Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection Lee Kong Chian School Of Business | Lee Kong Chian School of Business

5-2018

Competing on speed

Emiliano Sebastian PAGNOTTA
Singapore Management University, epagnotta@smu.edu.sg

Thomas PHILIPPON

Follow this and additional works at: https://ink.library.smu.edu.sg/lkcsb_research

 Part of the Finance and Financial Management Commons, and the Portfolio and Security Analysis Commons

Citation

PAGNOTTA, Emiliano Sebastian and PHILIPPON, Thomas. Competing on speed. (2018). *Econometrica*. 86, (3), 1067-1115.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/7028

This Journal Article is brought to you for free and open access by the Lee Kong Chian School of Business at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection Lee Kong Chian School Of Business by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Check for updates

Econometrica, Vol. 86, No. 3 (May, 2018), 1067–1115

COMPETING ON SPEED

EMILIANO S. PAGNOTTA
Department of Finance, Imperial College Business School

THOMAS PHILIPPON
Department of Finance, New York University Stern School of Business, National Bureau of Economic Research, and Centre for Economic Policy Research

We analyze trading speed and fragmentation in asset markets. In our model, trading venues make technological investments and compete for investors who choose where and how much to trade. Faster venues charge higher fees and attract speed-sensitive investors. Competition among venues increases investor participation, trading volume, and allocative efficiency, but entry and fragmentation can be excessive, and speeds are generically inefficient. Regulations that protect transaction prices (e.g., Securities and Exchange Commission trade-through rule) lead to greater fragmentation. Our model sheds light on the experience of European and U.S. markets since the implementation of Markets in Financial Instruments Directive and Regulation National Markets System.

KEYWORDS: Trading speed, exchanges, liquidity, fragmentation, segmentation, vertical differentiation, search, high-frequency trading, regulation, trade-through rule, investor participation, entry.

1. INTRODUCTION

THE SECURITIES EXCHANGE INDUSTRY has changed deeply over the past decade. Entry of new venues has led to fragmentation of trading, particularly in the United States and in Europe. Trading speed has increased a lot in some markets (equities and standardized derivatives in particular), but much trading still relies on human inputs. As a result, we now observe significant heterogeneity in trading across venues and asset classes. These evolutions have triggered heated debates in academic and policy circles. Why do venues compete on speed? Is there a connection between speed and fragmentation? What are the welfare consequences of these changes? What are the appropriate regulations? To shed light on these issues, we propose a model of the market for markets, that is, we analyze

Emiliano S. Pagnotta: e.pagnotta@imperial.ac.uk
Thomas Philippon: tphilipp@stern.nyu.edu

We are particularly grateful to Joel Hasbrouck for his insights and to Jonathan Brogaard, Thierry Foucault, Albert Menkveld, Guillaume Rocheteau, Pierre-Oliver Weill (discussants). We thank Yakov Amihud, Ivan Canay, Darrell Duffie, Xavier Gabaix, Ricardo Lagos, Lasse Pedersen, Marti Subrahmanyam, Dimitri Vayanos, Mao Ye, and seminar participants at HEC Paris, the NYU Stern School of Business, the London School of Economics, the Toulouse School of Economics, Rochester University, the Tinbergen Institute, the University of Illinois at Chicago, Imperial College Business School, the University of Lugano, the University of Amsterdam, IESE Business School, the University of San Andrés, the Federal Reserve Bank of New York, the Bank of England, and seminar participant at the following conferences: the 2011 Society of Economic Dynamics, the Fourth Annual Conference on Money, Banking and Asset Markets at University of Wisconsin-Madison, the Third Annual Conference of Advances in Macro Finance Tepper-LAEF, the Western Finance Association (Las Vegas), the Finance Theory Group Meeting (Harvard Business School), the Cowles Foundation GE Conference (Yale University), the Econometric Society NA Meetings (Evanston), the 2012 Financial Management Association Napa Conference, the Fourth Annual Hedge Fund Euronext Conference, the 2013 Conference of the Paul Woolley Centre for the Study of Capital Market Disfunctionality (London School of Economics), the 2014 CAFIN Conference, and the 2017 American Economic Association Meetings (Chicago). We are grateful to market participants at Citibank, Société Générale, and the TABB Group for their feedback. We acknowledge the support of the Smith Richardson Foundation.

© 2018 The Econometric Society https://doi.org/10.3982/ECTA10762

competition among trading venues offering differentiated trading services. For simplicity, we refer broadly to the quality of these services as *speed*, by which we mean a feature that reduces the time between the occurrence of a desire to trade and the execution of the trade.[1]

Our analysis requires modeling four distinct elements: (i) why and how investors value speed; (ii) how differences in speed affect competition among trading venues and the affiliation choices of investors; (iii) how trading regulations affect (i) and (ii); and (iv) how these choices affect investment in speed and equilibrium fragmentation. These requirements explain our modeling choices and the structure of our paper. We consider a dynamic infinite-horizon model where investors buy and sell a single security. Gains from trade arise from random shocks to the marginal utility (or marginal cost) of holding the asset.[2] High-marginal-utility investors are natural buyers, while low-marginal-utility investors are natural sellers of the asset. Higher speed allows investors to realize a larger fraction of the potential gains from trade. Investors differ in the volatility of their marginal utility process, and thus in their gains from trade and their demand for speed (Proposition 1).

Given the structure of the demand for speed, we can then analyze the supply of trading services as a sequential game. Venues first decide whether to enter or not (entry game), then invest in trading technologies (speed choices), and finally compete on fees to attract investors (affiliation game). In the equilibrium of the affiliation game, we show that faster venues charge higher fees and attract speed-sensitive investors, and that competition leads to lower fees and greater investor participation (Proposition 4).

We then turn to speed choices and entry. The key point is that choosing different speeds allows the venues to offer vertically differentiated products. We find that speed choices are inefficient because differentiation relaxes price competition and because venues do not internalize the welfare gains of infra-marginal investors (Propositions 5 and 6). In this context, we show that a regulator would find it optimal to impose a minimum speed requirement, but not a maximum speed limit (Proposition 7). Finally, in the entry game, we highlight the tension between business stealing, competition, and product diversity. Regardless of fixed entry costs, as long as the cost of speed is not too high, we show that entry by a second venue always enhances welfare (Proposition 8). However, excess entry is possible in a general oligopoly and we study one such extension.[3]

A distinct contribution of our paper is the equilibrium analysis of price fragmentation in a multi-venue market. We consider two polar cases. In the *segmented* case, a venue only executes the orders of its investors and trades occur at different prices in different venues. In the *integrated* case, there is a unique price and venues offer different "gates of entry" to

[1]Our notion of speed is broad and includes not only communication latencies, but also various technological innovations that make trading more convenient and more reliable, such as user-friendly software or data feeds and reliable hardware. In addition to pure physical speed, most traders emphasize convenience and reliability as important features for a trading platform. All these factors affect the *total expected* time and effort between the decision to trade and the execution of the trade. When we use the term *speed*, it is with this broad interpretation in mind.

[2]As is well understood in the literature, these shocks can capture liquidity demand (i.e., the need for cash), financing costs, hedging demand, portfolio rebalancing, or any other personal use of assets, including specific arbitrage opportunities (for a discussion, see Duffie, Garleanu, and Pedersen (2007)). The important point is that these shocks affect the private value of an asset, not its common value. The shocks, therefore, generate gains from trade.

[3]Diversity is beneficial in models of horizontal differentiation but not necessarily so in models of vertical differentiation such as ours. The familiar excess entry theorem of Mankiw and Whinston (1986) cannot be used in our environment.

a single asset market. We view an order price protection rule—such as the trade-through rule in Regulation National Market System (Reg NMS; see Section 2)—as a move from segmentation to integration. Such regulations, we find, affect the gains from trade and therefore have an impact on all the stages of the model: affiliation, speed, and entry. In the affiliation game, protection acts as a subsidy to the slow venue because its investors enjoy interacting with investors from the fast venue who are eager to trade. The slow venue, therefore, charges higher fees and enjoys higher profits under protection. This fact then encourages entry and increases trade volume fragmentation. We thus find that its welfare consequences crucially depend on the ability to impact entry decisions. We show that this occurs for a range of economies with intermediate entry costs (Proposition 8). To the best of our knowledge, this is the first formal analysis of this issue.

We provide a comprehensive calibration of the model for three asset classes associated with different trading speeds: corporate bonds, individual stocks, and equity index futures. Our primary benchmark for welfare comparisons is a planner subject to the same technological constraints as private venues. The calibration also allows us to analyze the welfare consequences of various regulations. We show that lower technological costs can dramatically increase trading speed and volume, but the associated welfare gains are small. However, welfare gains from enforcing a minimum speed can be significant. We can decompose welfare losses from lack of investor participation, misallocation of investors among venues, and misallocation of assets among investors. Interestingly, we find that, because venues differentiate, welfare losses can be significant even when the fast venues trade at extreme speeds. As a result, a secular reduction in speed costs never makes the economy converge to the frictionless outcome.

The recent sharp increase in market fragmentation in developed countries has encouraged a new wave of empirical studies whose results appear to be consistent with the predictions of our model. Foucault and Menkveld (2008), O'Hara and Ye (2011), and Degryse, De Jong, and van Kervel (2015), among others, found that an increase in trading fragmentation is associated with lower costs and faster execution speeds in a given asset class. Our result that price integration increases entry and thus fragmentation helps to rationalize (i) the sharp increase in fragmentation experienced in U.S. equity markets following Reg NMS in 2007 and (ii) the fact that fragmentation levels in the United States are among the highest in the world. The normative analysis, on the other hand, suggests limited welfare gains from purely technological improvements in fragmented markets and highlights the importance of sound regulations. We discuss further implications for regulators in Section 9.

Our paper relates to several strands of the literature in economics and finance. In the industrial organization literature, Gabszewicz and Thisse (1979) and Shaked and Sutton (1982, 1983) pioneered the analysis of vertically differentiated oligopolies. Our framework enriches the classical two-stage competition environment by endogenizing preferences for product quality (trading delays here) through a micro-founded model of dynamic trading. Because we model a market for markets, we show how the degree of price fragmentation in the downstream market (the asset market) affects the allocation of investors in the upstream market (that for venue services). This approach allows one to study how changes in trading protocols, fees, or regulations affect features of the industrial organization like the number of active venues.

Early theoretical analyses of fragmentation include those of Mendelson (1987) and Pagano (1989). These static models focus on the tradeoff between liquidity externalities, market power, and trading costs. This tension was of the first order of importance when different marketplaces were not as integrated as they are nowadays. Venues can differentiate in areas other than speed. For example, Santos and Scheinkman (2001) studied

competition in margin requirements, and Foucault and Parlour (2004) and Chao, Yao, and Ye (2017) studied competition in listing and make-take fees, respectively. These papers consider static frameworks and do not analyze speed differentiation. By contrast, we develop a dynamic model where speed plays an explicit role. We also provide the first equilibrium analysis of price protection.

Our trading model builds on the recent literature that models dynamic trading with friction, spurred by Duffie, Garleanu, and Pedersen (2005), and is closest to that of Lagos and Rocheteau (2009, LR09 hereafter).[4] We follow these models in that private valuations change randomly. We do not, however, encompass all trading mechanisms. In contrast to Duffie, Garleanu, and Pedersen (2005), we do not model decentralized OTC trades through random search, a specific matching function, and a bargaining game. For tractability and cleanness of analysis, we adopt a Walrasian clearing protocol that is subject to a random delay in the execution of a trade. A distinctive feature of our model is that the distribution of such delays across venues arises endogenously and is explicitly affected by the competition and regulatory environment. To the best of our knowledge, this paper offers the first model of a market for markets, with a joint determination of trading and market structure. We are also the first to propose a complete calibration in such environment. The asset pricing implications were studied by Pagnotta (2014).

Our work complements the recent literature that analyzes HFT (e.g., Aït-Sahalia and Saglam (2013), Budish, Cramton, and Shim (2015), Foucault, Hombert, and Rosu (2016), Biais, Foucault, and Moinas (2015)). The literature models speed-related advantages by introducing a form of short-lived asymmetric information that stems from stale prices when the common value component of the asset changes (e.g., news arrival). Naturally, one can argue that the desire to take advantage of information is one reason behind the observed increase in speed, and thus it would be interesting to extend the model to accommodate for this possibility. We do not analyze asymmetric information directly, but we provide a "macro" building block where positive and normative issues related to investors with different speed capacities can be examined. As an illustration, Budish, Cramton, and Shim (2015) argued that moving away from continuous trading toward periodic auctions would eliminate the speed investment frenzy. Our results suggest that these reforms could mitigate but are unlikely to stop this phenomenon. First, as the cost of speed decreases, venues' speeds increase and become more differentiated (Section 8.4). Second, lower costs may encourage the entry of faster venues that enjoy higher profits, as in the three-venue example in Section 7.2, increasing the average trade speed in the market, even without order front-running.[5]

Furthermore, we note that the joint increase in speed differentiation and fragmentation is not a phenomenon that began with HFT (see Section 2) and it has been observed in virtually every asset class, information-sensitive ones or otherwise.[6] Current models of front-running in a single venue can explain why some traders have an incentive to become *individually* faster, but they do not account for the distribution of investors across trading venues with different speeds. Nothing prevents the formation of a relatively slow

[4] Weill (2007) used a related framework to analyze market making in exchanges. Vayanos and Wang (2007) and Weill (2008) studied the concentration of liquidity across assets instead of venues. Many additional contributions were surveyed by Lagos, Rocheteau, and Wright (2017).

[5] In a sequential entry-exit setting, entry of a faster venue may require the simultaneous exit of the slowest one, in an analogous fashion to Gabszewicz and Thisse (1980), further increasing the market average speed.

[6] Even investors who are not interested in any front-running still decide to trade in fast exchanges, such as IEX, or fast venues with Request For Quote (RFQ) protocols. The popularity of exchange-traded funds and liquid index tracking instruments has likely strengthened this trend.

and cheap venue. If uninformed traders choose to join fast venues, they must value speed; otherwise, they would all join the slow venue, depriving the fast venue of liquidity. The idea that speed is provided exclusively to satisfy a fraction of informed traders seems to be inconsistent with free entry. Our model, on the other hand, captures a *fundamental part of the demand for speed*. Speed-sensitive gains from trade are required to rationalize the interplay of venue and investor choices. We certainly do not claim that asymmetric information is irrelevant, but we do argue that the building blocks of our model are required to analyze speed, fragmentation, and welfare, with or without asymmetric information.

2. SECURITIES TRADING: MOTIVATING FACTS AND TRENDS

Our model seeks to explain how changes in technology and market organization affect the ease and speed of trading. Here we provide a brief overview of these evolutions, from the telegraph in the 19th century to recent ultra-low latencies systems. We provide more details in Appendix A.

Trading Speed

Garbade and Silber (1977) studied two early developments of market infrastructure: the telegraph connecting New York with other American market centers in the 1840s, and a trans-Atlantic cable connecting New York and London in 1866.[7] Early in the 1900s, all European stock markets (except London) conducted periodic auctions, once or several times a day. The progressive—although not simultaneous—adoption of continuous trading represented a massive increase in trading frequencies.[8] The diffusion of personal computers in the 1980s enabled electronic trading and the development of information systems, such as Bloomberg terminals. The crash of 1987 and subsequent regulation reforms pushed exchanges to adopt automatic execution systems (like the Small Order Execution System) that did not rely on traditional floor brokers (Lewis (2014)). These historical examples highlight the interactions between technology, competition, and market structure that are at the heart of our model.

Figure 1 summarizes the current trading landscape. Over the past ten years, market centers have made costly investments in trading infrastructure to reduce order execution and communication latencies (Table A.I lists several recent examples). This process has gone beyond equities and futures to reach options, bonds, and currencies. We want to emphasize two important stylized facts. First, trading speeds vary widely across markets, and much trading still relies on human input. For instance, electronic trading covered only 21% of the corporate bond market in 2014, while voice trading covered the remaining 79%. High-frequency trading (HFT) is not the norm in most markets. Second, it is important to distinguish the speed of quote updating from the speed of trades. According to a recent SEC study (see Figure A.2), more than half of fully executed orders (and 60% of partial trades) take place between 5 seconds and 10 minutes. The blazing fast speed advertised by many trading venues corresponds to quote revisions, not trades, and it is trading speed that matters in our model.

[7] They argued that these two innovations accelerated the search for liquidity in financial markets and significantly reduced order execution delays. In the context of our model, such developments also represent a move toward integration between markets that were previously segmented.

[8] Floor brokers in traditional continuous-time exchanges such as the NYSE enjoyed advantages in trading speed compared to off-floor investors. The cost of participating in the exchange floor is a type of speed-related fee (similar to q in our model).

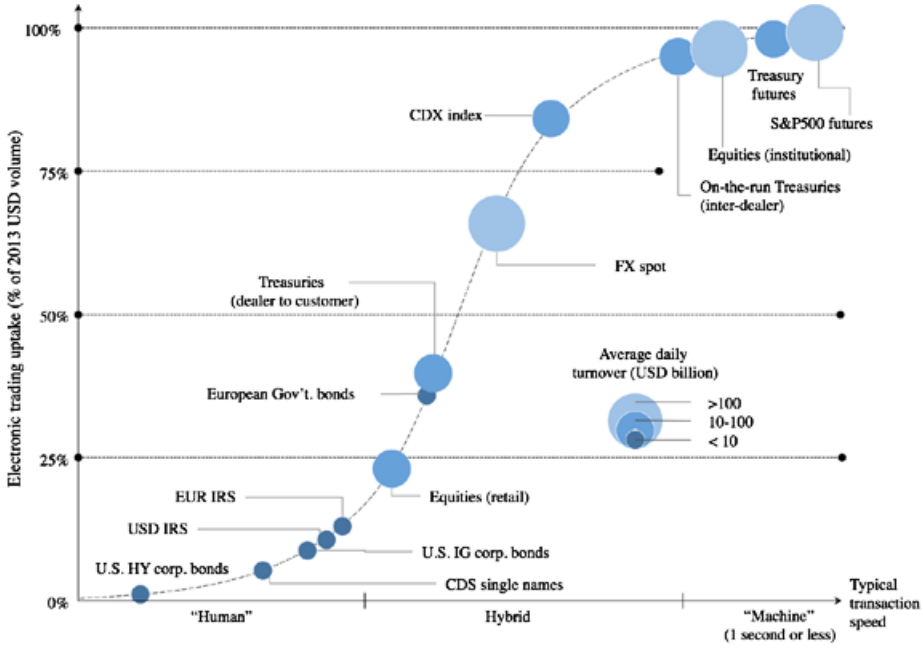

FIGURE 1.—Assets classes and trading speeds. Source: TABB Group and various sources.

Fragmentation and Differentiation

The second major feature of the current trading landscape is fragmentation (see Figure A.1). Traditional markets such as the London Stock Exchange (right panel) have lost market share to faster entrants such as Chi-X. The left panel shows an even more dramatic evolution: The fraction of NYSE-listed stocks traded on the NYSE decreased from 80% in 2004 to just over 20% in 2009. Most of the lost trading volume has been captured by new entrants (e.g., Direct Edge and BATS).[9] The link between entry and speed investments is also apparent in regions like Asia, Australia, and Latin America, where traditional trading venues have faced the threat of alternative trading platforms. Table A.I in Appendix A illustrates the global character of this phenomenon. Overall, fragmentation has become so prevalent that market participants keep track of *fragmentation indexes* across asset classes and countries (e.g., Fidessa's indices).

Besides the secular increase in average trading speed, we observe a lot of *differentiation* in every major asset class. Our model emphasizes the speed-related choices that investors must make and Table I presents some examples. In equity markets, investors might need to choose between two exchanges, such as the NASDAQ versus the NYSE in the United States, or ASX versus Chi-X in Australia.[10] A second, broader, interpretation is that investors sort themselves between "exchanges" and a range of "alternative trading venues." According to the SEC classification, U.S. investors can opt to direct their

[9]We focus here on the European and U.S. experiences, but our analysis and results apply to other recent international cases.

[10]Boehmer (2005) documented the tradeoff between execution speed and costs in U.S. markets before Reg NMS. He found that, analogously to venue 2 in our model, the NASDAQ is more expensive than the NYSE, but it is also faster. More recent data show that the NASDAQ was still significantly faster than the NYSE at the time of Reg NMS implementation in 2007 (Angel, Harris, and Spatt (2015)).

TABLE I

VENUE SPEED CHOICES IN ASSET MARKETS: EXAMPLES

Market	Slow Venues	Fast Venues
Equities (institutional)	Crossing networks, floor exchanges	Direct access to lit exchange, co-location
Equities (retail)	Retail bank (mutual funds)	Premium broker (ETF, index futures, etc.)
Foreign exchange (FX)	OTC dealer/bank (voice)	Currenex, EBS, Reuters
Corporate bonds	OTC voice trading	Aladdin, Tradeweb, Bonds.com, Liquidnet, NYSE Bonds, BrokerTec
Interest rate swaps (IRS)	OTC dealer/bank	SEFs. ICAP, BCG, Tradition
Credit swaps (CDS)	OTC dealer/bank	SEFs. Bloomberg, GFI, MarketAxxes

orders to registered exchanges, and a range of Alternative Trading Systems (ATS) that include Electronic Communication Networks (ECN), dark pools, broker/dealer Internalizers, and Crossing Networks.[11] Over-the-counter venues have made technical progress, but, as a group, organized exchanges typically offer investors the fastest communication and trading responses. To summarize, we can group broker-dealers/crossing networks and floor-driven exchanges as slow venues, and (lit) electronic exchanges as fast venues. We discuss additional asset classes in Appendix A.2.

Regulation of Entry and Price Protection Rules

Market regulators have not been passive witnesses to these evolutions. U.S. policymakers have encouraged fragmentation to reduce the market power of trading venues, prominently with the Regulation of Exchange and Alternative Trading Systems (Reg ATS) and Regulation National Market System (Reg NMS).[12] Encouraged by this experience, other economies started promoting competition between market centers. In Europe, for example, the Markets in Financial Instruments Directive (MiFID) transformed the trading landscape. Large-cap stocks that previously traded in one or two venues are now traded in almost 50 venues, including internalization pools and over-the-counter (OTC) venues.

Concerns about adverse effects of *price fragmentation*, in turn, motivated regulators to adopt investor protection rules that regulate order execution prices. Table A.II lists several examples. Under one approach, the *trade-through model*, market centers' systems are connected to one another, and they prevent trading through better prices available elsewhere. Price is thus the primary criterion for best execution. Such approach requires a complex and costly infrastructure as well as strong monitoring activity by market regulators. In the United States, Rule 611 of Reg NMS requires that venues execute their trades at the national best bid and offer quotes, thereby consolidating prices from scattered trading. In Canada, the Order Protection Rule (OPR) implemented by the Investment Industry Regulatory Organization shares the same spirit but aims to protect orders beyond the

[11] According to the SEC, these alternative venues jointly represent 33–36% of U.S. equity volume. Similarly, European regulators make a distinction between Regulated Markets, Multilateral Trading Facilities (MTFs), and Systematic Internalizers.

[12] For example, the U.S. Securities and Exchange Commission (SEC (2010)) states:

> 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 trading fees low.

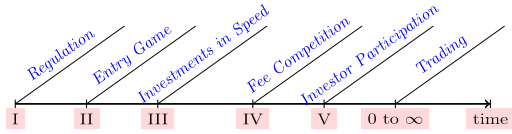

FIGURE 2.—Timing and structure of the model.

top level of the order book. Under a second approach, the *principles-based model*, criteria other than prices can be added to the best execution policy, such as investor type, listing exchange, liquidity, execution probability, and speed. Although this approach provides less transparency, it requires a simpler set of linkages between markets and may promote innovation by not enforcing uniformity. In Japan, 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." MiFID introduced a "transparency" regime in Europe, but no formal trade-through rule for which venues are held responsible. Both in Europe and Japan, sell-side firms are not required to monitor every active venue. Such a task is, rather, left to their clients.[13]

An additional type of market intervention that we study is the direct regulation of trading speed. We briefly discuss some examples of such policies in Section 6.

3. TRADING MODEL

The market structure determination sequence is depicted in Figure 2 and discussed in Section 4. This section analyzes the trading stage and equilibrium in one venue, taking as given venue choices and participation decisions. The key result of this section is a characterization of traders' value functions, providing an explicit micro foundation of how investors value speed in financial markets.

3.1. *Preferences, Technology, and Trading Equilibrium*

We start by describing the main building blocks of our model: investor preferences and trading technology. Preferences need to incorporate heterogeneity to create gains from trade as well as interesting participation decisions among venues. The trading technology must capture the role of speed in financial markets. Time is continuous and there is a continuum of heterogeneous investors, two goods, and one asset. The measure of investors is normalized to 1, and their preferences are quasi-linear. The numéraire good (cash) has a constant marginal utility normalized to 1 and can be freely invested at the constant rate of return r. The asset is in fixed supply, $\bar{a}$, which is also the (expected) endowment of each investor.[14] We restrict asset holdings to $a_t \in [0, 1]$. One unit of asset pays a fixed dividend equal to μ 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

$$u_{\sigma,\varepsilon_t}(a_t) = (\mu + \sigma\varepsilon_t)a_t,$$

[13] Arbitrageurs and smart routing technologies often work to undo price differentials between markets. However, pre MiFID I empirical evidence by Foucault and Menkveld (2008) suggests that this fact does not make price protection rules redundant as a significant proportion of trade-throughs remain even where there are no entry barriers to arbitrageurs.

[14] There are two interpretations: Either agents receive $\bar{a}$ or they own one unit with probability $\bar{a}$. Since all agents are risk neutral, the two interpretations are equivalent.

where (σ, ε_t) denotes the type of investor. The fixed component $\sigma \in [0, \bar{\sigma})$ is known at time 0 and distributed according to the twice-differentiable cumulative distribution G, with a log-concave density function g that is positive everywhere. The time-varying component $\varepsilon_t \in \{-1, +1\}$ changes randomly, and inter-arrival times between changes are distributed exponentially with parameter γ. Conditional on a change, ε is i.i.d. and each value has equal probability. As explained in the Introduction, ε may capture several sources of private value shocks such as time-varying liquidity demands or specific investment opportunities. The parameter σ then measures the size of these shocks.

The venue where investors trade the asset is characterized by the contact rate ρ (i.e., the average stochastic trading delay is ρ^{-1}). Conditional on being in contact, the market is Walrasian and clears at price p. That is, any trader in contact with the venue at time t can trade at the price p_t. Traders who are not in contact simply keep their holdings constant. Our assumptions about technology and preferences imply that the value function of a class-σ trader with current valuation ε_s and current asset holdings a at time t is

$$V_{\sigma,\varepsilon_t}(a,t) = \mathbb{E}_t\left[\int_t^T e^{-r(s-t)} u_{\sigma,\varepsilon_s}(a)\, ds + e^{-r(T-t)}\big(V_{\sigma,\varepsilon_T}(a_T, T) - p_T(a_T - a)\big)\right], \quad (1)$$

where the realization of the random type at time $s > t$ is ε_s and T denotes the next time the investor makes contact with the venue. Expectations are defined over the random variables T and ε_s and are conditional on the current type ε_t. We can show that the asset price remains constant during the trading game. The value functions are thus time independent. Letting $a^*_{\sigma,\varepsilon}$ denote the optimal choice of asset holding for type (σ, ε), equation (1) becomes simply

$$rV_{\sigma\varepsilon}(a) = u_{\sigma,\varepsilon}(a) + \frac{\gamma}{2}\sum_{\varepsilon'}\big[V_{\sigma\varepsilon'}(a) - V_{\sigma\varepsilon}(a)\big] + \rho\big[V_{\sigma\varepsilon}\big(a^*_{\sigma,\varepsilon}\big) - V_{\sigma\varepsilon}(a) - p\big(a^*_{\sigma,\varepsilon} - a\big)\big].$$

Following LR09, we define the adjusted holding utility as $\bar{u}(a; \sigma, \varepsilon) \equiv \frac{(r+\rho)u_{\sigma,\varepsilon}(a) + \gamma\mathbb{E}[u_{\sigma,\varepsilon'}(a)|\varepsilon]}{r+\rho+\gamma}$. Note that since ε is i.i.d. with mean zero, $\mathbb{E}[u_{\sigma,\varepsilon'}(a) \mid \varepsilon] = \mu a$ for any a and any ε. This expected utility over ε' does not depend on σ or ε. This result implies that

$$\bar{u}(a; \sigma, \varepsilon) = \left(\mu + \sigma\varepsilon\frac{r+\rho}{r+\rho+\gamma}\right)a. \quad (2)$$

We can then extend Proposition 1 of LR09 to take into account heterogeneity in σ. The equilibrium with constant price is characterized by the demand functions $a^*(p; \sigma, \varepsilon) = \arg\max_a\{\bar{u}(a; \sigma, \varepsilon) - rpa\}$ and the market clearing condition in flows is

$$\int_\sigma \rho \sum_{\varepsilon=\pm 1} \frac{a^*(p; \sigma, \varepsilon)}{2}\, dG(\sigma) = \rho\bar{a}. \quad (3)$$

On the right-hand side of equation (3), the asset supply (per capita) is $\bar{a}$ and a fraction ρ per unit of time is available for trading. Similarly, on the left-hand-side, we have the flow of demand. Note that the same ρ appears on both sides of the equation and we could therefore drop it. In this case with one venue, market clearing in stock is enough to ensure market clearing in flows.

There is symmetry around $\bar{a} = \frac{1}{2}$ since half the investors are of trading type $\varepsilon = 1$ and half are of type $\varepsilon = -1$. It is therefore sufficient to analyze a market where $\bar{a} \leq \frac{1}{2}$. In this case, supply is short, and low-σ types always sell their entire holdings when they contact the venue. Moreover, there is a *marginal trading type*, σ^{t}, that is indifferent between buying and not buying when $\varepsilon = 1$ and is given by

$$\sigma^{\mathrm{t}}(p, \rho) \equiv \frac{r + \rho + \gamma}{r + \rho}(rp - \mu). \tag{4}$$

The demand function is therefore $a^* = 1$ when $\varepsilon = +1$ and $\sigma \geq \sigma^{\mathrm{t}}$. It is $a^* = 0$ in all other cases. We can use these demand curves to rewrite the market clearing condition. All negative trading types $\varepsilon = -1$ want to hold $a = 0$ and they represent half of the traders. The trading types $\varepsilon = +1$ want to hold one unit if $\sigma > \sigma^{\mathrm{t}}$ and nothing if $\sigma < \sigma^{\mathrm{t}}$. The demand for the asset is $\frac{1}{2}(G(\bar{\sigma}) - G(\sigma^{\mathrm{t}}))$. The ex ante supply of the asset (per capita) is $\bar{a}$. The market clearing condition is therefore

$$\frac{1 - G(\sigma^{\mathrm{t}})}{2} = \bar{a}. \tag{5}$$

Note that the asset holdings of types $\sigma < \sigma^{\mathrm{t}}$ are non-stationary since they never purchase the asset. They sell their holding $\bar{a}$ on the first contact with the venue and never trade again. The fact that they stop trading, as opposed to trading repeatedly in smaller quantities, is just a consequence of linear preferences. We refer to traders with $\sigma < \sigma^{\mathrm{t}}$ as *light traders* and traders with $\sigma \geq \sigma^{\mathrm{t}}$ as *heavy traders*, implicitly linking to how 'heavy' their trading volume is.

Over time, the assets move from the low-σ to the high-σ types and then keep circulating among the 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. The gross supply of assets is always $\rho\bar{a}$. The gross demand from high-σ types is always $\rho(1 - G(\sigma^{\mathrm{t}}))/2$. From equation (5), the market always clears.[15]

We can now characterize the steady-state distribution among types $\sigma > \sigma^{\mathrm{t}}$. Let $\alpha_{\sigma,\varepsilon}(a)$ be the share of class-σ investors with trading type ε currently holding a units of asset. Consider first a type $(\varepsilon = +1, a = 1)$. This type is satisfied with its current holding and does not trade even if it contacts the venue. Outflows result only from changes of ε from $+1$ to -1, which occurs with intensity $\gamma/2$. There are two sources of inflow: types $(\varepsilon = -1, a = 1)$ that switch to $\varepsilon = 1$ and types $(\varepsilon = +1, a = 0)$ that purchase one unit when they contact the venue. In steady state, outflows must equal inflows:

$$\frac{\gamma}{2}\alpha_{\sigma,+}(1) = \frac{\gamma}{2}\alpha_{\sigma,-}(1) + \rho\alpha_{\sigma,+}(0). \tag{6}$$

The dynamics for types $(\varepsilon = -1, a = 0)$ are similar: $\frac{\gamma}{2}\alpha_{\sigma,-}(0) = \rho\alpha_{\sigma,-}(1) + \frac{\gamma}{2}\alpha_{\sigma,+}(0)$. For types $(\varepsilon = +1, a = 0)$ and $(\varepsilon = -1, a = 1)$, trade creates outflows, yielding $(\frac{\gamma}{2} + \rho)\alpha_{\sigma,+}(0) = \frac{\gamma}{2}\alpha_{\sigma,-}(0)$ and $(\frac{\gamma}{2} + \rho)\alpha_{\sigma,-}(1) = \frac{\gamma}{2}\alpha_{\sigma,+}(1)$. Finally, the shares must add up to 1: $\sum_{\varepsilon=\pm, a=0,1} \alpha_{\sigma,\varepsilon}(a) = 1$.

[15] In the case $\bar{a} = \frac{1}{2}$, the marginal type is not well defined and a range of prices can clear the market. More precisely, if $\sigma_{\min}$ is the lowest type in the market, then any price $p \in [\frac{\mu}{r} - \frac{\sigma_{\min}}{r}\frac{r+\rho}{r+\rho+\gamma}, \frac{\mu}{r} + \frac{\sigma_{\min}}{r}\frac{r+\rho}{r+\rho+\gamma}]$ is a market clearing price.

To summarize, the trading equilibrium is characterized by the price p and marginal trading type σ^{t} defined in equations (4) and (5), respectively. Light traders sell their initial holdings $\bar{a}$ and do not purchase the asset again. Heavy traders buy when $\varepsilon = 1$ and sell when $\varepsilon = -1$. The distribution of holdings among the latter converges to the steady-state distribution of well-allocated assets $\alpha_{\sigma,+}(1) = \alpha_{\sigma,-}(0) = \frac{1}{4}\frac{2\rho+\gamma}{\gamma+\rho}$ and misallocated assets $\alpha_{\sigma,+}(0) = \alpha_{\sigma,-}(1) = \frac{1}{4}\frac{\gamma}{\gamma+\rho}$.

We can formally define the instantaneous trade volume rate, $\mathcal{V}$, which in the steady state is given by

$$\mathcal{V} = \frac{\rho}{2}\big(\alpha_{\sigma,+}(0) + \alpha_{\sigma,-}(1)\big) \times \big(1 - G\big(\sigma^{\mathrm{t}}\big)\big). \tag{7}$$

The right-hand side of equation (7) is given by the product of the contact rate, the proportion of agents with misallocated assets, and the population of steady-state traders.

3.2. *Value Functions*

Our goal is to analyze the provision of speed in financial markets. We therefore need to estimate the value that investors attach to trading in each venue. We proceed in two steps. We first compute the steady-state value functions for traders. We then compute the ex ante values, taking into account the transition dynamics. Note that the no-trade outside option of any investor, W_{out}, only depends on her endowment $\overline{a}$ and the discounted value of her expected flow utility, μ/r. Therefore, $W_{\mathrm{out}} = \overline{a}\frac{\mu}{r}$. Consider the steady-state value functions for types $\sigma > \sigma^{\mathrm{t}}$. For the types holding the assets, we have

$$rV_{\sigma,-}(1) = \mu - \sigma + \frac{\gamma}{2}\big[V_{\sigma,+}(1) - V_{\sigma,-}(1)\big] + \rho\big(p + V_{\sigma,-}(0) - V_{\sigma,-}(1)\big),$$

and $rV_{\sigma,+}(1) = \mu + \sigma + \frac{\gamma}{2}[V_{\sigma,-}(1) - V_{\sigma,+}(1)]$. Analogous expressions hold for the value functions of the types not holding the assets, forming a system of equations that can be solved to compute their explicit form. The following proposition characterizes the ex ante value functions, that is, those of a trader that knows her permanent type σ but not the temporary preference shock, taking into account the transition dynamics leading up to the steady-state allocations.

PROPOSITION 1: *The ex ante value W for type σ of participating in a venue with speed ρ is the sum of the value of ownership and the value of trading*:

$$W\big(\sigma, \sigma^{\mathrm{t}}, s\big) - W_{\mathrm{out}} = \frac{s\sigma^{\mathrm{t}}}{r}\bar{a} + \frac{s}{2r}\max\big(0; \sigma - \sigma^{\mathrm{t}}\big), \tag{8}$$

where the marginal trading type σ^{t}, defined in equation (4), *increases in p and decreases in ρ and where effective speed s is defined by*

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

Proposition 1 provides the building block for our analysis of the industrial organization of financial markets. The net value of participation, $W - W_{\mathrm{out}}$, is composed of two parts. One is the option to sell the asset on the exchange: $\frac{s\bar{a}\sigma^{\mathrm{t}}}{r} = \frac{\rho}{r+\rho}\big(p - \frac{\mu}{r}\big)\bar{a}$. It is independent of σ and is the value that can be achieved by all light traders. The term $\frac{\rho}{r+\rho}$ is the discount

due to expected trading delays. The second part, $\frac{s}{2r}\max(0;\sigma-\sigma^{\mathrm{t}})$, is the value of trading repeatedly and it depends on the type σ. Importantly, this part of the value function is super-modular in (s,σ) and thus induces sorting of high-σ types into venues offering higher speeds.

4. MARKET STRUCTURE EQUILIBRIUM AND WELFARE

The market structure is determined as a sequential game where, taking regulations as a given, venues decide whether to enter, select trading speeds, and post membership fees. Venues make these decisions in Stages II to IV in Figure 2. We introduce a fixed entry cost κ to analyze the entry game in Stage II. Venues face the same increasing and convex speed investment cost function $C(s)$ in Stage III. Venues compete in fees à la Bertrand in Stage IV. Let q_i be the membership fee posted by venue i and let n_i be the number of investors who join venue i. The total net profits of venue i are therefore $q_i n_i - C(s_i) - \kappa$. Given venues' decisions, investors decide which venue to join in Stage V. Participation decisions are described by a mapping $\mathcal{P}$ from types σ to active venues, $[0,\bar{\sigma}]\mapsto\{0,1,\ldots,I\}$, where $\mathcal{P}(\sigma)=i$ means joining venue i and $\mathcal{P}(\sigma)=0$ means staying out. If an investor joins venue i, it pays a membership fee q_i and is then allowed to use the trading venue. Staying out costs nothing: $q_0=0$ and $W=W_{\mathrm{out}}$. Let G_i^{p} be the c.d.f. of types that participate in venue i. If all potential investors join venue i, we simply have $n_i=1$ and $G^{\mathrm{p}}=G$. In the generic case, however, we have $n_i G_i^{\mathrm{p}}\leq G$ since some investors do not participate. Indeed, we shall see in the multiple-venue model that the support of G^{p} is typically not connected.

Let us now formally define an equilibrium of the game.

DEFINITION 1: A market structure equilibrium is a set of participation decisions by traders and entry, speed, and fee strategies by trading venues, such that

- Venues maximize profits: The sequence of entry, speed, and fee strategies is a Nash equilibrium of each corresponding stage game (Stages II to IV).
- Participation decisions are optimal: For all σ and all i, $\mathcal{P}(\sigma)=i$ implies $W(\sigma,\sigma_i^{\mathrm{t}},s_i)-q_i\geq W(\sigma,\sigma_j^{\mathrm{t}},s_j)-q_j$ for all $j\neq i$.
- The distribution of types in venue i is consistent with individual participation decisions: $n_i=\int_{\mathcal{P}(\sigma)=i}dG(\sigma)$ and $g_i^{\mathrm{p}}(\sigma)=\frac{g(\sigma)}{n_i}\mathbf{1}_{\mathcal{P}(\sigma)=i}$ for all $\sigma\in[0,\bar{\sigma}]$.
- The venue affiliation market clears: $\sum_{i=0:I}n_i G_i^{\mathrm{p}}(\sigma)=G(\sigma)$ for all $\sigma\in[0,\bar{\sigma}]$.
- Subsequent asset prices and marginal types satisfy equations (4) and (5).

Sequential rationality of venue strategies is obtained by backward induction. We describe the fee-, speed-, and entry-stage payoff functions in Sections 5, 6, and 7, respectively.

Welfare and Regulation

Let $\mathcal{W}$ measure the welfare gains of a given market structure relative to the no-trade benchmark. From previous definitions, we have

$$\mathcal{W}\equiv\underbrace{\sum_{i=1:I}n_i\int_{\sigma}\big(W\big(\sigma,\sigma_i^{\mathrm{t}},s_i\big)-W_{\mathrm{out}}\big)\,dG_i^{\mathrm{p}}(\sigma)}_{\text{Total gains from trade}}-\underbrace{\sum_{i=1:I}\big(\kappa+C(s_i)\big)}_{\text{Entry and speed investment}}. \tag{9}$$

Welfare gains are given by the sum of investors' expected participation gains minus the fixed entry costs and the costs of investments in speed.

One natural welfare benchmark is a frictionless *Walrasian allocation* where investor participation is free ($q = 0$) and entry and speed costs are zero. Taking the limit $\rho \to \infty$ in equations (4) and (5), one obtains a marginal trading type given by $\sigma_W = G^{-1}(1 - 2\overline{a})$ and price $p_W = \frac{1}{r}[\mu + G^{-1}(1 - 2\overline{a})]$. The instantaneous volume rate equals $\mathcal{V}_W = \frac{\overline{a}\gamma}{2}$ and total gains from trade are given by $\frac{1}{2r}\int_{G^{-1}(1-2\overline{a})}^{\overline{\sigma}} \sigma \, dG(\sigma)$. To evaluate outcomes, however, we consider additional welfare benchmarks, a break-even planner, and a set of regulators (for speed, entry, etc.), as follows.

DEFINITION 2: The *break-even planner* maximizes welfare by choosing entry and investments subject to break-even constraints for each venue. *A regulator* can only affect one stage of the game, taking as given the equilibrium in all other stages.

We see the planner as an insightful benchmark to study the efficiency of the market allocation. Because the planner is subject to frictions and a break-even constraint, comparisons with market allocations will yield more conservative welfare losses than if we used the Walrasian allocation or allowed the planner to receive subsidies instead. The motivation to introduce the regulator as a second benchmark is to shed light on regulatory actions that better resemble the real world. Regulatory agencies typically do not decide on the entire market structure all at once but through regulations affecting specific aspects of it. Similarly, in our model, a regulator affecting speed choices in Stage III, for example, takes venues' entry and fee decisions as given in Stages II and IV.

Price Fragmentation and Order Protection Rule

Whenever more than one trading venue is active, the market observes some degree of *volume fragmentation*. With a duopoly, for example, volume fragmentation is maximal when each venue accounts for half of the total trades. In principle, different venues can also execute orders at different prices, leading to *price fragmentation*. Naturally, the degree of price fragmentation depends on the ability of traders in different venues to interact with each other. There are two extreme cases of analysis.

DEFINITION 3: We say that there is *segmentation* if venues do not execute orders from traders affiliated with another venue. If, instead, venues give access to the same market, with a single clearing price, we say there is *integration*.

In our model, a venue is an access gate to a market where transactions clear at a given price. Under segmentation, an investor joins a venue and never buys from—and never sells to—an investor from the other venue. Segmentation thus means that the venues give access to different markets and therefore to different market clearing prices. Integration means that two venues offer access to the same market with a single market clearing price. The trades cleared in that market come from both venues. A fast venue merely provides faster access to the market. Most real-world asset markets are, of course, somewhere in between these two polar cases. Arbitrage is imperfect because of many well-recognized frictions. For this reason, regulating agencies have designed specific rules that address the potential negative consequences of price fragmentation, such as the SEC's Trade-Through rule described in Section 2. In subsequent sections, we interpret integration as a stylized case that captures the outcome of an order price protection rule that is enforced by the regulator.

5. FEE COMPETITION AND VENUE AFFILIATION

In this section, we analyze fee competition among a given set of trading venues and the resulting allocation of investors across venues.

5.1. *Monopoly*

Consider the case of one venue charging a membership fee q. Recall that we have defined σ^{t} as the marginal trading type. We now define σ^{p} as the *marginal participating type*, that is, as the largest solution to $W(\sigma^{\mathrm{p}}, \sigma^{\mathrm{t}}, s) - W_{\mathrm{out}} = q$. The value function (8) is flat for all types below σ^{t}, and in any interior solution, the marginal trading type must also be the marginal participating type: $\sigma^{\mathrm{p}} = \sigma^{\mathrm{t}}$. This fact implies that

$$q = \frac{\bar{a}}{r} s \sigma^{\mathrm{p}}. \tag{10}$$

All types below σ^{p} are indifferent between joining and staying out. Let ℓ be the mass of light traders. Market clearing requires $\ell = (\frac{1}{2\bar{a}} - 1)(1 - G(\sigma^{\mathrm{p}}))$. That is, when $\bar{a} < \frac{1}{2}$, there are ℓ light traders who join and sell their asset but do not trade again.[16] The equilibrium is depicted in Figure 3. We have an interior solution (where some traders do not join) as long as $\ell < G(\sigma^{\mathrm{p}})$, that is, as long as $G(\sigma^{\mathrm{p}}) > 1 - 2\bar{a}$. In the remainder of the paper, we assume either that $\bar{a}$ is close enough to $\frac{1}{2}$ or that there is a sufficient mass of low-σ type investors to ensure the existence of interior solutions.

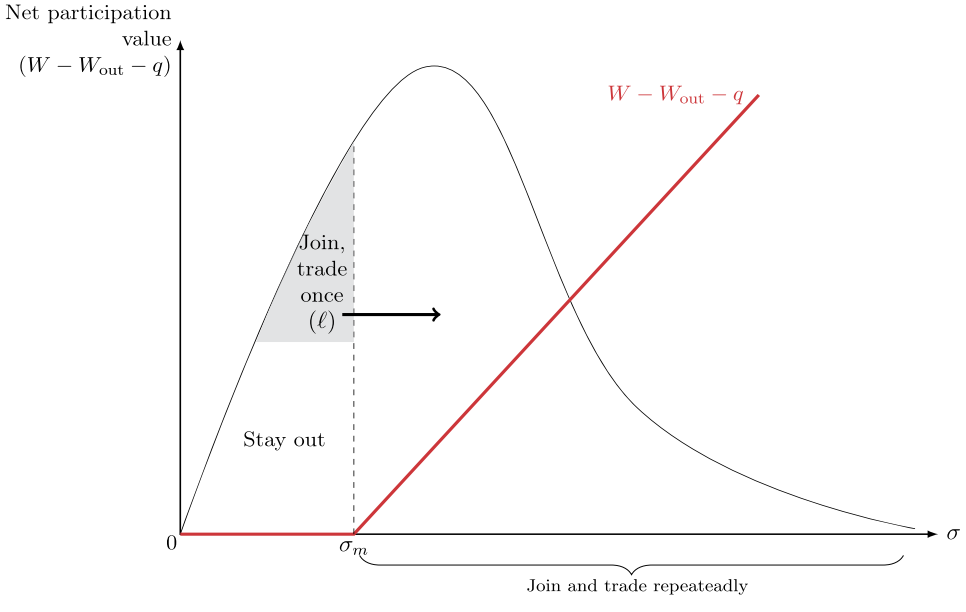

FIGURE 3.—Investor affiliation choices with one trading venue.

[16]There can also be a corner solution with full participation, characterized by the market clearing condition $G(\sigma_{\min}) = 1 - 2\bar{a}$. All investors pay the participation fee $q_{\min}$, which is also the total profit of the trading venue. Then, $G(\sigma_{\min})$ investors sell and drop out, while the remaining $1 - G(\sigma_{\min})$ investors trade in the market with a supply per capita of $1/2$. The participation condition is simply $\hat{V} - q \geq \mu \frac{\bar{a}}{r}$. There is full participation as long as $q \leq q_{\min} = \frac{s}{r} \bar{a} \sigma_{\min}$.

Consider now the profits of a monopolist that selects in this stage an access fee q_m. For simplicity of notation, we write $\sigma_m \equiv \sigma_m^{\mathrm{p}}$ hereafter. Total profits for the venue are given by $\pi_m = q_m(1 - G(\sigma_m) + \ell_m)$. Profit maximization with respect to q_m and subject to (10) and market clearing leads to the following lemma.

LEMMA 1: *The monopolist chooses a level of participation σ_m that is independent of its speed and satisfies*

$$1 - G(\sigma_m) = g(\sigma_m)\sigma_m. \tag{11}$$

First-order conditions are sufficient in this environment. Note that since g is positive and log-concave, it is also quasi-concave. Thus the tail distribution $1 - G$ is quasi-concave as well, which results in the quasi-concavity of $\sigma(1 - G(\sigma))$. The fact that σ_m is independent of the speed in the venue stems from our assumption that the marginal cost of adding traders to an existing venue is zero.[17] The monopoly fee q_m is proportional to the effective speed s_m that is determined in Stage III, as in equation (10).

5.2. *Duopoly Under Segmentation*

Since we assume that venues compete in fees à la Bertrand, a duopoly equilibrium without differentiation implies zero fees and zero profits. The interesting case arises for differentiation by speed. Without loss of generality, we take $s_2 > s_1$, so we refer to venue 2 as the fast venue. For simplicity of notation, we write the marginal participating type in venue $i = 1, 2$ as $\sigma_i \equiv \sigma_i^{\mathrm{p}}$ hereafter.

Consider first the case in which venues are segmented and thus prices can be different. Investors anticipate that each venue i will be characterized by its speed and price, which together define the marginal trading type σ_i^{t}. An investor of type σ estimates the net value from joining venue $i = 1, 2$ is $W(\sigma, \sigma_i^{\mathrm{t}}, s_i) - W_{\mathrm{out}} - q_i$. These value functions are depicted in the middle panel of Figure 4. We know that each venue must attract a mass ℓ_i of light traders and that their value functions are not super-modular in (σ, s). Because these types must be indifferent between joining and staying out, we must have $W(\sigma_i^{\mathrm{t}}, \sigma_i^{\mathrm{t}}, s_i) - W_{\mathrm{out}} - q_i = 0$ in both venues. In other words, as in the case of the monopoly, the marginal trading type σ_i^{t} must be indifferent between participating in venue i and not. Therefore, we must have

$$q_i = \frac{\bar{a} s_i \sigma_i^{\mathrm{t}}}{r}, \quad i = 1, 2. \tag{12}$$

Note, however, an important difference from the monopoly case. The marginal trader in venue 2, σ_2^{t}, would indeed be indifferent between joining venue 2 and not participating. But it is clear from Figure 4 that σ_2^{t} in fact joins venue 1. This means that, with two venues, marginal trading types and marginal participating types are not the same. They coincide only for the slow venue: $\sigma_1 = \sigma_1^{\mathrm{t}}$ but $\sigma_2 > \sigma_2^{\mathrm{t}}$. We then characterize a new marginal type, σ_2, that is indifferent between joining venue 1 and venue 2 and, therefore,

[17] If c were the marginal cost of adding a trader to the venue, profits would be $\pi = (q - c)\frac{1 - G(\sigma_m)}{2\bar{a}} = (\frac{s\bar{a}}{r}\sigma_m - c)(1 - G(\sigma_m))$ and the first-order condition would be $(1 - G(\sigma_m)) = g(\sigma_m)(\sigma_m - \frac{rc}{s\bar{a}})$. In this case, σ_m would depend on s. This effect does not add new insight, so we drop it. Arguably, such marginal costs related to adding an extra trader are less important in real markets than fixed investment costs in technology, software, and infrastructure.

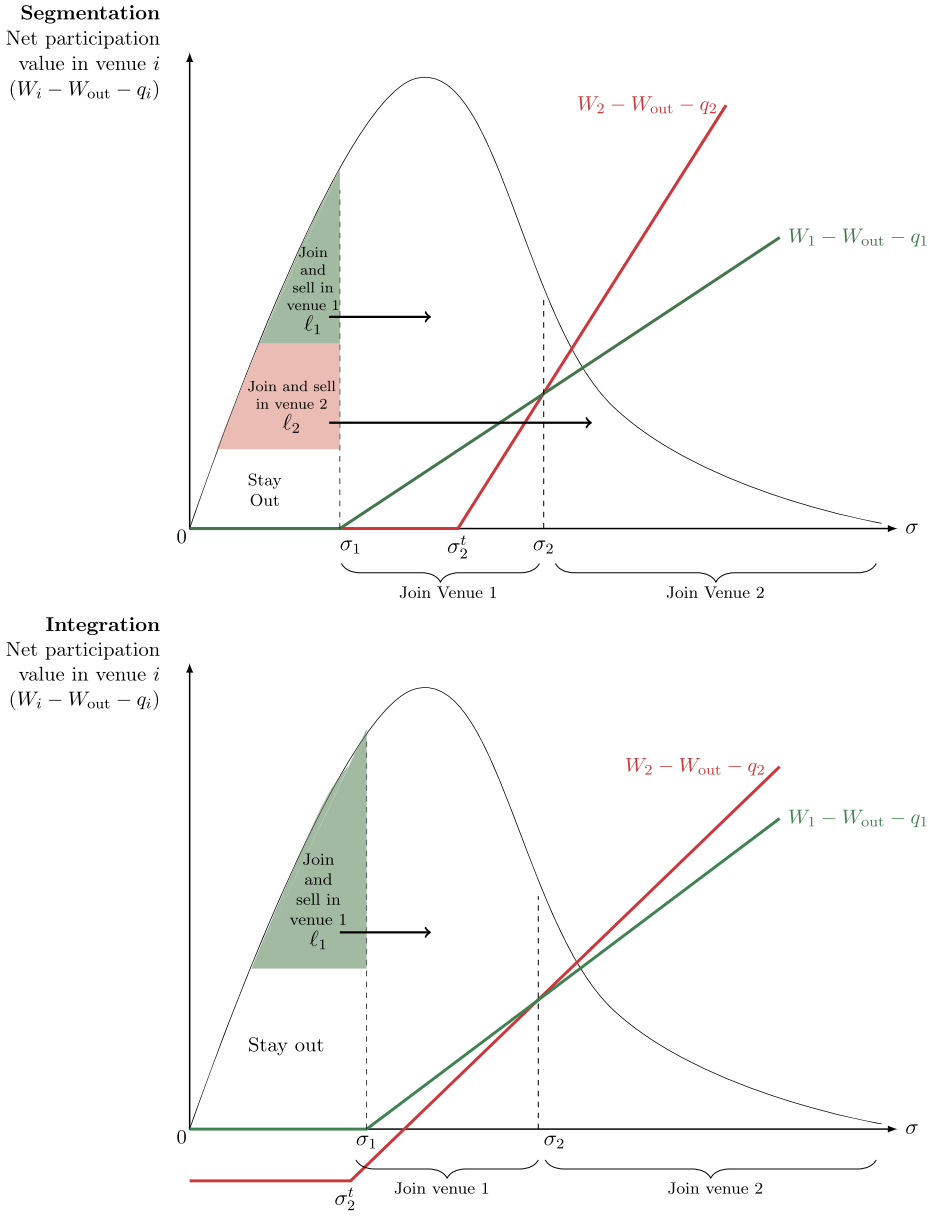

FIGURE 4.—Investor affiliation choice with two trading venues.

satisfies $W(\sigma_2, \sigma_2^{\text{t}}, s_2) - q_2 = W(\sigma_2, \sigma_1^{\text{t}}, s_1) - q_1$. Using equation (12), we then obtain

$$\sigma_2 = \frac{r}{\bar{a}} \frac{q_2 - q_1}{s_2 - s_1}. \tag{13}$$

Note that $\sigma_1 = \sigma_1^{\text{t}} < \sigma_2^{\text{t}} < \sigma_2$. The set of types that join venue 2 cannot be continuous over an interval. It is composed of all the types above σ_2 and some types below σ_1. The affiliation is depicted in the top panel of Figure 4.

Market clearing in venue 2 requires $(1 - G(\sigma_2) + \ell_2)\bar{a} = \frac{1-G(\sigma_2)}{2}$. The payoff for the fast venue under segmentation is $\pi_{2,\text{seg}} = q_2(1 - G(\sigma_2) + \ell_2) = q_2\frac{1-G(\sigma_2)}{2\bar{a}}$. Market clearing for the slow venue requires $(G(\sigma_2) - G(\sigma_1) + \ell_1)\bar{a} = \frac{G(\sigma_2)-G(\sigma_1)}{2}$. The payoff for the slow venue is $\pi_{1,\text{seg}} = q_1\frac{G(\sigma_2)-G(\sigma_1)}{2\bar{a}}$. The affiliation of investors to venues 1 and 2 is given by the marginal types described in (10) and (13), respectively. Venues 1 and 2 simultaneously select fees q_1 and q_2 so as to maximize their payoffs. The first-order conditions from the system result in the following lemma.

LEMMA 2: *In a segmented duopoly, marginal participating types* $(\sigma_{1,\text{seg}}, \sigma_{2,\text{seg}})$ *solve the system*

$$1 - G(\sigma_2) = g(\sigma_2)\left(\sigma_2 + \frac{\sigma_1}{\frac{s_2}{s_1} - 1}\right), \tag{14}$$

$$G(\sigma_2) - G(\sigma_1) = \left(g(\sigma_1) + \frac{g(\sigma_2)}{\frac{s_2}{s_1} - 1}\right)\sigma_1. \tag{15}$$

The price of the asset is higher in the fast venue, $p_2 > p_1$, *as long as* $\bar{a} < \frac{1}{2}$.

The system of equations (14) and (15) shows that equilibrium participation depends only on the degree of speed differentiation $\frac{s_2}{s_1} \in [1, \infty)$.

5.3. *Duopoly Under Integration*

Consider now the case in which both venues provide access to a single market asset and a single price p. The analysis of competing venues in this case is affected by short term price dynamics: the price must be relatively high initially to ensure market clearing when assets are concentrated in the slow venue.[18] Over time, assets migrate to the fast venue, the tradable supply increases, and the price decreases to its long run equilibrium. Formally, we obtain the following result.

LEMMA 3: *The transition dynamics are characterized by an asset migration rate* m *given by* $m(t) = (1 - G(\sigma_2))(\frac{1}{2} - \bar{a})(1 - e^{-\rho_2 t})$, *a price process* $p_t = \frac{\mu}{r} + \frac{\sigma^{\text{t}}_{1,t}}{r}\frac{r+\rho_1}{r+\gamma+\rho_1}$, *with* $\sigma^{\text{t}}_{1,\infty} = \sigma^{\text{t}}_1$ *as in equation* (12), *and the value function of light traders* $W_t = W + Be^{-\rho_2 t}$, *where* W *is the steady-state solution in Proposition* 1. *When the distribution of types is uniform,* $p_t = p + ke^{-\rho_2 t}$, *where* p *is the steady-state market clearing price.* B *and* k *are time-invariant quantities defined in Appendix D of the Supplemental Material* (*Pagnotta and Philippon* (2018)).

These transition dynamics can thus be computed in closed form, but they complicate the value functions without adding new insights. Moreover, they are not quantitatively important as long as r is small relative to ρ, which is clearly the case in practice. Using realistic values for r and ρ (see the quantitative sections below), we find that the degree of approximation of our value function is around 6% for the most affected traders (and

[18]We are grateful to a referee for pointing out that, with integrated prices, we need to distinguish the long run market clearing in stocks from the sequence of market clearing in flows.

much less for the other ones). For the sake of simplicity, we therefore ignore the transition dynamics and concentrate on the steady-state expressions of the value functions.

Let us analyze the equilibrium in the affiliation game. The marginal participating type in venue 1, σ_1, is still characterized by the indifference condition (12). The indifference condition for σ_2 is still $W(\sigma_2, \sigma_2^{\mathrm{t}}, s_2) - q_2 = W(\sigma_1, \sigma_1^{\mathrm{t}}, s_1) - q_1$, but the relation between marginal trading types is now different, affecting the derivation of σ_2. From equation (4), a single asset price implies that $(1+\frac{\gamma}{r+\rho_1})\sigma_2^{\mathrm{t}} = (1+\frac{\gamma}{r+\rho_2})\sigma_1^{\mathrm{t}}$, and thus $\sigma_2^{\mathrm{t}} < \sigma_1^{\mathrm{t}}$. The structure of the value functions under integration is depicted in the bottom panel of Figure 4 for the relevant case where $q_2 > q_1$. Therefore, the light traders join venue 1, where they can sell at a higher price than under segmentation because they can sell to investors in venue 2. The following result summarizes the allocation of traders.

PROPOSITION 2: *Consider a duopoly with $\rho_2 > \rho_1$. With integrated prices, $p_1 = p_2$, light traders join venue 1 only and market clearing requires $\ell_1 = (\frac{1}{2\bar{a}} - 1)(1 - G(\sigma_{1,\mathrm{int}}))$. The type that is indifferent between participating in venues 1 and 2 is given by $\sigma_{2,\mathrm{int}} = \frac{2r}{s_2 - s_1}(q_2 - \frac{z}{2\bar{a}}q_1)$, where $z \equiv 1 - \frac{1+\frac{r}{\rho_1}}{1+\frac{r}{\rho_2}}(1-2\bar{a})$. Moreover, in equilibrium, marginal participating types $(\sigma_{1,\mathrm{int}}, \sigma_{2,\mathrm{int}})$ satisfy*

$$1 - G(\sigma_2) = g(\sigma_2)\left(\sigma_2 + z\frac{\sigma_1}{\dfrac{s_2}{s_1} - 1}\right), \tag{16}$$

$$G(\sigma_2) - \frac{G(\sigma_1)}{2\bar{a}} = \left(\frac{g(\sigma_1)}{2\bar{a}} + z\frac{g(\sigma_2)}{\dfrac{s_2}{s_1} - 1}\right)\sigma_1 + 1 - \frac{1}{2\bar{a}}. \tag{17}$$

Note that the allocation under price integration converges to that under segmentation when $\bar{a} = \frac{1}{2}$. In the general case, however, price fragmentation in the 'downstream' market (the asset market) affects the distribution of investors in the 'upstream' market (the market for venue services).

5.4. *Break-Even Planner Allocation*

Let us consider the break-even planner problem. We first perform the analysis for a given value of s_1. The program is $\max_{s_2, q_1, q_2} \mathcal{W}$ subject to the break-even constraint $q_2(1 - G(\sigma_2)) \geq C(s_2)$. With financing constraints, one might expect the planner to open two venues in order to relax the break-even constraints by charging a high price for the fast venue while maintaining participation in the slower, but cheaper, venue. Perhaps surprisingly, however, we find that the planner chooses to operate only one venue. To understand the intuition, it is better to think of σ_1 and σ_2 as control variables instead of q_1 and q_2. We show in Appendix C.2 of the Supplemental Material that the planner's Lagrangian is

$$\begin{aligned}\mathcal{L}(s) = {} & s_1 \int_{\sigma_1}^{\overline{\sigma}} \sigma \, dG(\sigma) + (s - s_1) \int_{\sigma_2}^{\overline{\sigma}} \sigma \, dG(\sigma) - (1+\lambda) 2rC(s) \\ & + \lambda\big((s - s_1)\sigma_2 + s_1\sigma_1\big)\big(1 - G(\sigma_2)\big),\end{aligned}$$

where λ is the multiplier of the budget constraint of the fast venue, and we have used $q_2 = (s - s_1)\sigma_2 + s_1\sigma_1$. The welfare cost of raising σ_1 is $s_1\sigma_1 g(\sigma_1)$, and the financing gain

is $\lambda s_1(1 - G(\sigma_2))$. It is simple to show that the ratio of gains to costs is always higher for σ_1 than for σ_2. This implies that the planner chooses to increase σ_1 until it reaches σ_2. In other words, the slow venue is always inactive. The planner chooses a single venue even when $\kappa = 0$. The result is the same if the planner also chooses s_1, and it extends to the case where prices in the venues can be consolidated. We can then state the following result.

PROPOSITION 3: *The planner chooses one venue with higher participation than under monopoly.*

It is intuitive that participation is higher with the planner than with the monopoly. The comparison of speed choices is ambiguous. If the zero profit condition does not bind, the planner chooses a higher speed than the monopoly. If it binds, however, one can construct examples where the planner chooses a slower speed than the monopoly.

5.5. *Investor Participation in Different Market Structures*

We relate in this section the properties of the affiliation game in various market structures, taking as a given the entry decisions and speed choices. To prove some of our results, we need to make assumptions about the distribution of investor types. We maintain the following assumption throughout the paper.

ASSUMPTION 1: *The distribution of types σ is such that, for all σ,*

$$2g(\sigma) + g'(\sigma)\frac{1 - G(\sigma)}{g(\sigma)} \geq 0.$$

Assumption 1 is needed to prove a basic yet essential result. At the core of our analysis is the idea that vertical differentiation (via investment in trading technology) decreases fee competition. We then need to show that, in equilibrium, fees are higher and participation is lower when trading speeds are more differentiated. Assumption 1 is needed to prove this comparative static. Assumption 1 is not restrictive: It holds for all the distributions that we consider in our numerical analysis and many others.[19] Some results, however, can only be proven for specific classes of distributions and we use two such classes.

DEFINITION OF DISTRIBUTIONS: The exponential distribution is given by $G(\sigma) = 1 - e^{-\frac{\sigma}{\nu}}$. The uniform distribution is given by $G(\sigma) = \frac{\sigma}{\bar{\sigma}} 1_{\sigma \in [0, \bar{\sigma}]}$.

The following proposition characterizes the equilibrium of the affiliation game.

PROPOSITION 4: *The equilibrium of the affiliation game has the following properties*:

(i) Competition among venues increases participation. *With or without price integration and for a given speed, participation in the fast venue alone is higher than total participation under a monopoly, that is, $\sigma_2 < \sigma_m$. Total participation is even higher, since $\sigma_1 < \sigma_2$.*

(ii) Speed differentiation, defined as $\frac{s_2}{s_1}$, relaxes price competition. *Under Assumption 1, participation with a duopoly is lower (σ_1 and σ_2 are higher) when speeds are more differentiated.*

[19] For example, it holds for exponential, normal, log-normal, Pareto, Weibull, inverse Gaussian, gamma, and Kumaraswamy distributions. Assumption 1 is not standard in the vertical differentiation literature as virtually all of the contributions there assume that agent types are uniformly distributed.

(iii) Price integration increases the profits of the slow venue and decreases total participation, *that is*, $\pi_1^{\text{int}} \geq \pi_1^{\text{seg}}$ *and* $\sigma_{1,\text{int}} \geq \sigma_{1,\text{seg}}$. *Conditional on speed*, *price integration has an ambiguous impact on participation in the fast venue*. (*The proof is analytical for exponential and uniform distributions*, *and numerical in other cases*.)

The intuition for (i) is simply that fee competition fosters participation. A result that is perhaps less obvious is that participation in the fast venue *alone* is already higher than total participation with a monopoly. Point (ii) helps us understand how speed choice affects the affiliation game. Based on the system given by equations (14) and (15), we show in Appendix B that $\frac{\partial \sigma_1}{\partial(\frac{s_2}{s_1})} > 0$ and $\frac{\partial \sigma_2}{\partial(\frac{s_2}{s_1})} > 0$. This result is fundamental since it shows that differentiation decreases competition and therefore decreases trader participation.

Point (iii) shows that price integration has two main effects. First, it increases the comparative advantage of the slow venue because it allows its investors to trade with investors from the fast venue. The per-capita gains from trade are higher in the fast venue because of higher speed and because the fast venue attracts investors with high σ values, as can be seen from equation (8) and the fact that $\frac{s\bar{a}\sigma^{\text{t}}}{r} = \frac{\rho}{r+\rho}(p - \frac{\mu}{r})\bar{a}$. Under price integration, the investors in the slow venue benefit from these gains from trade, which makes traders more willing to join the slow venue in the first place. The second effect of integration is to soften the fee elasticity of the marginal type σ_2 by making the value function steeper. With more demand and less competition, venue 1 can charge a higher fee and make higher profits. Interestingly, the soft competition effect is strong enough that participation decreases $\sigma_{1,\text{int}} \geq \sigma_{1,\text{seg}}$. On the other hand, participation in the fast venue can go up or down. If venue speeds are relatively similar, then the decrease in the fee elasticity leads to a large increase in q_1 and an increase in participation in venue 2. If, instead, venue speeds are very different, the change in q_1 is smaller, and participation in venue 2 may decrease. In our calibrated asset markets, we find that price protection leads to an increase in participation for venue 2 (see Table VII).

6. TRADING SPEED

This section analyzes investment in trading technology, taking as a given the set of active venues. We focus on the case in which $\bar{a} = 1/2$ to separate this analysis from that of price integration in the previous section. Based on the analysis in Sections 4 and 5, we can rewrite equation (9) as the following lemma.

LEMMA 4: *Social welfare in a duopoly is*

$$\mathcal{W}(2) = \frac{s_1}{2r}\int_{\sigma_1}^{\sigma_2} \sigma \, dG(\sigma) + \frac{s_2}{2r}\int_{\sigma_2}^{\bar{\sigma}} \sigma \, dG(\sigma) - \sum_{i=1,2} C(s_i) - 2\kappa. \tag{18}$$

We show in the proof of Lemma 4 that $\mathcal{W}(2)$ contains an extra term under integration when $\bar{a} \neq 1/2$. Welfare in a monopoly, $\mathcal{W}(1)$, can be seen as a particular case of equation (18) with $s_1 = 0$ and only one entry cost κ. When convenient to derive closed-form solution, we assume that the cost of speed ρ is linear, $c\rho$, with $c \geq 0$. Given that $s \equiv \frac{\rho}{r+\gamma+\rho}$, this implies the following cost expression.

ASSUMPTION 2: *The cost of reaching the effective speed s is* $C(s) = c(r+\gamma)\frac{s}{1-s}$.

6.1. *Venue Speed Choices*

Monopoly

Lemma 1 shows that the participation cutoff σ_m chosen by a monopolist does not depend on its effective speed s_m. The monopolist chooses its speed to maximize $s\frac{\sigma_m}{2r}(1 - G(\sigma_m)) - C(s)$, as in the following proposition.

PROPOSITION 5: *The monopolist chooses a speed level* s_m *such that* $\frac{\partial C}{\partial s}(s_m) = (1 - G(\sigma_m))\frac{\sigma_m}{2r}$, *where* σ_m *is given by equation* (11). *Under Assumption* 2, *we have the following closed-form expressions. If types are exponentially distributed*, $s_m = 1 - \sqrt{2rc(r+\gamma)e/\nu}$. *If the types are uniformly distributed*, $s_m = 1 - \sqrt{8rc(r+\gamma)/\overline{\sigma}}$.

The effective speed s (or the contact rate ρ) decreases with the cost parameter c and increases with the average size of private preference shocks (e.g., an increase in ν and $\overline{\sigma}$). For instance, with an exponential distribution, when ν increases, the distribution has a fatter right tail, gains from trade increase, and the demand for speed also increases, as one would expect from Proposition 1. Moreover, s decreases with the frequency of preference shocks γ because when γ is high, the desired holding period shrinks. However, more interestingly, since $\rho = (r+\gamma)s/(1-s)$, the optimal contact rate ρ_m is concave in γ. Starting from a low γ, as the frequency of preference shocks increases, investors will want to reallocate their assets more frequently, which increases the demand for speed. When γ is very high, though, the holding period effect dominates.

Duopoly

In a duopoly, venues have an incentive to offer different speeds to reduce price competition. Recall that the revenue functions for venues 1 and 2 can be expressed as $\pi_1 = q_1(G(\sigma_2) - G(\sigma_1))$ and $\pi_2 = q_2(1 - G(\sigma_2))$, and that fees are given by $q_1 = \frac{1}{2r}s_1$ and $q_2 = \frac{1}{2r}(\sigma_2(s_2 - s_1) + \sigma_1 s_1)$, and the affiliation equilibrium is given in Lemma 2. Venues 1 and 2 then simultaneously solve $\max_s \Pi_i(s, s_{-i}) = \pi_i(s, s_{-i}) - C(s)$. The optimality conditions that determine the solutions (s_1, s_2) are presented in Appendix B.6.

Let us now compare the market equilibria under monopoly and duopoly. There is a fundamental tension between profitability and elasticity. On the one hand, the marginal return to speed depends on $\sigma(1 - G(\sigma))$ for both the monopolist and the fast duopolist. The monopoly chooses σ_m to maximize precisely this quantity; therefore, we know that $\sigma_m(1 - G(\sigma_m)) > \sigma_2(1 - G(\sigma_2))$. This profitability effect makes the monopolist more willing to invest in speed. On the other hand, competing venues have an incentive to differentiate their services. As s_2 increases, competition is relaxed, q_1 increases, and σ_2 decreases, resulting in greater participation and higher profits for venue 2. Which effect dominates depends on the distribution of types. With a uniform distribution of types, we can show in Appendix B that the second effect dominates and, therefore, that the equilibrium speed is higher under a duopoly.

PROPOSITION 6: *With uniformly distributed types, the fast duopolist venue chooses a higher speed than a monopoly does*: $s_2 \geq s_m$.

6.2. *Regulation of Speed*

Let us now study the welfare consequences of speed choices. We consider a game where the regulator can mandate speed bounds to increase social welfare, taking as a given venue fee choices and investor affiliation decisions.

DEFINITION: The regulator can set a minimum speed $\underline{s}$ and a maximum speed $\overline{s}$.

Assuming a uniform distribution of types, we obtain the following result.

PROPOSITION 7: *When types are uniformly distributed, it is optimal for the regulator to mandate a minimum speed but not a maximum speed. That is,* $\underline{s} > s_1$ *but* $\overline{s} = 1$.

Consider the monopoly first. The speed chosen by the monopolist is as in Proposition 5. The regulator seeks to maximize social welfare and thus solves $\max_s \frac{s}{2r}\int_{\sigma_m}^{\bar{\sigma}} \sigma\, dG(\sigma) - C(s)$ taking σ_m as a given. Since $\int_{\sigma_m}^{\bar{\sigma}} \sigma\, dG(\sigma) > \sigma_m(1 - G(\sigma_m))$, the regulator's optimal is greater than s_m for any distribution of types G. The regulator prefers a higher speed than the monopoly because she values the welfare gain for the infra-marginal types ($\sigma > \sigma_m$) while the monopolist does not.

Consider now the duopoly. Our first result is that maximum speed limits are not efficient. Under the duopoly, speed allows venues to differentiate and relax Bertrand competition. The regulator trades off efficiency for high-σ types against participation for low-σ types. The regulator's first-order condition is

$$2r\frac{\partial C}{\partial s}(s_2) = \int_{\sigma_2}^{\bar{\sigma}} \sigma\, dG(\sigma) - (s_2 - s_1)\sigma_2 g(\sigma_2)\frac{\partial \sigma_2}{\partial s_2} - s_1\sigma_1 g(\sigma_1)\frac{\partial \sigma_1}{\partial s_2}.$$

The term $\int_{\sigma_2}^{\bar{\sigma}} \sigma\, dG(\sigma)$ is the surplus of the high-σ types that the fast venue does not appropriate and therefore does not internalize. Allocation efficiency for types $\sigma > \sigma_2$ calls for higher speed. On the other hand, $\frac{\partial \sigma_2}{\partial s_2}$ and $\frac{\partial \sigma_1}{\partial s_2}$ capture the impact of s_2 on differentiation, which softens competition. The link between social welfare and speed depends on the tradeoff between participation and trading efficiency for the high-σ types. We show in Appendix B that the trading efficiency effect dominates when the types are uniformly distributed. This is particularly interesting, since we have shown in Proposition 6 that the fast duopolist chooses a higher speed than a monopoly does. Proposition 7 states that this is not enough and the regulator would like an even higher speed. Therefore, the regulator does not find it optimal to impose an upper limit on speed.

On the other hand, it is optimal for the regulator to impose a minimum speed requirement that is higher than that chosen by the slow venue. The intuition is that such a minimum speed increases the welfare of the low-σ types and at the same time intensifies fee competition among venues.[20] We provide calibrated welfare enhancement estimates in Section 8 and further discuss related regulations in Section 9.

7. ENTRY

This section completes the determination of the market structure by analyzing entry decisions.

[20]Our result for $\underline{s}$ can be seen as an extension of a result of Ronnen (1991), who analyzed minimum quality standards in a simpler static Shaked–Sutton (1982) framework with exogenous preferences for a final product quality. Note that, although we do not model venues offering menus of speeds to investors, our analysis could be extended in this direction. Champsaur and Rochet (1989) analyzed a multi-product oligopoly where firms produce a range of qualities. They showed that firms provide non-overlapping quality ranges. Given this paper's result, our intuition is that venues would likely offer non-overlapping menus of speed and that investors with low and high types would similarly sort across venues.

TABLE II
VENUES' ENTRY PAYOFFS AND PRICE FRAGMENTATION

Venue 1 ↓ and 2 →	In	Out
In	$\pi_1^{\text{seg/int}} - \kappa;\ \pi_2^{\text{seg/int}} - \kappa$	$\pi_m(s_1) - \kappa;\ 0$
Out	$0;\ \pi_m(s_2) - \kappa$	$0;\ 0$

7.1. *Entry and Efficiency in a Duopoly*

We analyze first the case of two potential entrants facing an entry cost $\kappa \geq 0$. We model entry as a simultaneous game with exogenous speeds s_1 and s_2, respectively, with the convention that $s_1 < s_2$.[21] A given venue i finds it optimal to enter whenever profits net of entry costs are nonnegative. Venues' revenues depend on price fragmentation, so we consider the functions π_i^{seg} and π_i^{int} as in Section 5. The payoffs of the entry game for venues $i = 1, 2$ are thus as shown in Table II, where the expression $\pi_m(s_i) \equiv \max_q q(1 - G(\sigma_m(q, s_i)))$ denotes the monopolist's optimal profit given speed s_i.

From the previous analysis, we know that, regardless of integration, $\pi_1 < \pi_2$ because venue 2 is faster and attracts participants with higher gains from trade. Furthermore, $\pi_1^{\text{seg}} < \pi_1^{\text{int}}$ from Proposition 4. Since integration increases the expected profit of the slow venue, it expands the range of values of κ for which a duopoly is a Nash Equilibrium of the entry game. In particular, the number of active venues is strictly larger for economies with intermediate entry costs: $\pi_1^{\text{seg}} < \kappa < \pi_1^{\text{int}}$.

Let us now consider the possibility of inefficient entry and ask when the regulator wants to encourage or restrict entry. For brevity, we abstract from price integration (e.g., $\overline{a} = \frac{1}{2}$) so the welfare function under duopoly is as given in Lemma 4 for endogenous speeds (s_1, s_2).[22] The welfare gain of moving from a monopoly to a duopoly, $\Delta\mathcal{W}$, is therefore

$$\Delta\mathcal{W} = \frac{1}{2r}\left(s_1 \int_{\sigma_1}^{\sigma_2} \sigma\, dG(\sigma) + s_2 \int_{\sigma_2}^{\overline{\sigma}} \sigma\, dG(\sigma) - s_m \int_{\sigma_m}^{\overline{\sigma}} \sigma\, dG(\sigma)\right) - \sum_{i=1,2} C(s_i) + C(s_m) - \kappa. \tag{19}$$

Entry is profitable for the slow venue if and only if $\Pi_1 > \kappa$, that is, if $\frac{s_1\sigma_1}{2r}(G(\sigma_2) - G(\sigma_1)) > \kappa + C(s_1)$. Excess entry thus occurs if and only if $\pi_1 > \kappa$ and $\Delta\mathcal{W} < 0$ hold simultaneously. Absent thick market externalities, equation (19) suggests that the source of excess entry can be related to cost duplication in speed investments or fixed entry costs. In this environment, we find that, regardless of entry costs, excess entry of a second venue is unlikely provided speed costs are relatively low.[23]

We summarize the entry results with two potential entrants in the following proposition.

[21] The analysis can easily be extended to endogenous speeds as long as speed choices do not depend on price fragmentation. The general case where $s_i^{\text{int}} \neq s_i^{\text{seg}}$ can only be solved numerically. Our calibrations suggest that speed choices are not significantly affected by price integration.

[22] The analysis easily extends to price integration.

[23] Mankiw and Whinston (1986) identified three general conditions under which excess entry occurs: (i) some form of economies of scale due to fixed costs, (ii) post-entry prices exceeding marginal cost, and (iii) enough business stealing to decrease average firm output. The first two conditions are easily verified in our environment. The third is not, however, since trading services are vertically differentiated. This is a fundamental difference between our model and the Hotelling models of Spence (1976) and Mankiw and Whinston (1986).

PROPOSITION 8: *The equilibrium of the entry game with two potential entrants has the following properties*. (*i*) *When the distribution of types is exponential or uniform*, *the number of entrants is weakly higher under price integration than under segmentation*. (*ii*) *When the distribution of types is uniform*, *for any fixed entry cost* κ, *entry of a second venue improves welfare as long as speed costs are not too high*.

Part (i) suggests that, instead of direct subsidies, a price protection rule can be used to encourage entry. Part (ii) says that moving from a monopoly to a duopoly is likely to be optimal as long as speed costs are not too high. It is indeed sufficient that the incremental cost of s_2 over s_m be small relative to the gains from trade, as it is the case in all our simulations. This fact might not be true when there are already *several* venues and we consider an additional entrant. We now consider this case.

7.2. *Generalized Oligopoly*

We consider first the affiliation game where venues compete in fees to attract investors. As before, let I denote the number of active venues and let σ_i be the lowest type that joins venue i. This marginal participating type is indifferent between venues i and $i-1$; therefore, $\sigma_i = \frac{r}{\bar{a}} \frac{q_i - q_{i-1}}{s_i - s_{i-1}}$. By repeated substitutions, it is easy to show that we must have $q_i = \frac{\bar{a}}{r} \sum_{j=1}^{i} \sigma_j (s_j - s_{j-1})$, where $s_0 \equiv 0$. Defining $\sigma_{I+1} \equiv \overline{\sigma}$, we can write the revenues of any venue $i \in \{1, \ldots, I\}$ as $\pi_i = q_i (G(\sigma_{i+1}) - G(\sigma_i))$. Taking first-order conditions with respect to q_i, we obtain the following result that generalizes the equilibrium with two venues described in Lemma 2.

LEMMA 5—Equilibrium of the Affiliation Game With I Active Venues: *For all* $i \in \{1, \ldots, I\}$, *the set of marginal participating types* $\{\sigma_i : i = 1 : I\}$ *satisfies*

$$G(\sigma_{i+1}) - G(\sigma_i) = \left(\frac{g(\sigma_i)}{s_i - s_{i-1}} + \frac{g(\sigma_{i+1})}{s_{i+1} - s_i} \right) \sum_{j=1}^{i} \sigma_j (s_j - s_{j-1}).$$

We can now study the entry stage. As explained above, moving from one to two venues is likely to be efficient since it introduces competition into a market where there is none. It is less clear whether moving from I to $I+1$ is efficient when I is already above 1. We explore this issue by analyzing a case with three venues. In particular, we consider the unexpected entry of a third venue in an existing duopoly. The incumbents have already chosen their speeds and paid their fixed entry costs, expecting to be in a duopoly. We then ask if the entry of a third venue would raise welfare. The third venue chooses its speed optimally, given the speeds of the existing duopoly. The considered approach allows us to bypass the issue of entry deterrence, which is beyond the scope of this paper. This case can only be solved numerically (see Appendix E.1). We highlight the main finding of this case as follows.

EXAMPLE 1—Inefficient Entry of a Third Venue: With two incumbent venues, entry of a third venue can reduce welfare when the speed of the new entrant is lower than the speed of the slow incumbent.

Entry always increases total participation but can also lead to misallocations, and this effect can be large if entry takes place at the low end of the range of incumbents' speeds.

Let us explain the intuition for this result. To keep our notation simple, we denote the low- and high-speed venues of the existing duopoly by (l, h) and the new entrant by e. In a duopoly, we have the mapping $(l, h) \to (1, 2)$. When we add venue e, the new ordering depends on the relative speed of the entrant. As explained above, here we consider the case in which $s_e < s_l$. In the oligopoly with three venues, the ranking in terms of Proposition 5 is therefore $(e, l, h) \to (1, 2, 3)$. It is clear that entry creates direct competition for venue l, which is forced to lower its fee. The important point is to consider the reaction of venue h. Venue h does not compete directly with venue e, but it competes with l. Venue h reacts to the induced drop in q_l. The optimal pricing condition for venue h is

$$1 - G(\sigma_h) = g(\sigma_h)\left(\sigma_h + 2r\frac{s_l}{s_h - s_l}q_l\right), \tag{20}$$

where σ_h is the marginal participating type in venue h. Under Assumption 1, the function $\frac{1-G(\sigma)}{g(\sigma)} - \sigma$ is decreasing in σ. Therefore, equation (20) implies that σ_h is a decreasing function of q_l. This result explains the potential inefficiency. Entry by the third venue forces the middle venue to lower its fee, but it is not profitable for the fast venue to fully accommodate this fee change. Therefore, $q_h - q_l$ increases and σ_h goes up. Traders who used to trade in the fast venue now trade in the middle venue. This is a misallocation since the planner would rather have more investors in the fast venue. Naturally, the private surplus increases for investors who move from h to l venue, but the profit losses of the fast venue are even greater. The analysis in Appendix E.1 shows that, on the other hand, allowing for a third venue can lead to significant welfare gains when entry takes place at the high end of the speed ladder.

8. CALIBRATION AND WELFARE ANALYSIS

We now calibrate our model to study the impact of speed, fees, and entry decisions on market participation, volume, and welfare. We consider three asset classes that capture the range of speeds discussed in Figure 1: corporate bonds, stocks, and Standard and Poor's (S&P) 500 index futures. We calibrate the model using secondary markets data, and we conduct comprehensive sensitivity analysis for the critical parameters.

8.1. *Method and Asset Classes*

The baseline parameters are displayed in Table III. Unless otherwise noted, these parameters are held constant across our experiments. We assume a uniform distribution of investor types, and we use Assumption 2 for the cost function. We set the fixed cost κ to zero and the asset supply to $\frac{1}{2}$ when we do not analyze price integration.[24] The rate r is based on the composite rate of long-term U.S. Treasury securities from January 2007 to December 2013. We set the asset holding cashflow, μ, to match the average S&P500 div-

[24]Futures are, of course, in zero net supply, but $\overline{a} > 0$ can be interpreted as the case in which the sell side is short the asset and we capture trades among buy-side investors. We could also allow for negative holdings as long as holdings are bounded.

TABLE III

BASELINE PARAMETER VALUES[a]

$g(\sigma)$	$\overline{\sigma}$	κ	$\overline{a}$	r	μ	n
$1/\overline{\sigma}$	$\mu/2$	0	0.5	3.75%/252	2.75/252	28,225

[a]This table displays the following parameters: asset supply ($\overline{a}$), discount rate (r), cash flow (μ), investor σ type density (g), maximum investor type ($\overline{\sigma}$), entry costs (κ), and number of potential investors (n).

idend yield over the same period (2.13%).[25] Trading days last 6.5 hours, as in U.S. equity markets, and we set the upper bound of the σ-type distribution to $\frac{\mu}{2}$.[26]

Choice of Number of Traders n

We have so far assumed a unit mass of investors. To compare the model-implied per-capita volume in equation (7) with the data, we thus need to specify the number of investors, n, as an additional parameter. We start from the number of institutional funds in the United States as a proxy for the size of the buy side of financial markets. According to Morningstar, at the end of 2007, there were 629 exchange-traded funds, 17,500 mutual funds, and 10,096 hedge funds. These 28,225 funds represent the set of potential investors. Let k denote an asset class (bonds, stocks, futures), and let n_k be the potential number of traders in that class. Market participation would be n_k in the first best allocation, $n_k/2$ with a monopoly, and an in-between value with a duopoly. The most difficult part of the calibration is to determine how many of these investors are active in each asset market. We motivate asset-class-specific values below.

Volume-Implied γ

Once we calibrate n_k, we use the observed volume $\mathcal{V}_k$ and speed ρ_k to back out the rate of preference shocks γ_k. For instance, with a single venue, the model-implied transaction rate for a *typical* asset in class k is $\mathcal{V}_k = n_k \frac{\gamma_k}{4} \frac{\rho_k}{\gamma_k+\rho_k}(1 - G(\sigma_m))$. Using a uniform distribution of types, an observed volume rate $\mathcal{V}_k$ implies that $\gamma_k = \frac{8\mathcal{V}_k\rho_k}{n_k\rho_k - 8\mathcal{V}_k}$. With two venues, we have

$$\mathcal{V}_k = n_k \times \frac{\gamma_k}{4}\left[\frac{\rho_{1,k}}{\gamma_k+\rho_{1,k}}\left(\frac{\sigma_{2,k}-\sigma_{1,k}}{\overline{\sigma}}\right) + \frac{\rho_{2,k}}{\gamma_k+\rho_{2,k}}\left(1-\frac{\sigma_{2,k}}{\overline{\sigma}}\right)\right]. \tag{21}$$

Given the considered stylized speed values, the calibrated value of n_k, the observed volume rate $\mathcal{V}_k$, and the marginal participating types as in Lemma 2, we invert equation (21) to recover γ_k.

[25]With $\mu = 2.75$ and $r = 0.0375$, relative to the Walrasian price, the dividend yield is $\frac{\mu}{p_w} \approx 2.13\%$. The value of μ affects the asset price in the model. Real asset prices obviously also reflect market risk exposure, among other factors, which is not the focus of this paper. This parameter also affects global welfare. However, our analysis focuses on the fraction of welfare that is earned in excess of the autarchy value, that is, $W(\sigma,\cdot) - W_{\text{out}}$, as in equation (9). This is the main reason why we do not calibrate μ separately for each asset class (and that we abstract from the fact that the futures contract yields no cash flow).

[26]This parameter is not easy to compute based on market data. Hence, we experimented with different values of $\overline{\sigma}$ for robustness. The results are qualitatively similar and consequently are omitted here. To illustrate the economic interpretation of these values, consider the median investor type, $\frac{1}{2}$, when $\mu = 2$. The annual holding flow utility under a temporary shock ε is $u_{\frac{1}{2},\varepsilon}(1) = 2 + \text{sign}(\varepsilon) \times \frac{1}{2}$. This implies that, when facing a negative or positive temporary shock, the annual flow utility equals 1.5 or 2.5 units of consumption, respectively.

TABLE IV

STYLIZED, IMPLIED, AND PREDICTED PARAMETER VALUES[a]

	Panel I: Stylized Values		
	Corporate Bonds	Stocks	S&P500 Futures
Volume	1.97	3,023.4	1,030,204
Number of assets	21,723	2,805	1
Stylized ρ_m	–	–	117,000
Stylized ρ_1	1	195	–
Stylized ρ_2	39	23,400	–

	Panel II: Model Implied and Predicted Values								
	Corporate Bonds			Stocks			S&P500 Futures		
Number of Traders	8	13	17	61	92	115	14,113	21,169	28,225
Implied γ	1.378	0.834	0.588	299.72	182.95	139.64	586.93	390.63	292.73
Implied c ($\times 10^{-3}$)	36.444	36.201	36.415	0.1605	0.1570	0.1564	2.7457	2.7503	2.7526
Predicted ρ_m	37.377	36.211	35.220	22,616	21,986	21,551	117,000	117,000	117,000
Predicted ρ_1	1.644	1.044	0.750	386.50	239.13	183.38	773.57	516.93	388.13
Predicted ρ_2	40.367	38.132	38.065	24,442	23,758	23,286	126,414	126,402	126,396

[a]Volume figures, speed parameters (ρ), and preference switching rate (γ) are expressed in daily values. Parameters ρ_m, ρ_1, and ρ_2 denote, respectively, the monopolist speed, the slow duopolist speed, and the fast duopolist speed. Parameter c is that of the cost function in Assumption 2. Stylized values are calibrated with market data. The calculation of implied values is described in Section 8.1. Given stylized and implied parameter values, predicted speeds are equilibrium values in the monopoly and duopoly cases.

Speed-Implied c

We compute the implicit cost parameter c that rationalizes ρ as an optimal speed, given all the other parameters. Inverting the monopoly first-order condition in Proposition 5 and assuming a uniform distribution of types, we obtain

$$c_k = \frac{\overline{\sigma}}{8r} \frac{n_k(n_k\rho_k - 8\mathcal{V}_k)\big(n_k r\rho_k + 8\mathcal{V}_k(\rho_k - r)\big)}{\big(n_k\rho_k(\rho_k + r) - 8r\mathcal{V}_k\big)^2}.$$

In the duopoly case, there are two first-order conditions. In principle, one could retrieve two different values of c. We concentrate on the fast venue condition (see equation (25)) and use the resulting c as the single parameter for both venues. The speed of the second venue is by far the most important in terms of welfare. In addition, stylized speed values are more likely to be accurate for fast venues which are more likely to be transparent about their trading delays.

Table IV presents the calibration parameters by asset class. Our calibration is consistent with the common wisdom about the relative allocative efficiency across classes. In particular, the ratio ρ/γ is highest for E-Mini futures and lowest for corporate bonds. Given the challenging choice of n_k, we test the sensitivity of our results to values that are one-third lower or higher than the benchmark.

S&P500 Index Futures

The Chicago Mercantile Exchange (CME) has a monopoly over its E-mini futures contracts, and thus we employ here the monopoly formulas in Section 8.1. There is one contract and, as a benchmark, we consider that most investors are active in this market:

$n_{\text{Emini}} = \frac{3}{4}n$. We use the average number of daily trades on May 6, 2010, as reported by Kirilenko et al. (2017).[27] The stylized speed considered here is equivalent to an average delay of 200 ms (117,000 is five times the number of seconds in a 6.5-hour trading day). The implied γ means that there are 390 shocks per trading day and investor, or approximately one shock every minute. When $\pm\frac{1}{3}n_{\text{Emini}}$, there is one shock every 40 second or 80 seconds instead. We choose c to match the contact rate. If the market were a duopoly, our model would then imply trading speeds for the slow and fast venues such that average delays would be 45 seconds (consistent with human intervention in, say, a traditional trading pit) and 185 ms (automated platform).

Corporate Bonds

All trades for 2013Q4 are collected from the Trade Reporting and Compliance Engine (TRACE) data set. The average daily number of trades for each of these bonds is 1.97, reflecting the fact that most corporate bonds trade infrequently. Our sample contains 21,723 bonds. The non-transparent nature of the corporate bond market makes it difficult to estimate the participants. According to the *Investment Company Fact Book*, out of 8,000 mutual funds surveyed in 2007, about 800 are bond funds, so we assume that 10% of investors are active in corporate bonds. As a starting point, we assume that each participant is active in 100 individual bonds, which is consistent with anecdotal evidence. This gives us $n_{\text{bonds}} = 0.1 * 100 * 28{,}225/21{,}723 = 13$ traders per bond. We perform robustness checks with $n_{\text{bonds}} = 8$ and $n_{\text{bonds}} = 17$, as explained above. We calibrate the corporate bond market as a duopoly. Corporate bonds trade in traditional voice-based OTC broker networks, the slow venue, or in modern electronic platforms, the fast venue. The stylized contact rates are one and 39, respectively. The first is equivalent to an average trading delay of a day, consistent with values in a traditional voice-based OTC network. The latter value represents an average delay of 10 minutes, closer to that of an electronic platform based on the RFQ protocol. The implied γ means that there is 0.834 preference shock per trading day per investor. The implied c value is harder to judge based on intuition alone. Interestingly, after computing the full duopoly equilibrium using this value, we obtain predicted rates $(\rho_1, \rho_2) = (1.04, 38.13)$ which are very close to the stylized values in this market.

Stocks

We calibrate the model to 2007 because that was when Reg NMS was implemented. According to data from the NYSE Group (www.nyxdata.com), the average number of daily trades for a representative NYSE-listed stock in 2007 was equal to 3,023. The number of listed stocks in 2007 was 2,805. According to the *Investment Company Fact Book*, 46% of funds are U.S. equity funds. There is no simple way to estimate n_{stocks} because many equity-related trades are index trades, not trades on individual stocks. We choose the typical number of actively traded stocks to capture the intuitive idea that the stock market lies somewhere in between the bond market and the futures market regarding efficiency and number of traders per asset. Assuming that a trader is active in 20 individual stocks, we obtain a benchmark $n_{\text{stocks}} = 92$. We calibrate the equity market as a duopoly,

[27]This date corresponds to the so-called flash crash and displays both a large volume and a large number of investors trading. Using instead the reported values for May 3 to May 5, 2010, yields a lower value for γ. In our calibration, participation in the monopoly then equals $\frac{n_{\text{Emini}}}{2} \approx 10{,}584$. Kirilenko et al. (2017) reported the number of active daily traders to be between 11,875 and 15,422 in their CME sample.

given the prevalence of the NYSE and the NASDAQ at the time of Reg NMS implementation. To calibrate the stylized contact rate parameters, we consider SEC Rule 605 data for the NYSE for 2007, before the full implementation of Reg NMS. The value for the fast venue matches the average execution delay of 1 second in 2007 for small automated orders. The value for the slow venue represents a human broker–dealer round-trip delay of 1 minute and is consistent with the SEC data presented in Figure A.2 in Appendix A. The implied γ means that there is one shock per investor every 128 seconds in this market.[28] As for the case of corporate bonds, we use the model implied cost parameter c to compute the duopoly equilibria and obtain predicted rates that are reasonably close to its stylized values.

8.2. *Welfare Analysis*

Table V shows the main equilibrium outcomes in the benchmark case with $\overline{a} = \frac{1}{2}$. All the values in the table are relative to the constrained first best as given by the break-even planner (bep) in Definition 2. Panels I to III of Table V, respectively, display the outcomes

TABLE V

CALIBRATION OUTCOMES (PLANNER CASE = 100)[a]

	Corporate Bonds			Stocks			S&P500 Futures		
	Investor Partic.	Trading Volume	Welfare	Investor Partic.	Trading Volume	Welfare	Investor Partic.	Trading Volume	Welfare
I. $\frac{2}{3}n_k$	$\gamma = 1.378$, $c = 0.0364$			$\gamma = 299.72$, $c = 0.000160$			$\gamma = 586.93$, $c = 0.00275$		
Monopoly	50.64	50.11	74.32	50.23	50.04	74.75	50.09	50.02	74.91
Venue 1	29.46	16.45	9.01	29.29	16.65	9.06	29.22	16.67	9.04
Venue 2	58.93	58.46	82.02	58.58	58.41	82.44	58.43	58.37	82.57
Duopoly	88.39	74.91	91.03	87.87	75.07	91.50	87.65	75.04	91.61
II. n_k	$\gamma = 0.834$, $c = 0.0362$			$\gamma = 182.95$, $c = 0.000157$			$\gamma = 390.63$, $c = 0.00275$		
Monopoly	50.40	50.07	74.57	50.15	50.03	74.84	50.06	50.01	74.94
Venue 1	29.37	16.59	9.05	29.25	16.67	9.05	29.20	16.67	9.04
Venue 2	58.74	58.45	82.28	58.50	58.39	82.52	58.40	58.36	82.59
Duopoly	88.11	75.04	91.33	87.74	75.06	91.57	87.60	75.03	91.63
III. $\frac{4}{3}n_k$	$\gamma = 0.588$, $c = 0.0364$			$\gamma = 139.64$, $c = 0.000156$			$\gamma = 292.73$, $c = 0.00275$		
Monopoly	50.29	50.05	74.69	50.11	50.02	74.88	50.04	50.01	74.95
Venue 1	29.32	16.64	9.06	29.23	16.67	9.05	29.19	16.67	9.04
Venue 2	58.64	58.43	82.39	58.46	58.38	82.55	58.38	58.35	82.60
Duopoly	87.96	75.07	91.45	87.69	75.05	91.60	87.58	75.02	91.64

[a]Each cell is normalized relative to the break-even planner outcome as in Definition 2.

[28]It is important to keep several factors in mind when interpreting γ. First, we calibrate our model using institutional investors, who indirectly represent multiple agents (such as retail investors), and it is natural to think of institutional investors as receiving frequent shocks. There is no reliable information about the direct participation of private corporations and wealthy individuals but, of course, if we included those, N_k would increase and γ would decrease. Finally, and most importantly, the common practice of order splitting increases the number of reported trades. It is not possible to identify which trade represents a new trading shock as opposed to a fraction ("child order") of a larger trade. Our model offers a stylized description of the trading process where the incentive for order splitting, namely the price impact, is absent. A more sophisticated specification with order splitting would naturally imply a lower fundamental γ for the same observed volume.

corresponding to the parameters implied by the low, medium, and high values of n_k (see Table IV).

Participation

Total investor participation under a monopoly is slightly above one-half that of the planner. Participation increases dramatically to around 88% when two venues compete. We verify numerically that participation in the second venue alone is always greater than in the monopoly case, as predicted by the theory. Participation levels are similar across asset classes because the degree of *relative* differentiation $\frac{s_2}{s_1}$ is similar across asset classes.[29]

Trading Volume

Even in markets with high speeds, the duopoly fails to realize a major fraction of the potential trades as volume represents about 75% of the planner's. This fact reflects lack of full participation in the duopoly. But, importantly, it also reflects the inefficient allocation of the asset across investors in the first venue due to speed differentiation. The slow venue volume share is roughly one-half of its relative investor participation for all asset classes. In other words, with speed differentiation, the market equilibrium displays much lower levels of volume fragmentation relative to the (also endogenous) distribution of investors across venues.

Welfare

Although the monopoly only attracts nearly half of all investors, it achieves almost three-fourths of the planner's attainable welfare. This is chiefly because those investors who choose to participate benefit from large gains from trade and, as reflected in Table IV, the monopolist offers a relatively high trading speed. The calibration suggests large social gains associated to encouraging entry. Welfare typically increases by at least 15 percentage points when transitioning from one to two venues. Naturally, the gains from trade are disproportionately distributed across the slow and fast venue. For example, in the benchmark case for stocks, the slow venue only contributes 10% of the total gains from trade.

The welfare associated with a particular market structure is driven by both technology and market power frictions. To further understand their relative importance, we decompose the market-planner welfare gap. For the monopoly, the decomposition is as follows:

$$\mathcal{W}_{\text{bep}} - \mathcal{W}(1) = \underbrace{\frac{s_{\text{bep}}}{2r}\int_{\sigma_{\text{bep}}}^{\sigma_m} \sigma\, dG(\sigma)}_{\text{Participation loss}} + \underbrace{\frac{s_{\text{bep}} - s_m}{2r}\int_{\sigma_m}^{\bar{\sigma}} \sigma\, dG(\sigma)}_{\text{Speed loss}} + \underbrace{C(s_{\text{bep}}) - C(s_m)}_{\text{Speed cost differential}}. \tag{22}$$

Analogous expressions can be derived for two or more venues. The decomposition is fairly intuitive. Keeping the efficient technology s_{bep} constant, imperfect competition distorts the lowest active type from σ_{bep} to σ_m. This generates a *limited participation* welfare loss. In turn, for any given level of market participation, the market equilibrium is less efficient in reallocating the asset from low to high types using the suboptimal technology level

[29] Remember that s is given by $\rho/(\rho + r + \gamma)$, so a similar ratio s_2/s_1 across assets does not imply similar ρ_2/ρ_1 ratios. The ratio s_2/s_1 lies in between 1.5 and 2 for all assets, whereas ρ_2/ρ_1 ranges from a lower bound of roughly 24 for corporate bonds to over 300 for S&P500 index futures.

$s_m < s_{\text{bep}}$. This generates a *trading speed* welfare loss. The total effect of the distortion in technology choices also depends on speeds costs. Furthermore, the participation and speed losses can be computed for the planner relative to Walrasian frictionless benchmark in analogous fashion.

Table VI shows the calibrated value of the sources of welfare loss for the planner, monopoly, and duopoly outcomes. The planner participation loss is minimal as it is only due to the need to finance the speed investment. For the monopoly, however, it is substantial, in the order of 25% of the efficient welfare level. Competition among venues dramatically decreases the participation loss. In the duopoly case, its biggest value is only 1.64%.

Let us now consider speed losses. The misallocation loss for the planner, which reflects the cost of the trading technology, is the only meaningful loss of welfare. Its value ranges from a maximum of 2.53% of the frictionless gains from trade for corporate bonds to only 0.177% in the case of the large index futures markets (for which c is small). The misallocation loss of the monopoly is relatively modest in value. Interestingly, however, the latter represents the bulk of the welfare losses when there is competition among venues. The speed loss is between four and five times greater than the participation loss even with only two venues. This is because the duopoly equilibrium forces venues to decrease fees significantly, allowing near-efficient total participation, but it allocates nearly one-third of the active investors to the slow speed venue. Note that the speed cost differential has different signs for the planner and market outcomes. The positive sign for the planner is simply a result of considering a Walrasian benchmark for which costs are zero. The negative sign in the case of the monopoly is a consequence of $s_m < s_{\text{bep}}$ and for the duopoly is a feature of the calibration outcomes displaying $\sum_{i=1,2} C(s_i) < C(s_{\text{bep}})$.

8.3. *Is Price Protection Socially Desirable?*

Price fragmentation affects not only traders but competition among venues and, ultimately, welfare. If a regulator can enforce price integration, it can be interpreted as the outcome of a price protection rule (see the discussion in Sections A and 5.3). There are three different cases to consider to understand the welfare consequences of such policy, but two have already been analyzed. When entry costs κ are larger than π_1^{int}, only one venue can enter, making the policy irrelevant. When $\pi_1^{\text{seg}} < \kappa \leq \pi_1^{\text{int}}$, price protection increases the number of venues (see Proposition 8) and we have seen in Section 8.2 that this has a large positive effect on welfare. Our goal in this section is to quantify the potential consequences of price protection when $\kappa \leq \pi_1^{\text{seg}}$ so that it does not affect the entry game, but it distorts competition among venues.

Table VII presents our estimates of the welfare cost of these distortions setting $\overline{a} = 0.45$ and, as in Table V, displaying values that are relative to the planner's. To facilitate the connections with the propositions in Section 5, we keep the same speeds regardless of price protection (endogenizing speeds has a second-order effect relative to the fee distortions). Participation at time 0 includes all the traders. Over time, the light traders drop out and, at time ∞, only the heavy traders remain. This process, however, is the same in the planner's allocation so, although fewer traders are active in the market solution, the participation ratios reported in the Segmented panel of Table VII do not change.[30]

[30]Participation at time $t = 0$ represents total investor affiliation (which drives venue revenue). For the integrated case is $(2\bar{a}G(\sigma_2) - G(\sigma_1) + 1 - 2\bar{a})$ for venue 1 and $1 - G(\sigma_2)$ for venue 2. At time $t = \infty$, participation in venue 1 is $G(\sigma_2) - G(\sigma_1)$. Participation at time $t = 0$ in the segmented case is $\frac{1}{2\bar{a}}(G(\sigma_2) - G(\sigma_1))$ and

TABLE VI

SOURCES OF WELFARE LOSS[a]

	Corporate Bonds				Stocks				S&P500 Futures			
	Partic. Loss	Speed Loss	Cost Diff.	Total Loss	Partic. Loss	Speed Loss	Cost Diff.	Total Loss	Partic. Loss	Speed Loss	Cost Diff.	Total Loss
I. $\frac{2}{3}n_k$	$\gamma = 1.378$, $c = 0.0364$				$\gamma = 299.72$, $c = 0.000160$				$\gamma = 586.93$, $c = 0.00275$			
Planner	0.016	2.53	2.435	4.981	0.002	0.927	0.914	1.843	0.000	0.353	0.351	0.705
Monopoly	25.63	0.809	−0.758	25.68	25.23	0.291	−0.274	25.25	25.1	0.11	−0.104	25.1
Duopoly	1.644	7.861	−0.5345	8.971	1.583	7.125	−0.209	8.498	1.57	6.9	−0.082	8.39
II. n_k	$\gamma = 0.834$, $c = 0.0362$				$\gamma = 182.95$, $c = 0.000157$				$\gamma = 390.63$, $c = 0.00275$			
Planner	0.006	1.598	1.560	3.164	0.000	0.584	0.579	1.16	0.000	0.235	0.235	0.47
Monopoly	25.4	0.506	−0.474	25.43	25.15	0.183	−0.172	25.2	25.1	0.0733	−0.069	25.1
Duopoly	1.605	7.414	−0.350	8.669	1.574	6.99	−0.134	8.43	1.57	6.86	−0.055	8.37
III. $\frac{4}{3}n_k$	$\gamma = 0.588$, $c = 0.0364$				$\gamma = 139.64$, $c = 0.000156$				$\gamma = 292.73$, $c = 0.00275$			
Planner	0.003	1.164	1.144	2.311	0.000	0.456	0.453	0.909	0.000	0.177	0.176	0.353
Monopoly	25.29	0.366	−0.3445	25.31	25.1	0.142	−0.134	25.1	25	0.055	−0.518	25
Duopoly	1.59	7.224	−0.2602	8.554	1.57	6.94	−0.105	8.4	1.57	6.83	−0.413	8.36

[a]Friction values for the planner case are normalized relative to the (frictionless) Walrasian outcome. Friction values for the monopoly and duopoly are normalized relative to the constrained planner values.

TABLE VII

SEGMENTED AND PROTECTED EQUILIBRIA OUTCOMES (PLANNER CASE=100)[a]

	Corporate Bonds					Stocks					S&P500 Futures				
	Participation					Participation					Participation				
Time	$t=0$	∞	Π	$\mathcal{V}$	$\mathcal{W}$	$t=0$	∞	Π	$\mathcal{V}$	$\mathcal{W}$	$t=0$	∞	Π	$\mathcal{V}$	$\mathcal{W}$
Segmented															
Venue 1	36.26	36.26	10.48	18.44	11.18	36.11	36.11	10.37	18.52	11.18	36.05	36.05	10.32	18.58	11.16
Venue 2	60.07	60.07	59.04	53.80	78.15	59.83	59.83	59.36	53.75	78.43	59.74	59.74	59.53	53.73	78.51
Duopoly	96.33	96.33	69.52	72.24	89.33	95.94	95.94	69.73	72.27	89.61	95.79	95.79	69.86	72.25	89.67
Integrated															
Venue 1	35.36	28.75	11.17	14.62	8.25	35.21	28.62	11.06	14.68	8.26	35.15	28.58	11.01	14.68	8.24
Venue 2	59.52	66.13	59.28	59.22	82.92	59.28	65.87	59.6	59.17	83.17	59.18	65.76	59.77	59.14	83.24
Duopoly	94.88	94.88	70.45	73.84	91.17	94.49	94.49	70.66	73.85	91.43	94.34	94.34	70.78	73.83	91.48

[a]The parameter values are the same as in Panel II of Table IV, except for $\overline{a}$, which equals 0.45 here. Participation at $t=0$ includes both light and heavy investors and is equal to total affiliation. Participation at $t=\infty$ includes only those investors who trade in the steady state (types $\sigma \geq \sigma_i$ in venue i). The terms $\mathcal{V}$ and $\mathcal{W}$ denote trading volume and welfare in the steady state. Profits (Π) are normalized using 100 for the monopolist in the baseline specification.

For the bond market, for instance, participation is always 96.3% that of the constrained efficient participation. Price protection, in turn, affects both total participation and the distribution of investors across venues. We observe that $\sigma_{1,\text{int}} > \sigma_{1,\text{seg}}$ as in Proposition 4 and that the relative affiliation to venue 1 increases under price protection, but $\sigma_{2,\text{int}} < \sigma_{2,\text{seg}}$ so affiliation to the fast venue decreases. However, as assets migrate from the slow venue to the fast venue, as described in Appendix D, the proportion of active investors in the fast venue increases and in the steady state is nearly 10% higher than in under segmentation. The total drop in investor participation is about 1.45%. The effects of price protection on equilibrium profits of the slow venue are even larger: For the three asset classes considered, we find that its profits increase by more than 6.5%. Table VII also shows that the welfare impact of price protection (around 1.8%) is low relative to the welfare impact of market structure changes in Table V.[31]

8.4. *Speed and Welfare*

Table VIII analyzes the welfare consequences of speed regulations. Panel I reviews the outcomes with baseline parameters and n_k investors. Panel II shows the effect of a 50% reduction in the cost of speed parameter c. Speed increases dramatically in the fast venue but barely moves in the slow venue. Welfare increases, but only slightly. For corporate bonds, welfare increases by 42 basis points, from 91.03% to 91.45%. For stocks, the welfare gains are 10 basis points. The important aspect is that welfare benefits are small, even for asset classes that are initially slow. Even when the cost of speed decreases by 90%, welfare gains are less than 1%, and the first venue barely accelerates, while the fast venue selects a speed that is several times as fast.

Panel III of Table VIII, on the other hand, shows the effect of enforcing a minimum speed requirement, $\underline{\rho}$, that is 50% higher than ρ_1 in the unregulated equilibrium ($\underline{\rho} = 1.5\rho_1$). The increase in welfare is much more significant in this case: 275 basis points for bonds and near 240 basis points for stocks and futures. Forcing the slow stock venue to reduce trading delays from 2 minutes to around 1 minute increases welfare twenty times more than what is achieved by a 50% decrease in the cost of speed. Competition and participation explain much of the welfare gains, while better asset allocation is less important. Table VIII also shows that the reduction in profits disproportionately affects the fast venue.

9. DISCUSSION OF IMPLICATIONS

Let us briefly summarize our key findings in relation to industry facts and regulatory debates. On the normative side, we clarify the circumstances under which competition, fragmentation, and speed improve or reduce welfare. We also comment on the model limitations and highlight opportunities for subsequent work.

$\frac{1}{2\overline{a}}(1 - G(\sigma_2))$ for venues 1 and 2. At time $t = \infty$, participation is $1 - G(\sigma_2)$ and $G(\sigma_2) - G(\sigma_1)$ for venues 1 and 2. These terms represent the same fraction of the constrained planner case for all t. The welfare expressions for the segmentation case are given in Lemma 4 and those for the integration case are given in Appendix C.2.

[31]The social value of price protection technically depends on the value of $\overline{a}$ (0.45 here), but welfare differences remain small in any case.

TABLE VIII

SPEED COST, SPEED REGULATION, AND SOCIAL OUTCOMES (PLANNER CASE = 100)[a]

	Corporate Bonds					Stocks					S&P500 Futures				
	ρ	Π	$\mathcal{P}$	$\mathcal{V}$	$\mathcal{W}$	ρ	Π	$\mathcal{P}$	$\mathcal{V}$	$\mathcal{W}$	ρ	Π	$\mathcal{P}$	$\mathcal{V}$	$\mathcal{W}$
I. Baseline	$\gamma = 0.834$, $c = 0.0362$					$\gamma = 182.95$, $c = 0.000157$					$\gamma = 390.63$, $c = 0.00275$				
Monopoly	36.211	100.00	50.64	50.11	74.32	21,986	100.00	50.23	50.04	74.75	117,000	100.00	50.09	50.02	74.91
Venue 1	1.044	8.47	29.46	16.45	9.01	239.13	8.34	29.29	16.65	9.06	516.93	8.35	29.22	16.67	9.04
Venue 2	38.132	57.68	58.93	58.46	82.02	23,758	58.01	58.58	58.41	82.44	126,402	58.22	58.43	58.37	82.57
Duopoly	–	66.15	88.39	74.91	91.03	–	66.41	87.87	75.07	91.50	–	66.58	87.65	75.04	91.61
II. $c \downarrow$	$\gamma = 0.834$, $c = \frac{1}{2}0.0362$					$\gamma = 182.95$, $c = \frac{1}{2}0.000157$					$\gamma = 390.63$, $c = \frac{1}{2}0.00275$				
Monopoly	51.555	101.4	50.28	50.05	74.07	31,169	100.5	50.10	50.02	74.89	165,625	100.2	50.04	50.01	75.02
Venue 1	1.066	8.56	29.32	16.64	9.06	240.6	8.42	29.22	16.67	9.05	518.11	8.37	29.19	16.67	9.04
Venue 2	55.719	58.58	58.63	58.43	82.39	33,677	58.38	58.45	58.37	82.56	178,924	58.34	58.38	58.35	82.61
Duopoly	–	67.14	87.95	75.07	91.45	–	66.80	87.67	75.04	91.6	–	66.71	87.57	75.02	91.64
III. $\underline{\rho} \uparrow$	$\gamma = 0.834$, $c = 0.0362$					$\gamma = 182.95$, $c = 0.000157$					$\gamma = 390.63$, $c = 0.00275$				
Venue 1	1.565	8.11	30.23	20.05	10.06	358.69	8.06	30.09	20.05	10.04	775.40	8.02	30.04	20.02	10.02
Venue 2	40.538	46.74	60.47	60.21	83.71	24,587	47.45	60.18	60.09	83.91	130,767	47.77	60.08	60.04	83.96
Duopoly	–	54.85	90.70	80.26	93.78	–	55.51	90.28	80.14	93.95	–	55.79	90.11	80.06	93.98

[a]Venue speeds and profits are given by ρ and Π, respectively. Profits are normalized using 100 for the monopolist in the baseline specification. The terms $\mathcal{P}$, $\mathcal{V}$, and $\mathcal{W}$ denote participation, trading volume, and welfare, respectively, and are normalized using 100 for the break-even planner.

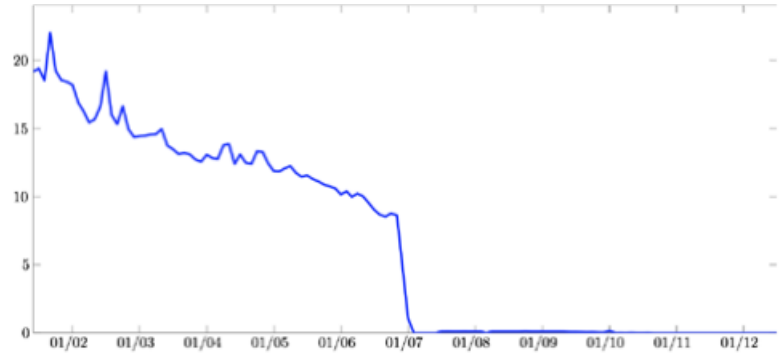

FIGURE 5.—Average small order execution time for the NYSE, executed at the top of the book (seconds). Source: SEC Rule 605 reports.

Speed

We find that, barring front-running issues, it is not optimal to limit venue speed. In fact, market-lead speed levels are lower than what the regulator finds optimal. This fact does not imply, however, that there is much to gain from purely technological improvements in trading speeds for asset classes with high penetration of electronic trading. Perhaps one of the most striking results of our quantitative analysis is that a reduction in the cost of speed leads to a vast increase in speed by the fast venue, almost no increase by the slow venue, and, as a result, modest welfare gains. On the other hand, we find that it can be optimal to increase both speed and competition by pushing the slow venues to upgrade their technology. Slow and inefficient markets, such as that for corporate bonds, could benefit the most from such an intervention. We argue that Reg NMS included an important de facto minimum speed requirement for the NYSE. Given that the Trade-Through rule protects only automated quotes, the NYSE was forced to adopt automation. Figure 5 illustrates this fact: At the time of implementation in 2007, the average execution delay sharply declined from human to machine-driven speeds. The model result may further capture the effects of regulations that push for more electronic trading for OTC derivatives (e.g., the Dodd–Frank Act, the European Market Infrastructure Regulation, and MiFID II).

For tractability purposes, trading venues in the model make entry decisions and choose their speeds once. Although adjusting trading technologies and communication systems is undoubtedly costly, as these costs decrease, venues could, in principle, re-optimize. The comparative statics analysis in Section 8 suggests that changes in c can induce big changes in individual venues' speed. Allocations and welfare, however, are not altered substantially given that the equilibrium driving ratio, s_2/s_1, remains at similar levels for plausible changes in c. If venues faced different speed adjustment costs over time (e.g., if they were also competing in R&D for trading technologies), however, an incumbent slow venue may try to become the fast one when presented with that opportunity. This possibility would likely induce a reallocation of investors among venues and have a higher welfare impact.

Entry

We find that the welfare gains of challenging monopolies are substantial and that, relative to the constrained first best, welfare losses are still significant in a duopoly (8–9%). Entry welfare gains can be lower in cases where fixed entry costs are nontrivial. Beyond

a duopoly, we find that entry by a fast third venue is likely to increase welfare, while entry by a slow venue might not. The case with three venues could capture the (unobserved) welfare consequences of having new venues, such as BATS or Direct Edge, enter and challenge the incumbents (NYSE and NASDAQ) after Reg NMS. The new venues entered with, arguably, better technologies than the incumbents had. Overall, our entry results are relevant for several regulatory agencies around the world that have fostered venue entry, such as the SEC or the European Securities and Markets Authority (ESMA).

Although we consider the possibility of a sequential setting in Section 7, we do not model entry deterrence. There is, of course, an extensive literature that studies this aspect. For instance, Donnenfeld and Weber (1995) considered a vertically differentiated duopoly facing the threat of entry by a third firm. They showed that incumbent firms can deter entry by choosing quality levels that reduce ex post differentiation relative to an unchallenged duopoly. Therefore, entry deterrence may improve welfare, even without actual entry. We conjecture that, everything else being constant, our calculations may overestimate participation-related welfare losses due to this possibility.

Order Price Protection

Price protection affects entry incentives and post-entry competition. Its welfare effects can be substantial when it encourages entry. The effects are ambiguous and more modest when the active number of venues is not affected, as one would expect in markets that are highly fragmented ex ante. These are valuable insights in light of the debate regarding the impact of the SEC's Trade-Through rule on market quality (e.g., as expressed by the Equity Market Structure Advisory Committee) and in other economies that have adopted such rules (e.g., Canada's Order Protection Rule). According to our model, price protection likely had a positive impact on welfare in the United States because it encouraged the entry of new venues. On the other hand, for markets that are already fragmented, as in most of Europe, the adoption of a similar rule may not affect entry and thus have limited value from a welfare perspective. In fact, if, as considered in Section 5 and Appendix B, $\sigma_{2,\text{int}} > \sigma_{2,\text{seg}}$, the welfare impact of price protection can be negative.

Of course, we do not capture welfare gains from mitigating execution price uncertainty, if they exist, given that our traders are risk neutral. On the other hand, the implementation of such rules is arguably costly (we do not model this cost) and its ability to eliminate price fragmentation has been compromised by the proliferation of the so-called make-take fees. Our analysis does not include all of these features, but we provide a consistent framework in which they can be added for evaluation in future work.

10. CONCLUDING REMARKS

We have provided an equilibrium analysis of entry, investment in speed, and fee competition among trading venues that seek to attract traders who are subject to temporary random preference shocks. We have shown that the framework has rich positive and normative implications that help to rationalize evolutions in equities, derivatives, and other capital markets in recent years. The empirical implementation of the model permits a quantitative assessment of changes in both the competitive environment and regulations.

Let us conclude with further ideas for future work. We do not model all possible sources of differentiation. It would be interesting to incorporate alternative trading mechanisms like limit order books (e.g., Biais, Hombert, and Weill (2014)) or RFQ and allow venues to differentiate across this dimension. Also interesting is to incorporate sources of differentiation related to price transparency, such as between lit and dark trading venues,

as well as endogenizing contract offerings between trading venues and liquidity providers (e.g., HFT firms). By introducing such features, one could study the performance of incentive schemes for liquidity creation and stability, a key concern in the regulation debate over designated market makers. In unregulated markets like blockchain-based assets, exchanges can differentiate by offering distinct levels of cyber security or quality of custodial services. Overall, we hope that our model provides others with a tractable benchmark to integrate asset trade dynamics with the determination of the industrial organization in financial markets.

APPENDICES

APPENDIX A: REMARKS ON THE SECURITY EXCHANGE INDUSTRY

A.1. *Supplement to Section 2: Figures and Tables*

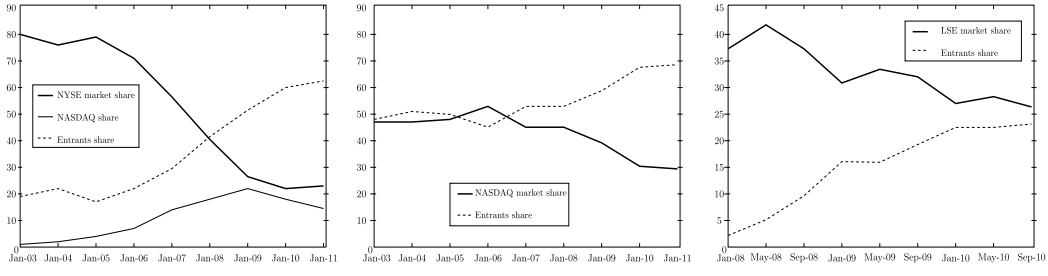

FIGURE A.1.—Equity market volume fragmentation by listing venue: NYSE, NASDAQ, and the London Stock Exchange (LSE). (Source: Barclays Capital Equity Research.)

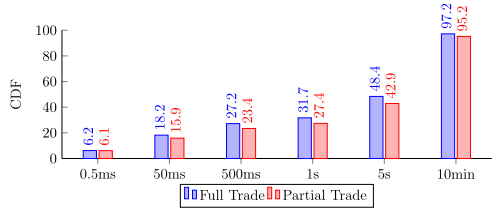

FIGURE A.2.—Distribution of trading speed in U.S. equity markets. (Source: SEC, October 2013.)

A.2. *Slow and Fast Venues: Examples Across Asset Classes*

Corporate Bonds

The corporate bond market has traditionally operated in a decentralized fashion and over the phone (Duffie, Garleanu, and Pedersen (2005)). Since the last financial crisis, institutional investors have begun migrating some of their orders execution away from voice and toward the electronic request for quote protocol (eRFQ).[32] The eRFQ represented

[32]MarketAxxes, a leading venue, introduced the list-based e-RFQ in 2002.

TABLE A.I

SELECTED SPEED INVESTMENTS BY WORLD EXCHANGES (2008–2012)[a]

Exchange	Quarter	Investment	Latency Reduction (as Reported)	Asset Class
NYSE Euronext	Q4 2008	Universal Trading Platform	150–400 microseconds from 1.5 ms	Bonds
	Q1 2009	Universal Trading Platform		Cash Equities
NYSE	Q2 2009	Super Display Book System Platform	5 ms from 105 ms (350 in 2007)	Cash Equities
NYSE Amex	Q3 2009	Super Display Book System Platform	5 ms from 105 ms (350 in 2007)	Cash Equities
NYSE, NYSE Arca, NYSE Amex	Q4 2009	Universal Trading Platform	from 5 to 1.5 milliseconds	Cash Equities
Tokyo Stock Exchange	Q4 2009	Tdex + System	to 6 millisecond	Options
	Q1 2010	Arrowhead Platform	5 millisecond from 2 seconds	Cash Equities
	Q4 2011	Tdex + System	5 milliseconds	Futures
Turquoise (LSE's)	Q4 2009	Millenium Exchange Platform	Latency of 126 microsecond	Derivatives
NASDAQ OMX (Nordic + Baltic)	Q1 2010	INET Platform	to 250 microsec	Cash equities
Johannesburg Stock Exchange	Q1 2011	Millenium Exchange Platform	400 times faster to 126 microsecond	Cash equities
London Stock Exchange	Q4 2010	Millenium Exchange Platform		Cash equities
Singapore Stock Exchange	Q3 2011	Reach Platform		Cash equities
Hong Kong Stock Exchange	Q4 2012	HKEx Orion		Cash equities

[a]Source: Hand collected list from various sources.

TABLE A.II

VENUE COMPETITION AND INVESTOR PROTECTION IN SELECTED COUNTRIES[a]

Economic Area	Reg. Agency	Regulation	Year	Investor Protection Model
USA	SEC	Reg.NMS	2005	Trade-through (top of the book)
Europe	ESMA	MiFID I	2007	Principles-based
Japan	FSA, FIEA	FIEA	2007	Principles-based
Canada	IIROC, CSA	OPR	2011	Trade-through (full book)
South Korea	FSC	FSCMA	2011	Principles-based
Australia	ASIC	MIR	2011	Principles-based

[a]Source: www.fidessa.com and regulating agencies' websites.

an evolutionary step toward efficiency, not a market structure change. Similar to picking up a phone and calling a handful of dealers, it allows investors to gather a pool of potential liquidity providers. This mechanism can then be seen as the electronic version of the status quo. More recently, there has been a proliferation of electronic trading venues, some of them operating a central limit order book as well (CLOB). We list some of them below:

- Slow Venue: Voice trading using traditional dealer banks.
- Fast Venue (mainly eRFQ): Bank-sponsored electronic bond trading networks: GSessions (Goldman Sachs), Bond Pool (Morgan Stanley), Price Improvement Network (UBS), Aladdin Trading Network (BlackRock's), BondPoint (Knight). Bond trading platforms: Bloomberg, MarketAxxes, Tradeweb, Bonds.com.
- Faster Venues (mainly CLOB): ICAP's BrokerTec, GFI, NYSE Bonds.

Despite the recent innovation in trading systems, slow voice trading is still dominant. The TABB Group estimates that, as of 2014, approximately 15%–16% of the notional volume for investor-initiated (otherwise known as dealer-to-client) trading is executed via some electronic medium (approximately 21% if accounting for retail transactions). The scope for further platforms development and growth is illustrated by the fact that near four-fifths of the notional volume are still transacted in the "old fashioned" way. In this market, trading protocols are still evolving and it is still challenging to find liquidity in off-the-run corporate bond issues (which account for most of the market).

Foreign Exchange (FX)

FX is global and trades 24 hours. A large number of financial institutions, individuals, and corporations that are active in this market select to trade in venues with different speeds. We can group venues in two stylized groups.

- Slow Venue: Traditional banks/trading desks acting as voice brokers/dealers, trading at human speeds.
- Fast Venue: Multiple venues operating with different technologies. Inter-dealer electronic brokers platforms (EBS, Reuters, in London); ECNs (such as Currenex), 10–15 single-banks platforms. Trading speed is sub-second.

Despite the rapid growth of electronic venues in the FX market, by the end of 2012 only 60% of the global trading volume was electronic[33] (from 51% in 2010).

An important fraction of market centers' speed investment is in the form of locating trading venues where customers congregate (trading hubs such as Chicago, New York, or London). Thus, a large part of the speed premium that clients pay is in the form of co-location and developing trading infrastructure in multiple cities. The FX market is traditionally highly unregulated and opaque. In particular, there is not a trade-through-like rule protecting execution prices, resulting in a high degree of price fragmentation.

Swaps: IRS, CDS

Until recently, financial institutions and corporations participating in these markets were used to trading at much lower speeds than equities and paying higher commissions. The landscape has been transformed by the strong regulation force of the Dodd–Frank Act, which mandates electronic trading of large classes of derivatives—and the subsequent entry of new venues with modern trading platforms. We can conceptually group venues as follows:

- Slow Venue: OTC broker–dealer (such as UBS, Credit Suisse, and Morgan Stanley) trading over the phone, or traditional RFQ.
- Fast Venue: Inter-dealer electronic platforms as ICAP i-Swaps, Tradition and BCG for IRS, and Bloomberg's BSEF for credit default swaps; several electronic Swap Execution Facilities (SEF).

[33] Reported in Greenwich Associates' Global Foreign Exchange Services Study (2012).

As of April 2014, there were 24 SEFs registered with the CFTC operating across interest rate, credit, and foreign exchange asset classes, but only a handful have a market of more than five percent.

APPENDIX B: PROOFS OF PROPOSITIONS

B.1. *Proof of Proposition 1*

Consider the steady-state value functions for types $\sigma > \sigma^{\mathrm{t}}$. For the types holding the asset, we have

$$\begin{aligned} rV_{\sigma,+}(1) &= \mu + \sigma + \frac{\gamma}{2}\big[V_{\sigma,-}(1) - V_{\sigma,+}(1)\big], \\ rV_{\sigma,-}(1) &= \mu - \sigma + \frac{\gamma}{2}\big[V_{\sigma,+}(1) - V_{\sigma,-}(1)\big] + \rho\big(p + V_{\sigma,-}(0) - V_{\sigma,-}(1)\big). \end{aligned} \tag{23}$$

For those not holding the asset,

$$\begin{aligned} rV_{\sigma,-}(0) &= \frac{\gamma}{2}\big[V_{\sigma,+}(0) - V_{\sigma,-}(0)\big], \\ rV_{\sigma,+}(0) &= \frac{\gamma}{2}\big[V_{\sigma,-}(0) - V_{\sigma,+}(0)\big] + \rho\big(V_{\sigma,+}(1) - V_{\sigma,+}(0) - p\big). \end{aligned} \tag{24}$$

Define $I_{\sigma,\varepsilon} \equiv V_{\sigma,\varepsilon}(1) - V_{\sigma,\varepsilon}(0)$ as the value of owning the asset for type (σ, ε). Then, taking the differences of equations (23) to (24), we obtain

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

We can then solve $r(I_{\sigma,+} - I_{\sigma,-}) = 2\sigma - (\gamma + \rho)(I_{\sigma,+} - I_{\sigma,-})$ and obtain the gains from trade for type σ in venue ρ: $I_{\sigma,+} - I_{\sigma,-} = \frac{2\sigma}{r+\gamma+\rho}$. Therefore, we can compute $I_{\sigma,-} = \frac{\mu+\rho p}{r+\rho} - \frac{\sigma}{r+\gamma+\rho}$ and $I_{\sigma,+} = \frac{\mu+\rho p}{r+\rho} + \frac{\sigma}{r+\gamma+\rho}$. Moreover, the average values are

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

Let us now compute the ex ante value function W. Let us first consider types $\sigma < \sigma^{\mathrm{t}}$ who join the venue, sell at price p, and do not trade again. Averaging over types $\varepsilon = \pm 1$, the ex ante value function $\tilde{W}$ solves the Bellman equation $r\tilde{W} = \mu\bar{a} + \rho(p\bar{a} - \tilde{W})$, and thus $\tilde{W} = \frac{\mu+\rho p}{r+\rho}\bar{a}$. Since $\mu + \rho p = \frac{\mu}{r}(r+\rho) + \rho(p - \frac{\mu}{r})$, we can rewrite $\tilde{W} = \frac{\mu\bar{a}}{r} + \frac{\rho}{r+\rho}(rp - \mu)\frac{\bar{a}}{r}$. From the definition, we also know that $\frac{\rho}{r+\rho}(rp - \mu) = s(\rho)\sigma^{\mathrm{t}}$, with $s(\rho) \equiv \frac{\rho}{r+\gamma+\rho}$; therefore, $\tilde{W} = \frac{\mu\bar{a}}{r} + s\frac{\bar{a}}{r}\sigma^{\mathrm{t}}$. Of course, we also have $\tilde{W} = \bar{a}\bar{V}_{\sigma^{\mathrm{t}}}(1)$. Let us now consider the steady-state

types, $\sigma > \sigma^{\mathrm{t}}$. The ex ante value function is $W(\sigma) = \bar{a}\bar{V}_{\sigma}(1) + (1-\bar{a})\bar{V}_{\sigma}(0)$. Therefore,

$$\begin{aligned} W(\sigma) &= \frac{\bar{a}\mu}{r} + \bar{a}\frac{\rho}{2r}(p - I_{\sigma,-}) + (1-\bar{a})\frac{\rho}{2r}(I_{\sigma,+} - p) \\ &= \frac{\mu\bar{a}}{r} + \bar{a}\frac{\rho}{r}\frac{rp-\mu}{r+\rho} + \frac{1}{2r}\left(\frac{\rho}{r+\rho}(\mu - rp) + \frac{\rho}{r+\gamma+\rho}\sigma\right) \\ &= \frac{\mu\bar{a}}{r} + \frac{\bar{a}}{r}s(\rho)\sigma^{\mathrm{t}} + \frac{1}{2r}s(\rho)\big(\sigma - \sigma^{\mathrm{t}}\big). \end{aligned}$$

Therefore, when $\sigma > \sigma^{\mathrm{t}}$, we have $W(\sigma) = \tilde{W} + \frac{1}{2}s(\rho)\frac{\sigma-\sigma^{\mathrm{t}}}{r}$. Equation (8) simply generalizes the ex ante value expression for any type σ. *Q.E.D.*

B.2. *Proof of Proposition 2*

First notice that $W(\sigma_2^{\mathrm{p}}, \sigma_2^{\mathrm{t}}, s_2) - q_2 = W(\sigma_2^{\mathrm{p}}, \sigma_1^{\mathrm{t}}, s_1) - q_1$ can be written as

$$\frac{s_2\bar{a}\sigma_2^{\mathrm{t}}}{r} + \frac{s_2}{2r}\big(\sigma_2^{\mathrm{p}} - \sigma_2^{\mathrm{t}}\big) - q_2 = \frac{s_1\bar{a}\sigma_1^{\mathrm{t}}}{r} + \frac{s_1}{2r}\big(\sigma_2^{\mathrm{p}} - \sigma_1^{\mathrm{t}}\big) - q_1.$$

Since $q_1 = \frac{\bar{a}s_1\sigma_1^{\mathrm{t}}}{r}$, we get $\frac{s_2-s_1}{2r}\sigma_2^{\mathrm{p}} = q_2 - \frac{\bar{a}s_2\sigma_2^{\mathrm{t}}}{r} + \frac{s_2\sigma_2^{\mathrm{t}} - s_1\sigma_1^{\mathrm{t}}}{2r}$. Using $\sigma_2^{\mathrm{t}} = x\sigma_1^{\mathrm{t}}$, we get $\frac{s_2-s_1}{2r}\sigma_2^{\mathrm{p}} = q_2 - q_1(\frac{1}{2\bar{a}} - \frac{s_2}{s_1}x(\frac{1}{2\bar{a}} - 1))$, where $x \equiv \frac{1+\frac{\gamma}{r+\rho_2}}{1+\frac{\gamma}{r+\rho_1}}$. Since $\frac{s_2}{s_1}x = \frac{\rho_2}{\rho_1}\frac{r+\rho_1}{r+\rho_2}$, we get $\sigma_{2,\mathrm{int}}^{\mathrm{p}} = \frac{2r}{s_2-s_1}(q_2 - \frac{z}{2\bar{a}}q_1)$, where $z \equiv 1 - \frac{1+\frac{r}{\rho_1}}{1+\frac{r}{\rho_2}}(1-2\bar{a})$. Note that z is an increasing function of $\bar{a}$ satisfies $z \leq 1$. When $a \approx 0.5$, we have $z \approx 1$, and $z \approx 2\bar{a}$ when r/ρ is small (the realistic case). The profits of venue 1 are $\pi_1^{\mathrm{int}} = q_1(G(\sigma_2) - G(\sigma_1) + \ell_1)$. Therefore, in the integrated price case, venues simultaneously solve $\max_{q_1} \pi_1^{\mathrm{int}} = \frac{q_1}{2\bar{a}}(1 - 2\bar{a} + 2\bar{a}G(\sigma_2) - G(\sigma_1))$ and $\max_{q_2} \pi_2^{\mathrm{int}} = q_2(1 - G(\sigma_2))$. Developing the first-order conditions $\frac{\partial\pi_1^{\mathrm{int}}}{\partial q_1} = 0$ and $\frac{\partial\pi_2^{\mathrm{int}}}{\partial q_2} = 0$ leads to the system of equations (16) and (17). *Q.E.D.*

B.3. *Proof of Proposition 3*

The objective function of the break-even planner is

$$\max_{s_2,q_1,q_2} \frac{s_1}{2r}\int_{\sigma_1}^{\sigma_2} \sigma\, dG(\sigma) + \frac{s_2}{2r}\int_{\sigma_2}^{\overline{\sigma}} \sigma\, dG(\sigma) - C(s_2)$$

and the marginal types are given by (12) and (13), so we have $q_1 = s_1\frac{\sigma_1}{2r}$ and $q_2 = (s_2 - s_1)\frac{\sigma_2}{2r} + q_1$. The break-even constraint is $q_2(1 - G(\sigma_2)) \geq C(s_2)$, so the Lagrangian (scaled by $2r$) is

$$\begin{aligned} \mathcal{L} = {} & s_1\int_{\sigma_1}^{\overline{\sigma}} \sigma\, dG(\sigma) + (s - s_1)\int_{\sigma_2}^{\overline{\sigma}} \sigma\, dG(\sigma) - 2rC(s) \\ & + \lambda\big\{\big((s-s_1)\sigma_2 + s_1\sigma_1\big)\big(1 - G(\sigma_2)\big) - 2rC(s)\big\}. \end{aligned}$$

The first-order conditions of the break-even planner problem are

$$\sigma_1^{\mathrm{bep}} g\big(\sigma_1^{\mathrm{bep}}\big) = \lambda\big(1 - G\big(\sigma_2^{\mathrm{bep}}\big)\big),$$

$$\sigma_2^{\text{bep}} g\big(\sigma_2^{\text{bep}}\big) = \frac{\lambda}{1+\lambda}\left(1 - G\big(\sigma_2^{\text{bep}}\big) - \frac{s_1}{s^{\text{bep}} - s_1} g\big(\sigma_2^{\text{bep}}\big)\sigma_1^{\text{bep}}\right),$$

$$2rC'\big(s^{\text{bep}}\big) = \frac{1}{1+\lambda}\int_{\sigma_2^{\text{bep}}}^{\overline{\sigma}} \sigma\, dG(\sigma) + \frac{\lambda}{1+\lambda}\big(1 - G\big(\sigma_2^{\text{bep}}\big)\big)\sigma_2^{\text{bep}},$$

and the break-even constraint is simply $2rC(s^{\text{bep}}) = (1-G(\sigma_2^{\text{bep}}))((s^{\text{bep}}-s_1)\sigma_2^{\text{bep}} + s_1\sigma_1^{\text{bep}})$. From the first two conditions, it is immediate that $\sigma_1^{\text{bep}} g(\sigma_1^{\text{bep}}) > \sigma_2^{\text{bep}} g(\sigma_2^{\text{bep}})$. From the second-order conditions, we know that $\sigma g(\sigma)$ is increasing in σ at the optimum. Therefore, $\sigma_1^{\text{bep}} > \sigma_2^{\text{bep}}$, which is inconsistent with our assumption that venue 1 is active. We conclude that there must be a single venue.

This result can be extended to the case where the planner operates the two venues with one budget constraint. In this case, the constraint is $(G(\sigma_2) - G(\sigma_1))q_1 + (1-G(\sigma_2))q_2 > C(s_2)$ and the Lagrangian is

$$\mathcal{L} = s_1\int_{\sigma_1}^{\overline{\sigma}} \sigma\, dG(\sigma) + (s-s_1)\int_{\sigma_2}^{\overline{\sigma}} \sigma\, dG(\sigma) - 2rC(s)$$
$$+ \lambda\big(\big(1-G(\sigma_1)\big)s_1\sigma_1 + \big(1-G(\sigma_2)\big)(s-s_1)\sigma_2 - 2rC(s)\big),$$

and the first-order conditions for affiliations are $1 - G(\sigma_1^{\text{bep}}) = g(\sigma_1^{\text{bep}})\frac{1+\lambda}{\lambda}\sigma_1^{\text{bep}}$ and $1 - G(\sigma_2^{\text{bep}}) = g(\sigma_2^{\text{bep}})\frac{1+\lambda}{\lambda}\sigma_2^{\text{bep}}$. The optimal speed satisfies the same equation as before. In this case, we see that $\sigma_1^{\text{bep}} = \sigma_2^{\text{bep}}$, so venue 1 is still inactive.

With one active venue, the Lagrangian of the planner is

$$\mathcal{L} = s\int_{\sigma^{\text{p}}}^{\overline{\sigma}} \sigma\, dG(\sigma) - 2rC(s) + \lambda\big(s\sigma^{\text{p}}\big(1 - G\big(\sigma^{\text{p}}\big)\big) - 2rC(s)\big).$$

From the previous analysis, it is immediate that $1 - G(\sigma^{\text{bep}}) = g(\sigma^{\text{bep}})\frac{1+\lambda}{\lambda}\sigma_1^{\text{bep}}$. Since the monopoly solution is $\frac{1-G(\sigma_m)}{g(\sigma_m)} = \sigma_m$, it is clear that $\sigma_m > \sigma^{\text{bep}}$. Regarding speed, the planner solution satisfies

$$2rC'\big(s^{\text{bep}}\big) = \frac{1}{1+\lambda}\int_{\sigma^{\text{bep}}}^{\overline{\sigma}} \sigma\, dG(\sigma) + \frac{\lambda}{1+\lambda}\big(1 - G\big(\sigma^{\text{bep}}\big)\big)\sigma^{\text{bep}},$$

while the monopoly chooses $2r\frac{\partial C}{\partial s}(s_m) = (1 - G(\sigma_m))\sigma_m$. If $\lambda = 0$, it is clear that $s^{\text{bep}} > s_m$, as expected. However, when the break-even constraint binds, the relation is ambiguous. *Q.E.D.*

B.4. *Proof of Proposition 4*

We prove points (i) and (ii) analytically under Assumption 1, and point (iii) for the cases of uniform distribution of types. The proof of part (iii) with an exponential distribution is available upon request.

Proof of Point (i)

Let us introduce some notations to simplify the exposition: $\alpha \equiv 2\bar{a}$, $k \equiv \frac{s_1}{s_2 - s_1}$, $\nu(\sigma) \equiv \frac{1-G(\sigma)}{g(\sigma)}$. Rearranging the first-order conditions, the marginal participating types under seg-

mentation $(\sigma_{1,\text{seg}}, \sigma_{2,\text{seg}})$ are the solution to

$$\sigma_2 = \nu(\sigma_2) - k\sigma_1,$$
$$\sigma_1\left(\frac{g(\sigma_1)}{g(\sigma_2)} + k\right) = \frac{g(\sigma_1)}{g(\sigma_2)}\nu(\sigma_1) - \nu(\sigma_2).$$

The integrated market types $(\sigma_{1,\text{int}}, \sigma_{2,\text{int}})$ are the solution to

$$\sigma_2 = \nu(\sigma_2) - z(\alpha)k\sigma_1,$$
$$\sigma_1\left(\frac{g(\sigma_1)}{g(\sigma_2)} + \alpha z(\alpha)k\right) = \frac{g(\sigma_1)}{g(\sigma_2)}\nu(\sigma_1) - \alpha\nu(\sigma_2).$$

Note that $z \equiv 1 - \frac{1+\frac{r}{\rho_1}}{1+\frac{r}{\rho_2}}(1-\alpha)$ is increasing in α. Note also that, since the monopoly marginal participating type is the solution to $\sigma_m = \nu(\sigma_m)$, we have $\sigma_2 < \sigma_m$ irrespective of price fragmentation. Participation in venue 2 alone is higher than under monopoly.

Proof of Point (ii)

We use the following notations to simplify the algebra $x \equiv \sigma_1$, $y \equiv \sigma_2$. The duopoly system with segmented prices is

$$G(\sigma_2) - G(\sigma_1) = \big(g(\sigma_1) + kg(\sigma_2)\big)\sigma_1,$$
$$1 - G(\sigma_2) = g(\sigma_2)(\sigma_2 + k\sigma_1).$$

We can differentiate the system with respect to k:

$$g(\sigma_2)\,d\sigma_2 - g(\sigma_1)\,d\sigma_1 = g(\sigma_1)\,d\sigma_1 + kg(\sigma_2)\,d\sigma_1 + \sigma_1\big(g'(\sigma_1)\,d\sigma_1 + g(\sigma_2)\,dk + kg'(\sigma_2)\,d\sigma_2\big),$$
$$-g(\sigma_2)\,d\sigma_2 = g'(\sigma_2)\,d\sigma_2(\sigma_2 + k\sigma_1) + g(\sigma_2)\big(d\sigma_2 + d[k\sigma_1]\big).$$

After some manipulations and simplifications, we get

$$d\sigma_1\big(2g(\sigma_1) + \sigma_1 g'(\sigma_1) + kg(\sigma_2)\theta\big) = -\sigma_1 g(\sigma_2)\theta\,dk,$$
$$d\sigma_2 = -g(\sigma_2)\frac{d[k\sigma_1]}{2g(\sigma_2) + g'(\sigma_2)(\sigma_2 + k\sigma_1)},$$

where $\theta \equiv \frac{3g(\sigma_2)+\sigma_2 g'(\sigma_2)}{2g(\sigma_2)+g'(\sigma_2)(\sigma_2+k\sigma_1)}$. We can then prove that under Assumption 1, we have $\frac{\partial\sigma_2}{\partial k} < 0$ and $\frac{\partial\sigma_1}{\partial k} < 0$. Note that venue 2 first-order condition implies that $\frac{1-G(\sigma_2)}{g(\sigma_2)} = \sigma_2 + k\sigma_1$. Therefore, under Assumption 1, we have $2g(\sigma_2) + g'(\sigma_2)(\sigma_2 + k\sigma_1) \geq 0$. This shows that the denominator in θ is strictly positive. Let us study the numerator and show that it is also strictly positive. Either $g'(\sigma_2) > 0$ and then $3g(\sigma_2) + \sigma_2 g'(\sigma_2) > 0$; or $g'(\sigma_2) < 0$, but then, since $k\sigma_1 > 0$,

$$3g(\sigma_2) + \sigma_2 g'(\sigma_2) > 2g(\sigma_2) + \sigma_2 g'(\sigma_2) > 2g(\sigma_2) + (\sigma_2 + k\sigma_1)g'(\sigma_2) > 0.$$

Therefore, $\theta > 0$. It is then easy to see that $\frac{\partial\sigma_1}{\partial k} < 0$. We also have $\frac{\partial\sigma_2}{\partial k} < 0$ since $\partial\sigma_1\left(\frac{2g(\sigma_1)+\sigma_1 g'(\sigma_1)}{g(\sigma_2)\theta} + k\right) + \sigma_1\partial k = 0$, $\sigma_1 > 0$, and $2g(\sigma_1) + \sigma_1 g'(\sigma_1) > 0$ under Assumption 1.

Proof of Point (iii)

Here it is convenient to define speed differentiation as $d \equiv \frac{s_2}{s_1}$. With a uniform distribution over $[0, \overline{\sigma}]$, the integration first-order conditions become $2\sigma_2 = \overline{\sigma} - \frac{z(\alpha)\sigma_1}{d-1}$ and $\sigma_1(2 + \frac{\alpha z(\alpha)}{d-1}) = (1-\alpha)\overline{\sigma} + \alpha\sigma_2$. Thus, $\sigma_2 = \frac{\overline{\sigma}(d-1) - z(\alpha)\sigma_1}{2(d-1)}$ and $\sigma_1 = \frac{(1-\frac{\alpha}{2})(d-1)}{2(d-1)+\frac{3}{2}\alpha z(\alpha)}\overline{\sigma}$. The solution implies that σ_1 is decreasing in α. We obtain the segmentation equations when $\alpha = 1$ and the integration equations when $\alpha < 1$ so σ_1 goes up under integration. The impact on σ_2 is ambiguous since $\sigma_2 = [1 - z(\alpha)\frac{1-\frac{\alpha}{2}}{2(d-1)+\frac{3}{2}\alpha z(\alpha)}]\frac{\overline{\sigma}}{2}$ and $z(\alpha) = 1 - \frac{1+\frac{r}{\rho_1}}{1+\frac{r}{\rho_2}}(1-\alpha)$. Clearly, if α is small, then σ_2 decreases with α. But if α and d are both close to 1, this can be reversed.

Comparing Profits

It is convenient to define a system that nests integration and segmentation as special cases. First, define the scaled controls, $t_1 \equiv \frac{2r}{\alpha s_1}q_1$ and $t_2 \equiv \frac{2r}{s_1}q_2$; and the scaled profits, $F_i \equiv \frac{2r}{s_1}\pi_i$. With these notations, the revenue functions are

$$F_1(t_1, t_2, \alpha) = t_1\big(1 - \alpha + \alpha G(\sigma_2) - G(t_1)\big),$$
$$F_2(t_1, t_2, \alpha) = t_2\big(1 - G(\sigma_2)\big),$$

and we have $\sigma_2 = k(t_2 - z(\alpha)t_1)$ and $\sigma_1 = t_1$.

The general system is the one with integrated prices with $\alpha < 1$ and $z(\alpha) < 1$ as explained above. Let us now derive the first-order conditions. Using $\frac{\partial \pi_1^{\text{int}}}{\partial t_1} = 0$ and $\frac{\partial \pi_2^{\text{int}}}{\partial t_2} = 0$, we get

$$1 - \alpha + \alpha G(\sigma_2) - G(\sigma_1) = t_1\big(\alpha z(\alpha) k g(\sigma_2) + g(\sigma_1)\big),$$
$$1 - G(\sigma_2) = t_2 k g(\sigma_2).$$

With a uniform distribution, we have $\sigma_2 = \frac{\overline{\sigma} - z(\alpha)k\sigma_1}{2}$ and $\sigma_1 = \frac{1-\frac{\alpha}{2}}{2+\frac{3}{2}\alpha z(\alpha)k}\overline{\sigma}$. Thus, $F_1 = \sigma_1(1 - \alpha + \alpha\frac{\sigma_2}{\overline{\sigma}} - \frac{\sigma_1}{\overline{\sigma}})$ and $F_2 = (\frac{\sigma_2}{k} + z\sigma_1)(1 - \frac{\sigma_2}{\overline{\sigma}})$, which implies that

$$F_1 = \frac{1}{2 + \frac{3}{2}\alpha z(\alpha)k}\left(1 - \frac{\alpha}{2}\right)^2\left(1 - \frac{1}{2}\frac{2 + \alpha z(\alpha)k}{2 + \frac{3}{2}\alpha z(\alpha)k}\right)\overline{\sigma},$$

$$F_2 = \left(\frac{1}{k} + \frac{z}{2}\frac{1 - \frac{\alpha}{2}}{2 + \frac{3}{2}\alpha z(\alpha)k}\right)\left(\frac{1}{2} + \frac{z(\alpha)k}{2}\frac{1 - \frac{\alpha}{2}}{2 + \frac{3}{2}\alpha z(\alpha)k}\right)\overline{\sigma}.$$

So it is easy to see that $\frac{\partial F_1}{\partial \alpha} < 0$. The impact of α on F_2 is ambiguous, however. *Q.E.D.*

B.5. *Proof of Proposition 5*

The first-order condition for speed is $C'(s) = (1 - G(\sigma_m))\frac{\sigma_m}{2r}$. Using Assumption 2, we have $C'(s) = \frac{c(r+\gamma)}{(1-s)^2}$. Under exponential distribution of types, we have $\sigma_m = \nu$ and thus $(1 - G(\sigma_m))\frac{\sigma_m}{2r} = \frac{\nu}{2er}$. Combining these expressions yields $s_m = 1 - (2rc(\gamma + r)e/\nu)^{\frac{1}{2}}$. Under

uniform distribution of types, we have $\sigma_m = \overline{\sigma}/2$ and $(1 - G(\sigma_m))\frac{\sigma_m}{2r} = \frac{\overline{\sigma}}{8r}$. Thus, $s_m = 1 - (8rc(r+\gamma)/\overline{\sigma})^{\frac{1}{2}}$. *Q.E.D.*

B.6. *Proof of Proposition 6*

Given the profit functions $\Pi_i(s_i, s_j)$, let us first characterize the speed choices that satisfy the first-order conditions in the duopoly case:[34]

$$\big(G(\sigma_2) - G(\sigma_1)\big)\frac{\partial q_1}{\partial s_1} + q_1\left(g(\sigma_2)\frac{\partial \sigma_2}{\partial s_1} - g(\sigma_1)\frac{\partial \sigma_1}{\partial s_1}\right) = \frac{\partial C}{\partial s}\big(s_1^*\big), \tag{25}$$

$$\frac{\partial q_2}{\partial s_2}\big(1 - G(\sigma_2)\big) - \frac{\partial \sigma_2}{\partial s_2}g(\sigma_2)q_2 = \frac{\partial C}{\partial s}\big(s_2^*\big). \tag{26}$$

The solution to the system of equations (25) and (26) implicitly characterizes a function $S_i(s_j)$ that represents the best response of venue i to s_j. Figure B.1 displays the speed choices. The 45-degree line represents the case in which there is no product differentiation, which would lead to Bertrand competition and would be inconsistent with entry by both venues for any arbitrarily small entry cost. The actual equilibrium satisfying equations (25) and (26) is at point (s_1^*, s_2^*) where the best response functions intersect. In this equilibrium, there is a fast venue and a slow venue.

Now consider venue 2's program: $\max_{s_2} \frac{1}{2r}(\sigma_2(s_2 - s_1) + \sigma_1 s_1)(1 - G(\sigma_2)) - C(s_2)$. It is immediate that this program converges to the monopolist's when $s_1 \to 0$. We then have $\lim_{s_1 \to 0} S_2(s_1) = s_m$. Thus, to show that $s_2 > s_m$, it suffices to show that $S_2'(s_1) > 0$. Differ-

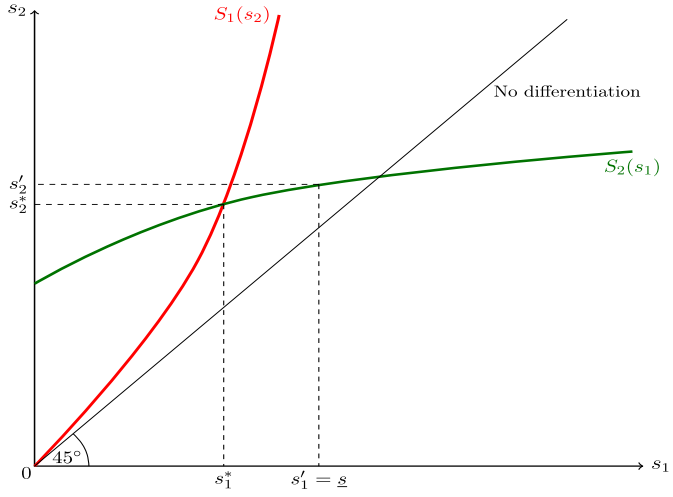

FIGURE B.1.—Regulation of speed and venue differentiation. The function $S_i(s_j)$ denotes venue i's best speed response to s_i; s_1^* and s_2^* are the (unregulated) equilibrium speed choices, $\underline{s}$ represents the minimum speed that the regulator may want to impose, and s_1' and s_2' are the optimal speed choices when $\underline{s} > s_1^*$.

[34]To the best of our knowledge, the literature does not offer an existence result for the first stage of competition in vertically differentiated oligopolies. In Section 8, we verify numerically that, for any parameter set, first- and second-order conditions are satisfied.

entiating venue's 2 profit with respect to s_1, and rearranging, yields

$$S_2'(s_1) = \frac{\partial^2 \pi_2}{\partial s_2 \partial s_1}\left(C''(s_2) - \frac{\partial^2 \pi_2}{\partial s_2^2}\right)^{-1}. \tag{27}$$

We now use the uniform distribution to sign the terms in the RHS of equation (27). Using equations (14) and (15,) we have $\sigma_1 = \overline{\sigma}\frac{s_2 - s_1}{4s_2 - s_1}$ and $\sigma_2 = \overline{\sigma}\frac{2s_2 - s_1}{4s_2 - s_1}$, which implies that revenue functions can be expressed as a function of (s_1, s_2): $\pi_1(s_1, s_2) = \frac{\overline{\sigma}}{2r}\frac{s_2 - s_1}{(4s_2 - s_1)^2}s_1 s_2$ and $\pi_2(s_1, s_2) = \frac{\overline{\sigma}}{2r}\frac{s_2 - s_1}{(4s_2 - s_1)^2}(2s_2)^2$. After some algebra, one can show that $\frac{\partial^2 \pi_2}{\partial s_2^2} = -\frac{4\overline{\sigma}}{r}\frac{s_1^2(s_1 + 5s_2)}{(4s_2 - s_1)^4} < 0$ and $\frac{\partial^2 \pi_2}{\partial s_2 \partial s_1} = \frac{4\overline{\sigma}}{r}\frac{s_1 s_2(s_1 + 5s_2)}{(4s_2 - s_1)^4} > 0$. These inequalities, together with convexity of C, yield $S_2'(s_1) > 0$. *Q.E.D.*

B.7. *Proof of Proposition 7*

Consider the maximum speed bound first. Holding market shares constant, the regulator seeks to maximize social welfare in each venue $i = 1, 2$. For venue 1, the optimality condition is $\frac{1}{2r}\int_{\sigma_1}^{\sigma_2} \sigma\, dG(\sigma) = C'(s_1)$. Using $d \equiv \frac{s_2}{s_1}$, we have $\sigma_1 = \overline{\sigma}\frac{d-1}{4d-1}$ and $\sigma = \overline{\sigma}\frac{2d-1}{4d-1}$ and thus $\frac{1}{2r}\int_{\sigma_1}^{\sigma_2} \sigma\, dG(\sigma) = \frac{\overline{\sigma}}{4r}\frac{d(3d-2)}{(4d-1)^2}$. From equation (25), we know that in an interior solution venue 1's marginal speed cost must equal its marginal revenue, which is given by $\frac{\overline{\sigma}}{2r}\frac{d^2(4d-7)}{(4d-1)^3}$. Straightforward calculations then show that $\frac{1}{2r}\int_{\sigma_1}^{\sigma_2} \sigma\, dG(\sigma) > C'(s_1)$, implying under provision of speed at the market equilibrium for venue 1. Similarly, the regulator optimality condition for venue 2 is $\frac{1}{2r}\int_{\sigma_2}^{\overline{\sigma}} \sigma\, dG(\sigma) = C'(s_2)$. Simple calculations imply that $\frac{1}{2r}\int_{\sigma_2}^{\overline{\sigma}} \sigma\, dG(\sigma) = \frac{\overline{\sigma}}{r}\frac{d(3d-1)}{(4d-1)^2}$. From equation (26), we know that the marginal cost $C'(s_2)$ must be equal in a market solution to $\frac{2\overline{\sigma}}{r}\frac{d^2(2d+1)}{(4d-1)^3}$. Straightforward calculations then show that $\frac{1}{2r}\int_{\sigma_2}^{\overline{\sigma}} \sigma\, dG(\sigma) > C'(s_2)$, implying under provision of speed at the market equilibrium for venue 2. We conclude that limiting speeds is not welfare enhancing.

Consider now the minimum speed bound. We start by computing the total derivative of the welfare function (9) with respect to s_1 at the market equilibrium:

$$\begin{aligned}\frac{\mathrm{d}\mathcal{W}}{\mathrm{d}s_1} = &\left(\frac{1}{2r}\int_{\sigma_1}^{\sigma_2} \sigma\, dG(\sigma) - C'(s_1)\right) + \left(\frac{1}{2r}\int_{\sigma_2}^{\overline{\sigma}} \sigma\, dG(\sigma) - C'(s_2)\right)S_2'(s_1)\\ &- \frac{1}{2r}\frac{\mathrm{d}d}{\mathrm{d}s_1}\left(\sigma_1 s_1 \frac{\partial \sigma_1}{\partial d} + \sigma_2(s_2 - s_1)\frac{\partial \sigma_2}{\partial d}\right).\end{aligned}$$

We have shown above that the bracketed expressions in the first two terms of the right-hand side of equation (28) are positive and also that $S_2'(s_1) > 0$. Thus, the sum of the first two terms is positive. Consider now the third term. It is immediate from Proposition 4 that $\frac{\partial \sigma_1}{\partial d}$ and $\frac{\partial \sigma_2}{\partial d}$ are positive under Assumption 1. Also notice that $\frac{\mathrm{d}d}{\mathrm{d}s_1} = \frac{S_2'(s_1) - d}{s_1}$. The revenue function for each venue is homogeneous of degree 1, implying that the marginal revenue functions are homogeneous of degree 0. By Euler's theorem, then $d = -\frac{\partial^2 \pi_2}{\partial s_2 \partial s_1}\Big/\frac{\partial^2 \pi_2}{\partial s_2^2}$ and $S_2'(s_1) < d$ given equation (27), which yields $\frac{\mathrm{d}d}{\mathrm{d}s_1} < 0$. We conclude that at the duopoly's speed choice equilibrium, we have $\frac{\mathrm{d}\mathcal{W}}{\mathrm{d}s_1} > 0$. *Q.E.D.*

B.8. *Proof of Proposition 8*

We analyze below the existence of Nash equilibrium (NE) in pure strategies of the normal-form entry game. The relation between entry costs κ and profits determines the number of active venues in equilibrium. Let $\overline{\pi}_i \equiv \max\{\pi_i^{\text{int}}, \pi_i^{\text{seg}}\}$ and $\underline{\pi}_i \equiv \min\{\pi_i^{\text{int}}, \pi_i^{\text{seg}}\}$ with $i \in \{1, 2\}$.

- *Two-venue equilibria*. By Proposition 4, we have that $\underline{\pi}_1 = \pi_1^{\text{seg}}$. It is immediate then that entry is always optimal for both venues when $\kappa \leq \pi_1^{\text{seg}}$ and that, for any $\pi_1^{\text{seg}} < \kappa \leq \pi_1^{\text{int}}$, we have $\pi_1^{\text{seg}} - \kappa < 0$ and $\pi_1^{\text{int}} - \kappa \geq 0$. A duopoly is never sustainable whenever $\kappa > \pi_1^{\text{int}}$.
- *Single-venue equilibria*. Suppose $\pi_1^{\text{int}} < \kappa \leq \pi_m(s_2)$.

– Case 1: $\pi_m(s_2) \geq \kappa > \pi_m(s_1)$. The only NE has the slow venue out and the fast venue entering.

– Case 2: $\overline{\pi}_1 \leq \kappa \leq \pi_m(s_1)$. In this case, there are two NE where only one venue enters, either the slow venue or the fast venue.

- *No-entry equilibrium*. Whenever $\kappa > \pi_m(s_2)$, the only NE has both venues out.

The analysis above shows that the number of entrants is weakly higher under price integration, proving part (i).

We now turn to part (ii). Entry is profitable for the slow venue if and only if $\pi_1 > \kappa$, that is, $\frac{s_1\sigma_1}{2r}(G(\sigma_2) - G(\sigma_1)) > \kappa + C(s_1)$. Excess entry happens only when $\pi_1 > \kappa$ and the value of expression (19) is negative. A *necessary* condition for excess entry is therefore

$$\frac{s_1}{2r}\int_{\sigma_1}^{\sigma_2} \sigma\, dG(\sigma) + \frac{s_2}{2r}\int_{\sigma_2}^{\overline{\sigma}} \sigma\, dG(\sigma) - C(s_2) - \frac{s_m}{2r}\int_{\sigma_m}^{\overline{\sigma}} \sigma\, dG(\sigma) + C(s_m) < \frac{s_1\sigma_1}{2r}\big(G(\sigma_2) - G(\sigma_1)\big),$$

which we can rewrite as

$$\frac{s_1}{2r}\int_{\sigma_1}^{\sigma_2} (\sigma - \sigma_1)\, dG(\sigma) + \left(\frac{s_2}{2r}\int_{\sigma_2}^{\overline{\sigma}} \sigma\, dG(\sigma) - \frac{s_m}{2r}\int_{\sigma_m}^{\overline{\sigma}} \sigma\, dG(\sigma)\right) < C(s_2) - C(s_m).$$

The first term of the left-hand side in the expression above is strictly positive. We know from Proposition 4 that $\sigma_2 < \sigma_m$ and from Proposition 6 that $s_2 \geq s_m$. Therefore, the bracketed second term is also strictly positive. Finally, the term on the right-hand side is near zero when speed costs approach zero or when $s_2 \approx s_m$. *Q.E.D.*

REFERENCES

AÏT-SAHALIA, Y., AND M. SAGLAM (2013): "High Frequency Traders: Taking Advantage of Speed," NBER Working Paper 19531. [1070]

ANGEL, J. J., L. E. HARRIS, AND C. S. SPATT (2015): "Equity Trading in the 21st Century: An Update," *Quarterly Journal of Finance*, 5, 1550002. [1072]

BIAIS, B., T. FOUCAULT, AND S. MOINAS (2015): "Equilibrium Fast Trading," *Journal of Financial Economics*, 116, 292–313. [1070]

BIAIS, B., J. HOMBERT, AND P.-O. WEILL (2014): "Equilibrium Pricing and Trading Volume Under Preference Uncertainty," *The Review of Economic Studies*, 81, 1401–1437. [1103]

BOEHMER, E. (2005): "Dimensions of Execution Quality: Recent Evidence for US Equity Markets," *Journal of Financial Economics*, 78, 553–582. [1072]

BUDISH, E., P. CRAMTON, AND J. SHIM (2015): "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *Quarterly Journal of Economics*, 130, 1547–1621. [1070]

CHAMPSAUR, P., AND J.-C. ROCHET (1989): "Multiproduct Duopolists," *Econometrica*, 57, 533–557. [1088]

CHAO, Y., C. YAO, AND M. YE (2017): "Why Discrete Price Fragments U.S. Stock Exchanges and Disperses Their Fee Structures," Working Paper, University of Illinois at Urbana-Champaign. [1070]

DEGRYSE, H., F. DE JONG, AND V. VAN KERVEL (2015): "The Impact of Dark Trading and Visible Fragmentation on Market Quality," *Review of Finance*, 19, 1587–1622. [1069]
DONNENFELD, S., AND S. WEBER (1995): "Limit Qualities and Entry Deterrence," *The RAND Journal of Economics*, 26, 113–130. [1103]
DUFFIE, D., N. GARLEANU, AND L. H. PEDERSEN (2005): "Over-the-Counter Markets," *Econometrica*, 73, 1815–1847. [1070,1104]
——— (2007): "Valuation in Over-the-Counter Markets," *Review of Financial Studies*, 20, 1865–1900. [1068]
FOUCAULT, T., AND A. J. MENKVELD (2008): "Competition for Order Flow and Smart Order Routing Systems," *Journal of Finance*, LXIII, 119–158. [1069,1074]
FOUCAULT, T., AND C. A. PARLOUR (2004): "Competition for Listings," *The Rand Journal of Economics*, 35, 329–355. [1070]
FOUCAULT, T., J. HOMBERT, AND I. ROSU (2016): "News Trading and Speed," *Journal of Finance*, LXXI, 335–382. [1070]
GABSZEWICZ, J., AND J.-F. THISSE (1979): "Price Competition, Quality and Income Disparities," *Journal of Economic Theory*, 20, 340–359. [1069]
——— (1980): "Entry (and Exit) in a Differentiated Industry," *Journal of Economic Theory*, 22, 327–338. [1070]
GARBADE, K. D., AND W. L. SILBER (1977): "Technology, Communication and the Performance of Financial Markets: 1840–1975," *Journal of Finance*, XXXIII, 819–832. [1071]
KIRILENKO, A., A. S. KYLE, M. SAMADI, AND T. TUZUN (2017): "The Flash Crash: The Impact of High Frequency Trading on an Electronic Market," *Journal of Finance*, LXXII, 967–998. [1094]
LAGOS, R., AND G. ROCHETEAU (2009): "Liquidity in Asset Markets With Search Frictions," *Econometrica*, 77, 403–426. [1070]
LAGOS, R., G. ROCHETEAU, AND R. WRIGHT (2017): "Liquidity: A New Monetarist Perspective," *Journal of Economic Literature*, 55, 371–440. [1070]
LEWIS, M. (2014): *Flash Boys: A Wall Street Revolt*. New York: Allen Lane. [1071]
MANKIW, N. G., AND M. D. WHINSTON (1986): "Free Entry and Social Inefficiency," *The RAND Journal of Economics*, 17, 48–58. [1068,1089]
MENDELSON, H. (1987): "Consolidation, Fragmentation, and Market Performance," *Journal of Financial and Quantitative Analysis*, 22, 187–207. [1069]
O'HARA, M., AND M. YE (2011): "Is Market Fragmentation Harming Market Quality?" *Journal of Financial Economics*, 100, 459–474. [1069]
PAGANO, M. (1989): "Trading Volume and Asset Liquidity," *Quarterly Journal of Economics*, 104, 255–274. [1069]
PAGNOTTA, E. S. (2014): "Speed, Fragmentation, and Asset Prices," Working Paper, Imperial College London. [1070]
PAGNOTTA, E. S., T. PHILIPPON (2018): "Supplement to 'Competing on Speed'," *Econometrica Supplemental Material*, 86, https://doi.org/10.3982/ECTA10762. [1083]
RONNEN, U. (1991): "Minimum Quality Standards, Fixed Costs, and Competition," *The RAND Journal of Economics*, 22, 490–504. [1088]
SANTOS, T., AND J. A. SCHEINKMAN (2001): "Competition Among Exchanges," *Quarterly Journal of Economics*, 116, 225–1061. [1069]
SECURITIES AND COMMISSION (2010): "Concept Release on Equity Market Structure," Release No. 34-61358. [1073]
SHAKED, A., AND J. SUTTON (1982): "Relaxing Price Competition Through Product Differentiation," *Review of Economic Studies*, 44, 3–13. [1069,1088]
——— (1983): "Natural Oligopolies," *Econometrica*, 51, 1469–1483. [1069]
SPENCE, M. (1976): "Product Differentiation and Welfare," *American Economic Review*, 66, 407–414. [1089]
VAYANOS, D., AND T. WANG (2007): "Search and Endogenous Concentration of Liquidity in Asset Markets," *Journal of Economic Theory*, 136, 66–104. [1070]
WEILL, P.-O. (2007): "Leaning Against the Wind," *Review of Economic Studies*, 74, 1329–1354. [1070]
——— (2008): "Liquidity Premia in Dynamic Bargaining Markets," *Journal of Economic Theory*, 140, 66–96. [1070]

Co-editor Lars Peter Hansen handled this manuscript.

Manuscript received 30 April, 2012; final version accepted 27 December, 2017; available online 11 January, 2018.