading order arrives in a given marketplace, only the amount of shares represented by the depth of the book at the national best bid and offer is protected. As an example, suppose that NASDAQ and NYSE are the only market centers and that an investor submits a market order to buy 100,000 shares of a given stock to NASDAQ. Currently the ask price at the NASDAQ is higher than the ask price at the NYSE (where the ask depth is 10,000 shares).

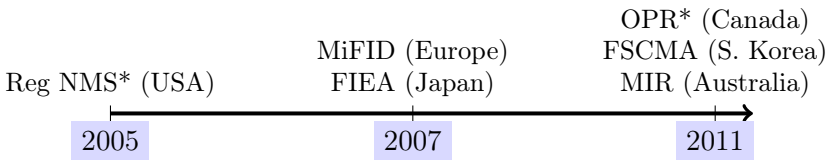

Figure 12: Timeline of the Introduction of Current Market Regulations
This timeline shows the year of introduction of several regulations related to competition between trading venues. Regulations marked by an asterisk (*) contain a trade-through provision. In the U.S. Reg NMS was not fully implemented until 2007. Source: www.fidessa.com

Then NASDAQ can either match the price at the NYSE or the first execution occurs at the NYSE for 10,000 shares. The remaining 90,000 shares "walk up" the book at the NASDAQ.

In Canada, the Order Protection Rule (OPR) implemented by the Investment Industry Regulatory Organization of Canada shares the same spirit, but aims to protect orders throughout the entire order book (as opposed to just the top level).

Principles-Based Model. Criteria other than price, such as the type of investor behind the trade, are included in the best execution policy here. Thus, this approach allows for more discretion and less transparency in the assessment of execution results. This approach best represents the spirit of Europe's MiFID and Japan's Financial Instruments and Exchange Act. For example, Article 40-2(1) of the Financial Instruments and Exchange Act defines best execution policy as a "method for executing orders from customers ... under the best terms and conditions." Some of the criteria to be considered are the listing trading venue, price, liquidity, execution probability, and execution speed. In Japan this system does not apply to professional investors. In both Europe and Japan, sell-side best execution policies are not obliged to consider every venue. The monitoring of execution quality is generally left to clients, which can be a problem in countries where investors have inadequate knowledge of financial markets. The claimed advantages of the principles-based approach lie in a much simpler set of linkages between markets and promoting innovation by not forcing uniformity.

B Generalized Asset Holdings

In this section we generalize investors' asset demand holdings allowing for asymmetric shock frequency and general asset holdings $a \geq 0$.

Let the preference shocks be represented by $\epsilon_i \in \{\epsilon_l, \epsilon_h\}$ and let ϕ be the probability of the high shock. An investor of type σ under state ϵ_i and with asset holdings $a \geq 0$ enjoys a

utility flow

$$u_{i\sigma}(a) = \theta_{i\sigma} u(a)$$

where $\theta_{i\sigma} = \mu + \epsilon_i \sigma$. The adjusted utility in Equation 39 becomes

$$\begin{aligned}\overline{u}_{i\sigma}(a) &= \frac{(r+\rho)\, u_{i\sigma}(a) + \gamma E_i\left[u_{i'\sigma}(a)\right]}{r+\gamma+\rho} \\ &= \underbrace{\left(\mu + \frac{(r+\rho)\,\epsilon_i\sigma + \gamma(2\phi-1)\,\sigma}{r+\gamma+\rho}\right)}_{\equiv\overline{\theta}_{i\sigma}} u(a)\end{aligned}$$

Optimal portfolio holdings in an interior solution satisfy $\overline{\theta}_{i\sigma} u'(a) = rp$. Thus, whenever u is invertible we have

$$a^*_{i\sigma} = \left(u'\right)^{-1}\left(\frac{rp}{\overline{\theta}_{i\sigma}}\right).$$

The market clearing condition is

$$\int_\sigma \sum_i \phi_i a^*_{i\sigma}(p)\, dG(\sigma) = \overline{a}.$$

Consequently, the equilibrium price p solves

$$\int_\sigma \sum_i \phi_i \left(u'\right)^{-1}\left(\frac{rp\,(r+\gamma+\rho)}{(r+\gamma+\rho)\,\mu + (r+\rho)\,\epsilon_i\sigma + \gamma(2\phi-1)\,\sigma}\right) dG(\sigma) = \overline{a}.$$

Example: CRRA utility Let $u(a)$ be given by $\frac{a^{1-\xi}}{1-\xi}$. The asset demand function then satisfies $\overline{\theta}_{i\sigma} a^{-\xi} = rp$, and thus

$$a^*_{i\sigma}(p) = (rp)^{-\frac{1}{\xi}}\left(\overline{\theta}_{i\sigma}\right)^{\frac{1}{\xi}}.$$

The market clearing condition is

$$\begin{aligned}&\int_\sigma \left\{\phi\,(rp)^{-\frac{1}{\xi}}\left(\overline{\theta}_{h\sigma}\right)^{\frac{1}{\xi}} + (1-\phi)\,(rp)^{-\frac{1}{\xi}}\left(\overline{\theta}_{l\sigma}\right)^{\frac{1}{\xi}}\right\} dG(\sigma) = \overline{a} \\ &\qquad \Rightarrow \int_\sigma \left\{\phi\left(\overline{\theta}_{h\sigma}\right)^{\frac{1}{\xi}} + (1-\phi)\left(\overline{\theta}_{l\sigma}\right)^{\frac{1}{\xi}}\right\} dG(\sigma) = \overline{a}\,(rp)^{\frac{1}{\xi}}\end{aligned}$$

To simplify things further, consider the normalization $\mu = 0$, $\phi = 1/2$, and assume A.1. Then, the market clearing condition becomes

$$\frac{1}{2}\left(\frac{r+\rho}{r+\gamma+\rho}\right)^{\frac{1}{\xi}}\left(\epsilon_l^{\frac{1}{\xi}}+\epsilon_h^{\frac{1}{\xi}}\right)\int_\sigma \sigma^{\frac{1}{\xi}}\frac{e^{-\frac{\sigma}{\nu}}}{\nu}d\sigma = \overline{a}\,(rp)^{\frac{1}{\xi}}\,.$$

Thus, the equilibrium price is given by

$$p = \frac{\nu}{r}\left[\frac{1}{2\overline{a}}\left(\frac{r+\rho}{r+\gamma+\rho}\right)\left(\epsilon_l^{\frac{1}{\xi}}+\epsilon_h^{\frac{1}{\xi}}\right)\Gamma\left(1+\frac{1}{\xi}\right)\right]^{\xi}, \tag{37}$$

where Γ denotes the Gamma function.

Note the following properties of equilibrium price 37. First, differently from the consolidated price in Section III, it decreases smoothly in the asset supply $\overline{a}$. Second, whether higher market speed raises the asset price depends on the elasticity of asset demand, driven by parameter ξ. Consistent with Proposition 5 in Lagos and Rocheteau (2009), $\xi \in (0,1)$ is a sufficient condition for $\frac{\partial p}{\partial \rho} > 0$.

Asset Pricing Frictions in Fragmented Markets

-Online Appendix-

Emiliano Pagnotta

New York University Stern School of Business

This Appendix comprises proofs of propositions and lemmas in the main paper.

Proof of Lemma 1

We follow Lagos and Rocheteau (2009) in expressing the optimal holdings problem recursively as follows[38]

$$a^{*}\left(p;\sigma,\epsilon\right)=\arg\max_{a\in\{0,1\}}\left\{\bar{u}\left(a;\sigma,\epsilon\right)-rpa\right\} \tag{38}$$

where $\overline{u}$, the adjusted holding utility, is given by

$$\bar{u}\left(a;\sigma,\epsilon\right)\equiv\frac{\left(r+\rho\right)u_{\sigma,\epsilon}\left(a\right)+\gamma\mathbb{E}\left[u_{\sigma,\epsilon'}\left(a\right)\mid\epsilon\right]}{r+\rho+\gamma}. \tag{39}$$

The RHS of Equation 39, denoted as the adjusted holding utility, represents the expected average utility when holding the asset, for a given ϵ. Note that since ϵ is i.i.d. with mean zero, we have $\mathbb{E}\left[u_{\sigma,\epsilon'}\left(a\right)\mid\epsilon\right]=\mu a$ for any a and any ϵ. This expected utility over ϵ' does not depend on σ or ϵ. This implies that

$$\bar{u}\left(a;\sigma,\epsilon\right)=\left(\mu+\sigma\epsilon\frac{r+\rho}{r+\rho+\gamma}\right)a. \tag{40}$$

From Equation 38 it is clear that $\hat{\sigma}$ satisfies

$$\bar{u}\left(a;\hat{\sigma},1\right)=rpa. \tag{41}$$

[38] See Lemma 1 there. The Lemma only needs to be adapted to take into account heterogeneity in σ.

Combining Equations 40 and 41 when $a = 1$, yields Equation 4. Thus, the displayed demand functions represent optimal holdings. □

Proof of Proposition 1

Case $\overline{a} < 1/2$. Combining Equations 4 and 5 yields $\hat{\sigma} = G^{-1}(1 - 2\overline{a})$. Replacing this value back in Equation 4, and re arranging, delivers the equilibrium price in Equation 6. Also note that

$$\lim_{\rho \to \infty} p = p_W = \frac{1}{r}\left[\mu + G^{-1}(1 - 2\overline{a})\right].$$

We analyze next equilibrium allocations. Consider first a type $(\epsilon = +1, a = 1)$. This type is satisfied with its current holding and does not trade even if it contacts the market. Outflows result only from changes of ϵ from $+1$ to -1, which happens with intensity $\gamma/2$. There are two sources of inflow: types $(\epsilon = -1, a = 1)$ that switch to $\epsilon = 1$ and types $(\epsilon = +1, a = 0)$ that purchase one unit when they contact the market. In steady state, outflows must equal inflows:

$$\frac{\gamma}{2}\alpha_{\sigma,+}(1) = \frac{\gamma}{2}\alpha_{\sigma,-}(1) + \rho\alpha_{\sigma,+}(0). \tag{42}$$

Dynamics for types $(\epsilon = -1, a = 0)$ are similar:

$$\frac{\gamma}{2}\alpha_{\sigma,-}(0) = \rho\alpha_{\sigma,-}(1) + \frac{\gamma}{2}\alpha_{\sigma,+}(0). \tag{43}$$

For types $(\epsilon = +1, a = 0)$ and $(\epsilon = -1, a = 1)$ trade creates outflows so we have

$$\left(\frac{\gamma}{2} + \rho\right)\alpha_{\sigma,+}(0) = \frac{\gamma}{2}\alpha_{\sigma,-}(0) \tag{44}$$

$$\left(\frac{\gamma}{2} + \rho\right)\alpha_{\sigma,-}(1) = \frac{\gamma}{2}\alpha_{\sigma,+}(1) \tag{45}$$

Finally, the shares must add up to one, therefore

$$\sum_{\epsilon=\pm, a=0,1} \alpha_{\sigma,\epsilon}(a) = 1 \tag{46}$$

To see the steady state allocations, add (42) and (45) to get $\alpha_{\sigma,-}(1) = \alpha_{\sigma,+}(0)$. This immediately implies $\alpha_{\sigma,-}(0) = \alpha_{\sigma,+}(1)$. Using (42), we obtain $\alpha_{\sigma,+}(1) = \left(1 + 2\frac{\rho}{\gamma}\right)\alpha_{\sigma,-}(1)$. We can then solve for the shares of each type $\alpha_{\sigma,+}(1) = \frac{1}{4}\frac{\gamma+2\rho}{\gamma+\rho}$ and $\alpha_{\sigma,+}(0) = \frac{1}{4}\frac{\gamma}{\gamma+\rho}$. Notice also that the market clearing condition among high types is

simply $\alpha_{\sigma,+}(1) + \alpha_{\sigma,-}(1) = 1/2$. It is immediate that when $\rho \to \infty$, the Walrasian limits of these allocations satisfy $\alpha_{\sigma,-}(1) = \alpha_{\sigma,+}(0) = 0$.

Case $\overline{a} \geq 1/2$. When $a > 1/2$, sellers become the short side of the market and the equilibrium price equals $p = \frac{\mu}{r} - \frac{\hat{\sigma}}{r}\left(\frac{r+\rho}{r+\gamma+\rho}\right)$. In the knife-edge case $\overline{a} = 1/2$, the equilibrium price belongs to the interval

$$\left[\frac{\mu}{r} - \frac{\hat{\sigma}}{r}\left(\frac{r+\rho}{r+\gamma+\rho}\right), \frac{\mu}{r} + \frac{\hat{\sigma}}{r}\left(\frac{r+\rho}{r+\gamma+\rho}\right)\right].$$

□

Proof of Lemma 2

This lemma is a particular case of Proposition 1 in Pagnotta and Philippon (2012), which applies to the costless participation environment here. We include the derivation of the function W for completeness. Define $I_{\sigma,\epsilon} \equiv V_{\sigma,\epsilon}(1) - V_{\sigma,\epsilon}(0)$ as the value of owning the asset for type (σ, ϵ). Then, taking differences of Equations 7-10 we get

$$\begin{aligned} rI_{\sigma,-} &= \mu - \sigma + \frac{\gamma}{2}\left(I_{\sigma,+} - I_{\sigma,-}\right) + \rho\left(p - I_{\sigma,-}\right) \\ rI_{\sigma,+} &= \mu + \sigma - \frac{\gamma}{2}\left(I_{\sigma,+} - I_{\sigma,-}\right) - \rho\left(I_{\sigma,+} - p\right) \end{aligned}$$

We can then solve $r\left(I_{\sigma,+} - I_{\sigma,-}\right) = 2\sigma - (\gamma+\rho)\left(I_{\sigma,+} - I_{\sigma,-}\right)$ and obtain the gains from trade for type σ in market ρ: $I_{\sigma,+} - I_{\sigma,-} = \frac{2\sigma}{r+\gamma+\rho}$. We then have $I_{\sigma,\epsilon} = \frac{\mu+\rho p}{r+\rho} + \epsilon\frac{\sigma}{r+\gamma+\rho}$ and the average values

$$\begin{aligned} \bar{V}_\sigma(0) &= \frac{\rho}{2r}\left(I_{\sigma,+} - p\right) \\ \bar{V}_\sigma(1) &= \frac{\mu}{r} + \frac{\rho}{2r}\left(p - I_{\sigma,-}\right) \end{aligned}$$

where $\bar{V}_\sigma(0) \equiv \frac{V_{\sigma,+}(0)+V_{\sigma,-}(0)}{2}$ and $\bar{V}_\sigma(1) \equiv \frac{V_{\sigma,+}(1)+V_{\sigma,-}(1)}{2}$.

Let us now compute the ex ante value functions. Let us first consider types $\sigma < \hat{\sigma}$. They join the market to sell at price p, and then do not trade again. Averaging over types $\epsilon = \pm 1$, we get the ex ante value function $\hat{W}$ that solves the Bellman equation

$$r\hat{W} = \mu\bar{a} + \rho\left(p\bar{a} - \hat{W}\right) \Longrightarrow \hat{W} = \frac{\mu+\rho p}{r+\rho}\bar{a}$$

Since $\mu + \rho p = \frac{\mu}{r}(r+\rho) + \rho\left(p - \frac{\mu}{r}\right)$ we can rewrite

$$\hat{W} = \frac{\mu\bar{a}}{r} + \frac{\rho}{r+\rho}(rp - \mu)\frac{\bar{a}}{r}$$

From the definition of $\hat{\sigma}$ in Equation 4 and $s(\rho) \equiv \frac{\rho}{r+\gamma+\rho}$ we then have $\hat{W} = \frac{\mu\bar{a}}{r} + s\frac{\bar{a}}{r}\hat{\sigma}$. Note that $\hat{W}$ does not depend on the type σ, but only on the price and speed of the market. Of course we also have $\hat{W} = \bar{a}\bar{V}_{\hat{\sigma}}(1)$.

Let us now consider the steady state types, $\sigma > \hat{\sigma}$. Since the probability of owning one unit of the asset is $\bar{a}$, we have

$$W(\sigma) = \bar{a}\bar{V}_{\sigma}(1) + (1-\bar{a})\bar{V}_{\sigma}(0).$$

Using the expression above, we get

$$\begin{aligned} W_{\sigma} &= \bar{a}\mu + \bar{a}\frac{\rho}{2r}(p - I_{\sigma,-}) + (1-\bar{a})\frac{\rho}{2r}(I_{\sigma,+} - p) \\ &= \frac{\mu\bar{a}}{r} + \bar{a}\frac{\rho}{r}\frac{rp-\mu}{r+\rho} + \frac{1}{2r}\left(\frac{\rho}{r+\rho}(\mu - rp) + \frac{\rho}{r+\gamma+\rho}\sigma\right) \\ &= \frac{\mu\bar{a}}{r} + \frac{\bar{a}}{r}s(\rho)\hat{\sigma} + \frac{1}{2r}s(\rho)(\sigma - \hat{\sigma}) \end{aligned}$$

Therefore, we have, when $\sigma > \hat{\sigma}$, we have

$$W(\sigma, \rho) = \hat{W} + \frac{1}{2}s\frac{\sigma - \hat{\sigma}}{r}$$

□

Proof of Proposition 2

We proceed in two steps, first analyzing the FOC of Equation 19 with respect to q and s, and then studying the price decomposition in Equation 21.

Step 1. Note that the FOC with respect to q is the same in the short and in the long-run equilibrium, as given by Equation 17. Under A.1, $\hat{\sigma}^{L}_{con} = \hat{\sigma}^{S}_{con} = \nu$ and thus $q_{con} = \frac{\bar{a}s}{r}\nu$. We can then write the single exchange program as

$$\max_{1 \geq s \geq 0} \frac{s}{2r}\frac{\nu}{e} - C(s).$$

The interior solution FOC with respect to s is $(1-s)^2 = \left(\frac{e}{\nu}\right)2rc(\gamma + r)$. The optimal effective speed is then

$$s_{con} = 1 - (2rce(\gamma + r))^{1/2}\nu^{-1/2}. \tag{47}$$

Step 2. Using Equation 16, and $\lambda \equiv \frac{r+\gamma s}{r+\gamma}$ we have

$$p^L_{con} = \frac{\mu}{r} + \frac{\hat{\sigma}_{con}}{r}\lambda_{con} \tag{48}$$

Adding and subtracting $\frac{1}{r}G^{-1}(1-2\overline{a})$ and $\frac{\hat{\sigma}_{con}}{r}$ to the RHS of Equation 48, and re arranging, we have

$$p^L_{con} = \frac{1}{r}\left[\mu + G^{-1}(1-2\overline{a})\right] + \frac{1}{r}\left[\hat{\sigma}_{con} - G^{-1}(1-2\overline{a})\right] - \frac{\hat{\sigma}_{con}}{r}(1-\lambda_{con}) \tag{49}$$

Note that the first term of the RHS of 49 corresponds to the expression for p_W in Proposition 1. Further, note from Equation 47 that $\lim_{c\to 0} s_{con} = 1$, which implies $\lim_{c\to 0}\lambda_{con} = 1$. Using this fact and Definition 2, we have that the limited participation distortion equals

$$\begin{aligned}\lim_{c\to 0}[p - p_W] &= \frac{1}{r}\left[\hat{\sigma}_{con} - G^{-1}(1-2\overline{a})\right] \\ &= \frac{\nu}{r}\left[1 + \ln(2\overline{a})\right],\end{aligned}$$

where the second equality uses A.1. Finally, the illiquidity discount is given by

$$\frac{\hat{\sigma}_{con}}{r}(1-\lambda_{con}) = \frac{\nu}{r}\left(\gamma\sqrt{\frac{2rc}{r+\gamma}\frac{e}{\nu}}\right).$$

where the second equality uses A.1 and Equation 47. □

Proof of Proposition 3

The decomposition of the equilibrium price is an in Proposition 2. To derive the components of the equilibrium price, we proceed in two steps. First, we derive the first order condition of the investor speed choice problem, which yields $\sigma \mapsto \theta(\sigma)$; and we compute the marginal type $\hat{\sigma}$ for a given participation fee q. Second, we solve the exchange optimization program. Finally, we analyze the effect of changes in $\overline{a}$ on s.

Step 1. Investors optimization.

Before trading, a type σ investor maximizes the RHS of Equation 24, which can be written as

$$\max_{0\leq\theta\leq\overline{\theta}}\left\{s\left(\theta\right)\left[\frac{\overline{a}}{r}\hat{\sigma}+\frac{1}{2r}\max\left(0;\sigma-\hat{\sigma}\right)\right]-c_I\theta\right\}.$$

The interior first order condition at $\sigma=\hat{\sigma}$ yields

$$\theta\left(\hat{\sigma}\right)=\max\left\{0,\min\left\{\overline{\theta},\left(\frac{\overline{a}}{r}\frac{\left(\gamma+r\right)}{c_I}\hat{\sigma}\right)^{\frac{1}{2}}-\left(r+\gamma+\rho-\overline{\theta}\right)\right\}\right\},\tag{50}$$

which by Equation 22 implies that, at an interior solution,

$$s\left(\hat{\sigma}\right)=1-\left[\frac{r}{\overline{a}}\left(\gamma+r\right)c_I\right]^{\frac{1}{2}}\hat{\sigma}^{-\frac{1}{2}}=1-\left(y\hat{\sigma}\right)^{\frac{1}{2}},\tag{51}$$

where $y\equiv\frac{r}{\overline{a}}\left(\gamma+r\right)c_I$. The marginal investor $\hat{\sigma}$ satisfies

$$q=\tilde{W}\left(\hat{\sigma},\hat{\sigma},s\left(\hat{\sigma}\right)\right)-W_{out}=s\left(\hat{\sigma}\right)\frac{\overline{a}}{r}\hat{\sigma}-c_I\theta\left(\hat{\sigma}\right).\tag{52}$$

Then we have

$$\tilde{W}=\frac{\overline{a}}{r}\left[\hat{\sigma}-2\left(y\hat{\sigma}\right)^{\frac{1}{2}}\right]+c_I\left(r+\gamma+\rho-\overline{\theta}\right).$$

Step 2. Exchange optimization.

Using Equation 52 we can write the exchange program as

$$\max_{\hat{\sigma}\in[0,\overline{\sigma}]}\left\{\left(\frac{\overline{a}}{r}s\left(\hat{\sigma}\right)\hat{\sigma}-c_I\theta\left(\hat{\sigma}\right)\right)\left[1-G\left(\hat{\sigma}\right)\right]\right\}.$$

Using $\frac{1-G(\sigma)}{g(\sigma)}=\nu$ by A.1, we can write the first order condition as follows

$$\nu\left(\frac{\overline{a}}{r}\left(s\left(\hat{\sigma}\right)+\hat{\sigma}s'\left(\hat{\sigma}\right)\right)-c_I\theta'\left(\hat{\sigma}\right)\right)=\frac{\overline{a}}{r}s\left(\hat{\sigma}\right)-c_I\theta\left(\hat{\sigma}\right).\tag{53}$$

Using Equations 50 and 51, we can express Equation 53 as the following polynomial equation

$$F\left(\hat{\sigma}\right)\equiv\hat{\sigma}^{\frac{3}{2}}-2y^{\frac{1}{2}}\hat{\sigma}+\left(\frac{r}{\overline{a}}c_I\left(r+\gamma+\rho-\overline{\theta}\right)-\nu\right)\hat{\sigma}^{\frac{1}{2}}+\nu y^{\frac{1}{2}}.\tag{54}$$

The polynomial 54 has three roots. Descartes' rule of signs suggest that $F\left(\hat{\sigma}\right)$ has two or zero positive roots. Note that this holds regardless of the sign of the $\hat{\sigma}^{\frac{1}{2}}$ term coefficient. A numerical analysis based on our baseline calibration indicates that F has two positive roots, only one of which satisfies the condition $s\left(\hat{\sigma}\right)>\underline{s}$. Substituting

this solution, σ_{hft}, in Equation 51, and using $\lambda \equiv \frac{r+\gamma s}{r+\gamma}$, yields λ_{hft}.

Computing $\frac{ds}{d\overline{a}}$.

Using 51, we can re express 54 as

$$H(s,\overline{a}) = \frac{r}{\overline{a}}(\gamma + r)\, c_I \frac{(2s-1)}{(1-s)^2} - \nu s + \frac{r}{\overline{a}} c_I \left(r + \gamma + \rho - \overline{\theta}\right). \tag{55}$$

Given 55, we have $\frac{ds}{d\overline{a}} = -\frac{H_{\overline{a}}}{H_s}$. Computing the corresponding derivatives, we obtain

$$\frac{ds}{d\overline{a}} = \frac{(2s-1)(1-s) + (1-s)^3\left(1 + \frac{\rho-\overline{\theta}}{r+\gamma}\right)}{2\overline{a}(3s-1)}$$

Note that if $s \approx 1$ then we have $\frac{ds}{d\overline{a}} > 0$. Consequently, when market frictions are relatively small, investors' investment in low latency technology increase with $\overline{a}$.

□

Proofs of Proposition 4

In the case of segmented markets prices are formed independently. We can then apply the single equilibrium formula in Equation 16 to market $i \in \{1,2\}$

$$p_i = \frac{\mu}{r} + \frac{\hat{\sigma}_i}{r}\lambda_i. \tag{56}$$

Adding and subtracting $\frac{1}{r}G^{-1}(1-2\overline{a})$ and $\frac{\hat{\sigma}_i}{r}$ to the RHS of Equation 56, and re-arranging, we have

$$p_i = p_W + \frac{1}{r}\left[\hat{\sigma}_i - G^{-1}(1-2\overline{a})\right] - \frac{\hat{\sigma}_i}{r}(1-\lambda_i), \tag{57}$$

which yields Equations 30 and 31. For temporary investors to be indifferent between joining and staying out, we must have $W(\hat{\sigma}_i, \hat{\sigma}_i, s_i) - W_{out} - q_i = 0$ in each market i. Otherwise all the low types would strictly prefer one market to another. For market 1 this condition is satisfied by Equation 26. For market 2 then we must have

$$q_2 = \frac{\overline{a} s_2 \hat{\sigma}_2}{r}. \tag{58}$$

Recall that the type that is indifferent between joining market 1 and 2 is given by Equation 28. When both markets are active we have $\hat{\sigma}_1 < \hat{\sigma}_2 < \hat{\sigma}_{12}$, which implies

that the limited participation distortion is higher in market 2. Note that whether the illiquidity premium is higher in market one depends on whether $\hat{\sigma}_1 (1-\lambda_1) \lesseqgtr \hat{\sigma}_2 (1-\lambda_2)$. Since $\hat{\sigma}_1 < \hat{\sigma}_2$ and $1-\lambda_1 > 1-\lambda_2$, the relative size of the illiquidity premiums is not determined a priori.

We now characterize the conditions that determine the equilibrium values of $(\hat{\sigma}_1^{seg}, \hat{\sigma}_{12}^{seg})$. In the fee competition stage, exchange $i \in \{1,2\}$ seek to maximize the profit function π_i^{seg} of Section IV with respect to q_i, taking as given the conditions describing the affiliation of investors to markets 1 and 2, that is Equations 26 and 28. The system of first-order conditions is then given by

$$1 - G(\hat{\sigma}_{12}) = g(\hat{\sigma}_{12})(\hat{\sigma}_{12} + \hat{\sigma}_1 k), \tag{59}$$

$$G(\hat{\sigma}_{12}) - G(\hat{\sigma}_1) = (g(\hat{\sigma}_1) + k g(\hat{\sigma}_{12}))\hat{\sigma}_1, \tag{60}$$

where $k \equiv \frac{s_1}{s_2 - s_1}$. Given the equilibrium values of $(\hat{\sigma}_1, \hat{\sigma}_{12})$, from Equation 28 we have

$$\hat{\sigma}_2 = \frac{1}{s_2}(\hat{\sigma}_{12}(s_2 - s_1) + s_1 \hat{\sigma}_1). \tag{61}$$

The types $\hat{\sigma}_1$, $\hat{\sigma}_2$ and $\hat{\sigma}_{12}$ are then defined by Equations 59-61.

□

Proof of Proposition 5

In the protected market case there is a single price p_{nb} satisfying

$$p_{nb} = \frac{\mu}{r} + \frac{\hat{\sigma}_1^{prot}}{r}\lambda_1. \tag{62}$$

Adding and subtracting $\frac{1}{r}G^{-1}(1-2\overline{a})$, $\frac{\hat{\sigma}_1^{seg}}{r}$, and $\frac{\hat{\sigma}_1^{seg}}{r}\lambda_1$ to the RHS of Equation 56, and re-arranging, yields Equation 33.

We now show that under A.1 the price protection distortion is positive, which is equivalent to $\sigma_1^{prot} - \sigma_1^{seg} > 0$. From Equation (4), a single asset price in both markets implies

$$\left(1 + \frac{\gamma}{r+\rho_1}\right)\hat{\sigma}_2 = \left(1 + \frac{\gamma}{r+\rho_2}\right)\hat{\sigma}_1. \tag{63}$$

This means that in the protected case. $\hat{\sigma}_2 < \hat{\sigma}_1$. For $\hat{\sigma}_{12}$ we have

$$\frac{s_2\bar{a}\hat{\sigma}_2}{r} + \frac{s_2}{2r}\left(\hat{\sigma}_{12} - \hat{\sigma}_2\right) - q_2 = \frac{s_1\bar{a}\hat{\sigma}_1}{r} + \frac{s_1}{2r}\left(\hat{\sigma}_{12} - \hat{\sigma}_1\right) - q_1.$$

We can write it as

$$\frac{s_2 - s_1}{2r}\hat{\sigma}_{12} = q_2 - q_1 + \left(\frac{s_1\hat{\sigma}_1}{r} - \frac{s_2\hat{\sigma}_2}{r}\right)\left(\bar{a} - \frac{1}{2}\right)$$

then we can use (63) to write $\hat{\sigma}_2 = \frac{1+\frac{\gamma}{r+\rho_2}}{1+\frac{\gamma}{r+\rho_1}}\hat{\sigma}_1$. Then

$$\frac{s_2 - s_1}{2r}\hat{\sigma}_{12} = q_2 - q_1 + \frac{s_1\hat{\sigma}_1}{r}\left(1 - \frac{s_2}{s_1}\frac{1+\frac{\gamma}{r+\rho_2}}{1+\frac{\gamma}{r+\rho_1}}\right)\left(\bar{a} - \frac{1}{2}\right)$$

and since $\hat{\sigma}_1 = \frac{rq_1}{\bar{a}s_1}$ we get

$$\frac{s_2 - s_1}{2r}\hat{\sigma}_{12} = q_2 - q_1\left(\frac{1}{2\bar{a}} - \frac{\frac{\rho_2}{r+\rho_2}}{\frac{\rho_1}{r+\rho_1}}\left(\frac{1}{2\bar{a}} - 1\right)\right).$$

Re-arranging the above expression yields

$$\hat{\sigma}_{12}^{prot} = \frac{2r}{s_2 - s_1}\left(q_2 - \frac{z}{2\bar{a}}q_1\right), \tag{64}$$

where $z \equiv 1 - \frac{1+\frac{r}{\rho_1}}{1+\frac{r}{\rho_2}}(1 - 2\bar{a})$.

The profits of market 1 are $q_1\left(G\left(\hat{\sigma}_{12}\right) - G\left(\hat{\sigma}_1\right) + \delta_1\right)$. To simplify the notation, let $\alpha \equiv 2\bar{a}$ and $k \equiv \frac{s_1}{s_2 - s_1}$. Using Equation 32, we can write exchanges' second stage programs as

$$\begin{aligned}\max_{q_1}\pi_1^{prot} =& \frac{q_1}{\alpha}\left(1 - \alpha + \alpha G\left(\hat{\sigma}_{12}\right) - G\left(\hat{\sigma}_1\right)\right)\\ \max_{q_2}\pi_2^{prot} =& q_2\left(1 - G\left(\hat{\sigma}_{12}\right)\right)\end{aligned}$$

The conditions $\frac{\partial\pi_1^{prot}}{\partial q_1} = 0$ and $\frac{\partial\pi_2^{prot}}{\partial q_2} = 0$ lead to

$$1 - G\left(\hat{\sigma}_{12}\right) = g\left(\hat{\sigma}_{12}\right)\left(\hat{\sigma}_{12} + zk\hat{\sigma}_1\right) \tag{65}$$

$$1 - \alpha + \alpha G\left(\hat{\sigma}_{12}\right) - G\left(\hat{\sigma}_1\right) = \left(g\left(\hat{\sigma}_1\right) + \alpha zkg\left(\hat{\sigma}_{12}\right)\right)\hat{\sigma}_1. \tag{66}$$

Using A.1, we can express the system above as

$$\begin{aligned} \frac{\hat{\sigma}_{12}}{\nu} &= 1 - zk\frac{\hat{\sigma}_1}{\nu} \\ \frac{\hat{\sigma}_1}{\nu}\left(e^{\frac{\hat{\sigma}_{12}-\hat{\sigma}_1}{\nu}} + \alpha zk\right) &= e^{\frac{\hat{\sigma}_{12}-\hat{\sigma}_1}{\nu}} - \alpha \end{aligned}$$

It is convenient to defined $\Delta \equiv (\hat{\sigma}_{12} - \hat{\sigma}_1)/\nu$ and $x \equiv \frac{\hat{\sigma}_1}{\nu}$, so that we can write the system in (x, Δ):

$$(1 + zk)\, x = 1 - \Delta \tag{67}$$

$$e^{\Delta} - \alpha = \left(e^{\Delta} + \alpha zk\right) x \tag{68}$$

The second equation of the system is $1 - x = \frac{\alpha(1+zk)}{e^{\Delta}+\alpha zk}$. This leads to a schedule x increasing in Δ. The issue is how it changes with α. We study the function on the RHS, namely: $\log\left(\frac{\alpha(1+zk)}{e^{\Delta}+\alpha zk}\right) = \log(\alpha) + \log(1+zk) - \log\left(e^{\Delta} + \alpha zk\right)$. Taking the derivative w.r.t. α

$$\frac{1}{\alpha} + \frac{kz'}{1+zk} - \frac{\alpha kz' + kz}{e^{\Delta} + \alpha kz} = \frac{1}{\alpha} - \frac{1}{\alpha + \frac{e^{\Delta}}{kz}} + kz'\left(\frac{1}{1+kz} - \frac{1}{\frac{e^{\Delta}}{\alpha} + kz}\right)$$

since $\frac{e^{\Delta}}{\alpha} > 1$ we have $\frac{1}{1+kz} - \frac{1}{\frac{e^{\Delta}}{\alpha}+kz} > 0$. Similarly $\frac{1}{\alpha} - \frac{1}{\alpha+\frac{e^{\Delta}}{kz}} > 0$. So $\frac{\alpha(1+kz)}{e^{\Delta}+\alpha zk}$ is increasing in α. Therefore the equilibrium condition $e^{\Delta} - \alpha = \left(e^{\Delta} + \alpha kz\right) x$ implies a schedule x increasing in Δ and decreasing in α. The first equilibrium condition $(1 + zk)\, x = 1 - \Delta$ gives a schedule x decreasing in Δ and decreasing in α. Straightforward analysis then shows that x must be decreasing in α. The free price structure corresponds to $\alpha = 1$, while the protected price structure corresponds to $\alpha = 2a < 1$. Therefore, since $\hat{\sigma}_1 = \nu x$, $\hat{\sigma}_1$ must be higher under price protection.

□

Proof of Proposition 6

To prove that the limited participation distortion is always larger is a consolidated market it suffices to show that $\hat{\sigma}_{con} > \hat{\sigma}_2$. Notice that given $\hat{\sigma}_1 < \hat{\sigma}_2 < \hat{\sigma}_{12}$, a sufficient condition is $\hat{\sigma}_{con} \geq \hat{\sigma}_{12}$. For the single exchange, $\hat{\sigma}_{con}$ satisfies Equation 17

$$\hat{\sigma}_{con} = \frac{1 - G(\hat{\sigma}_{con})}{g(\hat{\sigma}_{con})}. \tag{69}$$

From Equation 59 we have

$$\hat{\sigma}_{12} = \frac{1 - G(\hat{\sigma}_{12})}{g(\hat{\sigma}_{12})} - \hat{\sigma}_1 \frac{s_1}{s_2 - s_1}$$

Since $\hat{\sigma}_1 \frac{s_1}{s_2 - s_1} \geq 0$, the log-concavity of g implies $\hat{\sigma}_{con} \geq \hat{\sigma}_{12}$ and thus

$$\frac{1}{r}\left[\hat{\sigma}_{con} - G^{-1}(1 - 2\overline{a})\right] > \frac{1}{r}\left[\hat{\sigma}_2 - G^{-1}(1 - 2\overline{a})\right].$$

Note that given the pricing expressions in Section III-IV, $p_{con} \geq p_2$ depends on whether

$$\frac{\hat{\sigma}_{con}}{\hat{\sigma}_2} \geq \frac{r + \gamma s_2}{r + \gamma s_{con}}. \tag{70}$$

As market contact frictions become small ($\underline{s}$ approaches one), the RHS of inequality 70 approaches one, and the ability of exchanges to differentiate their services decreases, reducing $\hat{\sigma}_2$. On the other hand, Equation 69 shows that $\hat{\sigma}_{con}$ is unaffected. This implies that inequality 70 holds strictly for a large enough value of $\underline{s}$.

Finally, the inequality $s_2^{seg} > s_{con}$ follows from Proposition 5 in Pagnotta and Philippon (2012), which we provide below as a Lemma for completeness. This fact implies that in the long-run $1 - \lambda_2^{seg} < 1 - \lambda_{con}$ and thus the illiquidity discounts satisfy: $\hat{\sigma}_2^{seg}(1 - \lambda_2^{seg}) < \hat{\sigma}_{con}(1 - \lambda_{con})$.

Lemma 3. *Under A.1* $s_2^{seg} > s_{con}$.

Proof. Under A.1 and with $\alpha = 1$, we have $\hat{\sigma}_{12} = \nu - \frac{s_1}{s_2 - s_1}\hat{\sigma}_1$ and $q_2 = \frac{\nu}{2r}(s_2 - s_1)$. The profits of the fast venue are $\pi_2 = q_2(1 - G(\hat{\sigma}_{12}))$, and therefore

$$\pi_2 = \frac{\nu}{2r}(s_2 - s_1)(1 - G(\hat{\sigma}_{12}))$$

Note that this system is equivalent to the monopoly case when $s_1 = 0$. The FOC for speed is

$$2rC'(s_2) = \nu(1 - G(\hat{\sigma}_{12})) - \nu(s_2 - s_1)g(\hat{\sigma}_{12})\frac{\partial \hat{\sigma}_{12}}{\partial s_2} \tag{71}$$

The consolidated solution is $2rC'(\overline{s}_2) = \nu e^{-1}$. With two active venues we have $\frac{\partial \hat{\sigma}_{12}}{\partial s_2} = \frac{k}{s_2 - s_1}\hat{\sigma}_1 - k\frac{\partial \hat{\sigma}_1}{\partial s_2}$. Then,

$$\begin{aligned} 2rC'(s_2) &= \nu(1 - G(\hat{\sigma}_{12})) - \nu g(\hat{\sigma}_{12})\left[k\hat{\sigma}_1 - s_1\frac{\partial \hat{\sigma}_1}{\partial s_2}\right] \\ &= e^{-\frac{\hat{\sigma}_{12}}{\nu}}\left(\nu - \left[k\hat{\sigma}_1 - s_1\frac{\partial \hat{\sigma}_1}{\partial s_2}\right]\right). \end{aligned}$$

Using $x \equiv \frac{\hat{\sigma}_1}{\nu}, \Delta \equiv \frac{\hat{\sigma}_{12} - \hat{\sigma}_1}{\nu}$

$$2rC'(s_2) = \nu e^{kx-1}\left(1 - kx + s_1\frac{\partial x}{\partial s_2}\right). \tag{72}$$

Since C' is an increasing function, market 2 chooses a higher speed whenever the RHS of 72 is greater

than νe^{-1}. That is,

$$e^{kx}\left(1-kx+s_1\frac{\partial x}{\partial s_2}\right)-1>0. \tag{73}$$

Now we derive $\frac{\partial x}{\partial s_2}$. Differentiating the system 67-68 we have

$$\begin{aligned}(1+k)\,dx+d\Delta-\frac{k}{(s_2-s_1)}ds_2=0\\ \left(e^{\Delta}+k\right)dx+e^{\Delta}(x-1)\,d\Delta-\frac{k}{(s_2-s_1)}ds_2=0.\end{aligned}$$

After appropriate substitutions we get

$$s_1\frac{\partial x}{\partial s_2}=\frac{k^2x\left(1+e^{\Delta}(1-x)\right)}{e^{\Delta}(1+\Delta)+k\left(1+e^{\Delta}\right)}. \tag{74}$$

In order to verify 73, we plug 74 in 73, and define

$$S(k)\equiv e^{kx}\left(1-kx+\frac{k^2x\left(1+e^{\Delta}(1-x)\right)}{e^{\Delta}(1+\Delta)+k\left(1+e^{\Delta}\right)}\right)-1. \tag{75}$$

Re-arranging we have

$$S(k)=e^{kx}\left(\frac{e^{\Delta}(1+\Delta)+k\left(1+e^{\Delta}\right)-kxe^{\Delta}(1+\Delta-kx)}{e^{\Delta}(1+\Delta)+k\left(1+e^{\Delta}\right)}\right)-1. \tag{76}$$

To satisfy the inequality we need $S(k)>0$ for all $k>0$ and $S(0)=0$ (corresponding to the monopolist case where $s_1=0$). Let $x(k)$ and $\Delta(k)$ denote the solutions to the system 67-68 for a given $k\geq 0$. Since $x(k)$ and $\Delta(k)$ are continuous functions, $S(k)$ is continuous. Using 67-68 one can see that $\lim_{k\to\infty}x(k)=0$ and $\lim_{k\to\infty}\Delta(k)=\underline{\Delta}$, where $\underline{\Delta}$ is defined by $e^{\underline{\Delta}}+\underline{\Delta}=2$. Notice that $\lim_{k\to\infty}x(k)k=1-\underline{\Delta}$. Similarly, $\lim_{k\to 0}x(k)=1-\overline{\Delta}$ and $\lim_{k\to 0}\Delta(k)=\overline{\Delta}$, where $\overline{\Delta}$ is defined by $e^{\overline{\Delta}}\overline{\Delta}=1$. Taking limits of 75 we find $\lim_{k\to 0}S(k)=e^0-1=0$ and $\lim_{k\to\infty}S(k)=e^{1-\underline{\Delta}}-1>0$.

A sufficient condition for $S(k)>0$ for all $k>0$ is to show that the term between brackets in 76 is greater than one. This is the case whenever

$$e^{\Delta}(1+\Delta+k)+k+e^{\Delta}k\left[(1-x)+(xk)^2-x\Delta\right]>e^{\Delta}(1+\Delta+k)+k \tag{77}$$

Note from 67 that $1-x=kx+\Delta$. Then,

$$(1-x)+(xk)^2-x\Delta=kx+\Delta(1-x)+(xk)^2>0.$$

We conclude that $S(k)>0$ for all $k>0$.

□

Proof of Proposition 7

To prove the first part of the proposition we need to show that $\hat{\sigma}_{con}\lambda_{con} > \hat{\sigma}_1^{prot}\lambda_1$. Using the system of Equations 65-66, and following the steps of Proposition 6 it is easy to show that $\hat{\sigma}_{con} > \hat{\sigma}_1^{prot}$. Given the value of the default technology $\underline{s}$ we have $\lambda_{con} \geq \lambda_1$, with strict inequality any time that the single exchange invests in speed. We conclude that $p_{con} > p_{nb}$.

Following Proposition 5, under A.1 we have $\hat{\sigma}_1^{seg} < \hat{\sigma}_1^{prot}$ and thus $p_{nb} > p_1$. To prove that $p_{nb} < p_2$ it is sufficient to show that $\sigma_2^{seg} > \sigma_1^{prot}$. From Equation 28 we have that

$$\sigma_2^{seg}\frac{s_2}{s_2 - s_1} = \sigma_{12}^{seg} + \sigma_1^{seg}\frac{s_1}{s_2 - s_1}.$$

The RHS of this expression is according to Equation 59 equal to ν, and thus we have $\frac{\sigma_2^{seg}}{\nu} = \frac{s_2 - s_1}{s_2} = \frac{1}{1+k}$. According to Equation 65 we have $\frac{\hat{\sigma}_1^{prot}}{\nu} = \left(\frac{1-\Delta}{1+\alpha k}\right)$ and thus, provided $\overline{a}$ is large enough, we have $\sigma_2^{seg} > \sigma_1^{prot}$.

Finally note that when frictions are relatively large ($\underline{s}$ low) and/or costs are low, we have $s_2 \approx 1$ and $s_1 \ll s_2$. By Equation 28 this increases $\hat{\sigma}_{12}^{seg}$. The combined effect of these factors is to lower traded volume in market 1 (Equation 35) and to increase volume in market 2 (Equation 36), which makes p_{vw} closer to p_2^{seg} and thus higher than p_{nb}.

□

Proof of Proposition 8

I calculate the comparative statics for the short and long-term consolidated prices. The effect $\frac{\partial p_{con}^S}{\partial \nu} > 0$ is immediate. The expression $\frac{\partial p_{con}^L}{\partial \nu}$ is given by

$$\frac{\partial p_{con}^L}{\partial \nu} = \frac{\partial p_{con}^S}{\partial \nu} + \frac{\partial}{\partial \nu}\left(\frac{r + \gamma s_{con}}{r + \gamma}\right).$$

By Equation 47 we have $\frac{\partial s_{con}}{\partial \nu} > 0$ and thus $\frac{\partial p_{con}^L}{\partial \nu} > \frac{\partial p_{con}^S}{\partial \nu}$.

Note that in a consolidated market the effect of a change in γ only affects the illiquidity discount term. In the short-run price, $\frac{\partial p_{con}^S}{\partial \gamma}$ is given by $\frac{\nu(\underline{s}-1)}{(r+\gamma)^2}$, which is negative since $\underline{s} \leq 1$. In the long-run, the effect is given by

$$\frac{\partial p_{con}^L}{\partial \gamma} = -\left(\frac{ce\nu}{2r}\right)^{\frac{1}{2}}\frac{(2r+\gamma)}{(r+\gamma)^{\frac{3}{2}}},$$

which is clearly negative. It is immediate that the ratio $\left|\frac{\partial p^S_{con}}{\partial \gamma}\right| / \left|\frac{\partial p^L_{con}}{\partial \gamma}\right|$ is greater than one provided

$$1 - s_{con} < \frac{r}{r+\gamma}\left(1 - \underline{s}\right),$$

or equivalently using Equation 47

$$c < \left(\frac{1-\underline{s}}{2r+\gamma}\right)^2 \frac{2r\nu}{(r+\gamma)\, e}.$$

Finally, the cost parameter c effect on the price is given by

$$\frac{\partial p^L_{con}}{\partial c} = -\gamma \left(\frac{e\nu}{2cr\,(r+\gamma)}\right)^{\frac{1}{2}} < 0.$$

□