# Competing on Speed

Emiliano Pagnotta and Thomas Philippon[*]

July 2012

**Abstract**

Speed and fragmentation have reshaped global securities markets: Large-cap U.S. stocks can now be traded in almost 50 venues, and execution times are measured in milliseconds. We analyze these evolutions in a model where exchanges invest in trading speed and compete for investors who choose where and how much to trade. Faster venues charge higher fees and attract speed-sensitive investors. Competition among exchanges increases investor participation, traded volumes, and allocative efficiency but can lead to socially excessive levels of speed. Regulations that protect investors (e.g. SEC's trade-through) lead to more fragmentation and faster speeds, but may reduce welfare. Independently of technology and entry costs, the optimal design has a single operating exchange. Our model sheds light on the experience of European and U.S. markets since the implementation of MiFID and Reg NMS.

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

[*] Both authors are at the New York University Stern School of Business. Philippon is also at the National Bureau of Economic Research and the Centre for Economic Policy Research. We are particularly grateful to Joel Hasbrouck for his insights, and to Pierre-Oliver Weill, Guillaume Rocheteau, Albert Menkveld and Jonathan Brogaard (discussants). We thank Yakov Amihud, Darrell Duffie, Thierry Foucault, Ricardo Lagos, Lasse Pedersen, Marti Subrahmanyam, Dimitri Vayanos, Mao Ye, and seminar participants at HEC Paris, the NYU Stern School of Business, London School of Economics, Toulouse School of Economics, Rochester University, the Tinbergen Institute, the University of Illinois at Chicago, University of Lugano, Universidad de San Andres, and seminar participant a the following conferences: the 2011 Society of Economic Dynamics, 4th Annual Conference on Money, Banking and Asset Markets at UWM, Tepper-LAEF Advances in Macro Finance III, Western Finance Association (Las Vegas), Finance Theory Group Meeting (HBS), Cowles Foundation GE Conference (Yale), Econometric Society NA Meetings (Evanston), the 2012 Financial Managment Association Napa Valley, and Fourth Hedge Fund Euronext conference. We acknowledge the support of the Smith Richardson Foundation.

> "In this high-tech stock market, Direct Edge and the other exchanges are sprinting for advantage. All the exchanges have pushed down their latencies [...] Almost each week, it seems, one exchange or another claims a new record [...] The exchanges have gone warp speed because traders have demanded it. Even mainstream banks and old-fashioned mutual funds have embraced the change." *The New York Times*, January 1st 2011, The New Speed of Money, Reshaping Markets

# 1 Introduction

The securities exchange industry has been deeply transformed over the past decade. In particular, the speed at which investors trade has greatly increased and stock trading, particularly in the U.S. and Europe, has become significantly more fragmented. The consequences of these transformations are the subject of heated debates in academic and policy circles. This paper provides a joint analysis of trading speed, trading regulations, and market fragmentation.

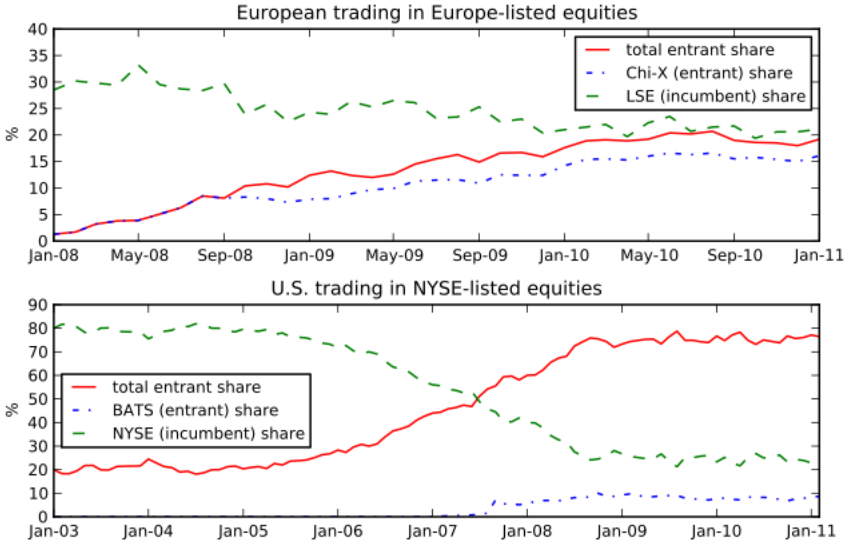
Let us first consider trading speed. Major market centers have made costly investments in fast computerized trading platforms to reduce order execution and communication latencies. This process has gone beyond stock exchanges to include futures, options, bonds, and currencies. Although these investments were chiefly first observed in the U.S., they have been recently undertaken in virtually any world economy, and at an accelerating rate during the second half of the 2000s.[1] The driving forces underlying this speed race are likely to be different than those of other historical periods. In the human-driven trading era, for example, higher execution speeds helped reduce moral hazard with floor brokers, but this aspect has become less relevant today.

The second major feature of the new trading landscape is fragmentation, illus-

[1] Table 5 in the Appendix collects several investments in latency reductions observed in recent years. Angel et al. (2011) displays the reduction in the execution times of small orders on the NYSE and NASDAQ over the last decade. Although we concentrate on trading venues, the speed investment frenzy has additional dimensions. One prominent example is the provision of connections between financial centers. Take Chicago–New York as an example. Spread Networks recently invested approximately $300 million in a new fiber optic cable that links these cities through the straightest possible route, saving about 100 miles with respect to existing ones. This allows the company to shave 6 milliseconds off their delay, for a total delay of 15 milliseconds. The success of this business model motivated McKay Brothers, a leading provider of low-latency wireless transport equipment, to contract the creation of a $300 million microwave-based network of towers connecting the same cities to Aviat Networks. This new technology is expected to reduce latencies even further.

trated by Figure 1 for Europe and the U.S. The top panel (Europe) shows that traditional markets such as the London Stock Exchange have lost market share to faster entrants such as Chi-X. The bottom panel (U.S.) shows an even more dramatic evolution: The fraction of NYSE-listed stocks traded at the NYSE has 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). Overall, fragmentation has increased so dramatically that market participants now keep track of fragmentation indexes across asset classes and countries.[2]

Figure 1: Market Fragmentation



source: Barclays Capital Equity Research and Federation of European Securities Exchanges

Source: Menkveld (2011)

Market regulators have not been passive witnesses of this process. In the U.S., policy makers have encouraged fragmentation to reduce the market power of exchanges and other intermediaries, prominently with Regulation National Market System (Reg NMS). For example, the Securities and Exchange Commission (2010) (SEC) 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

[2] See, for example, the Fidessa fragmentation indexes.

> for trading centers to create new products, provide high quality trading services that meet the needs of investors, and keep trading fees low."

The effects are tangible: Large-cap stocks that previously traded in one or two exchanges can now be traded in almost 50 venues (including internalization pools and over-the-counter, or OTC, venues). Encouraged by the recent U.S. experience, many other countries have started promoting competition between market centers. Concerns about adverse effects of trading fragmentation, in turn, motivated regulators to design rules that promote "investor protection." In the U.S., the idea of investor protection is implemented by the trade-through rule provided by Rule 611 of Reg NMS, which essentially requires that any venue executes its trades at the national best bid and offer, thereby consolidating prices from a scattered trading map.

But why do exchanges compete on speed? Is there a relation between the increase in trading speeds and the level of market fragmentation? What are the consequences of these changes? Does fragmentation achieve policy makers' goals? Should investor protection be fostered in the first place? We argue that technological advances, competition in the securities exchange industry and market regulations interact with each other affecting the trading landscape, asset prices, investor participation, and, ultimately, social welfare. Figure 2 captures this idea graphically. To shed light on these issues, we build on the insight that, everything else being constant, all investors are (at least weakly) better-off by trading faster, but they do not value speed equally. Thus exchanges competing to attract investors can vertically differentiate their intermediation services by catering to different clienteles, relaxing price competition.

Analyzing these issues is difficult because it requires modeling four separate components: (i) why and how investors value trading 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, which is depicted in Figure 3.

Our first task is to provide explicit microfoundations for how investors value speed in financial markets. We consider a dynamic infinite-horizon model where heterogeneous investors buy and a sell a single security. Ex post gains from trade arise from shocks to the marginal utility (or marginal cost) of holding the asset.[3] High–marginal-

[3] As is well understood in the literature, these shocks can capture liquidity demand (i.e., a need for cash), financing costs, hedging demand, or any other personal use of the asset, including specific

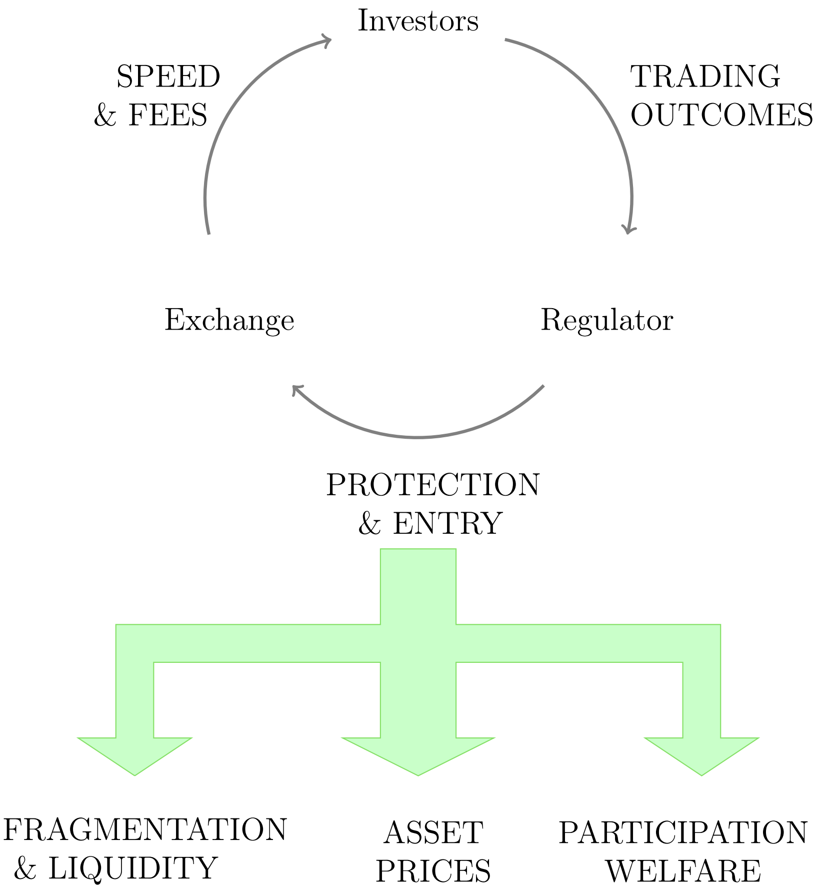


Figure 2: Security Exchange Industry Evolution and Aggregate Outcomes

utility investors are natural buyers, while low–marginal-utility investors are natural sellers of the asset. In this model, speed allows investors to realize a larger fraction of the ex post gains from trade (see Proposition 1).

Our second task is to analyze the allocation of investors across trading venues. To do so we model ex ante heterogeneity among venues and investors. Venues differ in their trading speeds and compete in prices. Investors differ by the volatility of their private value processes. We show that investors with high expected volatility attach a higher value to speed. We characterize the equilibrium with one venue (monopoly), and with two venues with different speeds (differentiated duopoly). Competition leads to lower fees and greater investor participation. Faster venues charge a higher price and attract speed-sensitive investors. The first contribution of our paper is to characterize the pricing decisions and equilibrium profits of trading venues and the participation and affiliation choices of investors (see Proposition 2).

Our third task is to analyze the impact of trading regulations aimed at protecting investors. We propose a stylized analysis of this regulation by considering two polar cases. In one case, which we refer to as "free segmentation," any venue can refuse to execute the trades of investors from the other venue. The venues are effectively segmented and trades occur at different prices. The other case corresponds to "price protection." We find that price protection acts as a subsidy for the relatively slow market. At the trading stage, investors in the slow venue enjoy being able to trade with investors from the fast venue. Anticipating this, they are more willing to join the slow venue under price protection than under free segmentation. An important contribution of our paper is to analyze how price protection affects ex ante competition among exchanges (see Propositions 2 and 3) and ex post aggregate outcomes. To the best of our knowledge, ours is the first formal analysis of this issue.

When we endogenize the speed and the market structure, we find that price protection encourages entry. In addition, we show that fragmentation leads to more investment in trading technologies and thus faster trading speeds. Putting these various pieces together, our model provides a consistent interpretation of the U.S. experience in recent years: After the implementation of Reg NMS, new market centers proliferated and trading speed increased rapidly (see Propositions 3, 4 and 5).

arbitrage opportunities (see Duffie et al. (2007) for a discussion). The important point is that these shocks affect the private value of the asset, not its common value. They therefore generate the gains from trade that are a required building block of any trading model.

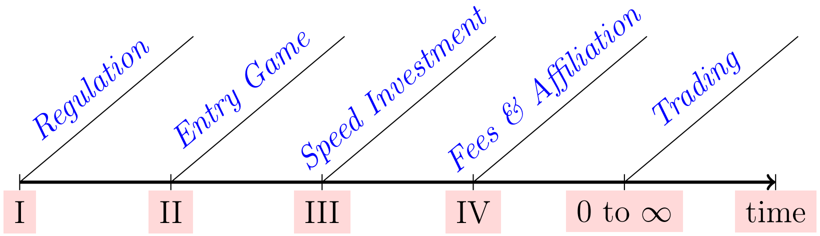


Figure 3: Timing and Structure of the Model

The model thus provides a rational two-way link between these two phenomena. In the absence of speed differentiation, exchanges Bertrand-compete and the market for liquidity becomes a natural monopoly for any negligible entry cost. Higher fragmentation, in turn, stimulates further speed production.

Finally, we analyze the welfare implications of entry, speed, and investor protection. As a benchmark we characterize the efficient outcome under the constraint that venues break even. Interestingly, we find that independently technology and entry costs, and in the absence of thick market externalities, a planner would choose to operate only one venue. Our model then allows us to answer three important questions: When does competition increase welfare? When does investment in trading speed increase welfare? Is price protection socially optimal?

We find that the market outcome is generally inefficient, but the inefficiencies play out differently depending on the market structure. In the monopoly case, participation is always too low and depends exclusively on the distribution of investors. Allowing for endogenous speed improves welfare, even though the speed chosen by the monopolist may be higher or lower than the one chosen by the planner. In the duopoly case, both entry and speed can be inefficient. Regarding entry, there is the usual tension between business stealing on the one hand, and competition and product diversity on the other. Entry typically improves welfare, but it can be excessive if entry costs are relatively high. Regarding speed choices, we find a fairly clear and intuitive condition: Allowing venues to compete on speed improves welfare if the default available speed is relatively low (e.g., purely human-based trading) but decreases welfare once the default speed reaches a certain threshold.

The welfare consequences of price protection depend crucially on entry. When protection increases entry, it has a first-order impact on welfare. There is a range of economies where entry costs are intermediate and the market equilibrium only sustains one exchange entering profitably. In such cases the implicit subsidy embedded in price protection may stimulate competition and result in higher investor participation and allocative efficiency. The effects are not necessarily welfare improving when combined entry cost offset social gains. When price protection does not increase entry, it has a small negative impact on welfare because it dampens competition. In all these cases, the endogenous affiliation decisions of investors play a crucial role.

## 1.1 Discussion of the literature

Our paper relates to several strands of the literature in economics and finance. To the best of our knowledge, the first work on the issue of separated markets and the role of speed is by Garbade and Silber (1977). Early theoretical analyses of fragmentation include those of Mendelson (1987), Pagano (1989), and Madhavan (1995). The focus of these early papers is the tradeoff between liquidity externalities, on the one hand, and market power and trading costs, on the other.[4] This tension was of first-order of important during the time in which different market places were not as integrated as they are nowadays. Competition between markets with different trading rules has been studied in Biais (1993), Glosten (1994), Hendershott and Mendelson (2000), Parlour and Seppi (2003), and Rust and Hall (2003). Models of competing exchanges include those of Santos and Scheinkman (2001), which study competition in margin requirements, and Foucault and Parlour (2004), which study competition in listing fees (but not competition for order flow). More recently, Colliard and Foucault (2012) study the effect of trading fees in a context where an exchange competes with an OTC dealer. Similar to our analysis, the authors endogenize both investors' trading decisions and exchanges' pricing decisions. In our paper we show how differentiated market center competition affects not only liquidity and trading costs but also technological acquisition, asset price levels, and investor participation levels. We also provide, to the best of our knowledge, the first formal analysis of price protection on equilibrium market structure and welfare.

The recent sharp increase in market fragmentation in developed countries has encouraged a new wave of empirical studies. O'Hara and Ye (2011) provide several references. Interestingly, these authors find that an increase in trading fragmentation is associated with lower costs and faster execution speeds for a given asset. Amihud

[4] For a thorough textbook analysis see Chapter 26 in Harris (2003).

et al. (2003) study the effects of trading consolidation of two virtually identical assets on price levels and find that consolidation increases asset prices.

Amihud and Mendelson (1986) pioneered the analysis of the effect of liquidity on asset prices. The literature of trading with search frictions has been fostered by Duffie et al. (2005). Our trading model is closest to that of Lagos and Rocheteau (2009). Weill (2007) uses a related framework to analyze exchanges. Vayanos and Wang (2007) study concentration of liquidity across assets. Our model contributes to this literature by simultaneously studying interactions between strategic trading venues and investors' affiliation decisions. Thus, we are able to explicitly study how (i) the entry of trading venues, (ii) speed investments, (iii) participation levels, and (iv) investor protection affect asset prices.[5] The full analysis of pricing implications is provided in a companion paper (see Pagnotta (2013)).

After the recent financial crisis there is an increasing interest in understanding the social value of financial markets. Our paper contributes to this discussion by emphasizing the welfare consequences of changes in market organization and regulation. For example, in the corporate finance literature it is often stated that higher asset prices are socially desirable since they reduce financing cost for firms in primary markets. By analyzing explicitly the welfare impact of trading in secondary markets, we highlight that observed prices are a poor statistic for welfare. For instance, prices can be inefficiently high due to limited investor participation. A complete analysis of the social value of financial markets needs then to incorporate possible tensions between primary and secondary markets effects.

Gabszewicz and Thisse (1979), Shaked and Sutton (1982), and Shaked and Sutton (1983) pioneered the analysis of vertically differentiated oligopolies. Our framework is enriched by having agents consuming a differentiated product first ("liquidity") and a homogeneous product (the asset itself) second. Consequently, we are able to endogenize the value of "quality" (trading delays here) through a microfounded trading game.

The rest of the paper is organized as follows. Section 2 presents our benchmark trading model derives the value functions of investors. Section 3 analyzes competition among trading venues with and without price protection. Section 4 analyzes trading

[5] Huang and Wang (2010) also study investor participation and welfare in a context where agents bear idiosyncratic risks, but they focus on liquidity externalities and the decision to become a liquidity provider. Trading costs are exogenous in that setting, while they are determined strategically in our paper.

venues' entry decisions and investment in speed. Section 5 characterizes the solutions to the planner's problem and studies the efficiency of the market equilibrium. Section 6 contains a numerical analysis of the model, and Section 7 concludes the paper.

# 2 Trading Model

We present our trading model in the case of one market. This section provides explicit microfoundations for how investors value speed in financial markets. The key result of this section is a characterization of value functions as a function of speed and investor characteristics.

## 2.1 Preferences and Technology

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 exchanges. The trading technology must capture the role of speed in financial markets.

Time is continuous and we fix a probability space. The model has a continuum of heterogeneous investors, two goods, and one asset. The measure of investors is normalized to one and their preferences are quasilinear. The numéraire good (cash) has a constant marginal utility normalized to one and can be freely invested at the constant rate of return $r$. The asset is in fixed supply, $\bar{a}$, which is also the (expected) endowment of each investor. We restrict asset holdings to $a_t \in \{0, 1\}$. One unit of asset pays a constant dividend equal to $\mu$ of a perishable non-tradable good. The flow utility that an investor derives from holding $a_t$ units of the asset at time $t$ is

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

where $(\sigma, \epsilon_t)$ denotes the type of investor. This type is defined by a fixed component $\sigma$ and a time-varying (random) component $\epsilon_t$. The fixed component $\sigma \in [0, \bar{\sigma})$ is known at time zero and distributed according to the twice-differentiable cumulative distribution $G$, with a log-concave density function $g$ that is positive everywhere. The type $\epsilon_t \in \{-1, +1\}$ changes randomly over time. The times when a change can occur are distributed exponentially with parameter $\gamma$. Conditional on a change, $\epsilon$ is i.i.d. and each value has equal probability.

As explained in the introduction, the $\epsilon$-shocks can capture time-varying liquidity demands, financing costs, hedging demands, or specific investment opportunities. For instance, a corporate investor may need to sell its financial assets to finance a real investment. A household may do the same for the purchase of a durable good or a house. The parameter $\sigma$ then simply measures the size of these shocks. In the context of delegated management, the shock represents the sum of the shocks affecting all the investors in a given fund or brokerage house. The parameter $\gamma$ measures the mean reversion of the utility flow process and is assumed for simplicity to be the same for all investors.[6]

Our paper focuses on the trading technology for the asset. For clarity, we describe here the case where all investors trade at the same speed (later we endogenize speed choices and consider markets with different speeds). The market where investors trade the asset is characterized by the constant contact rate $\rho$. Conditional on being in contact, the market is Walrasian and clears at the price $p_t$.[7] Any investor in contact with the market at time $t$ can trade at the price $p_t$. Investors that are not in contact simply keep their holdings constant.

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

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

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

[6] We introduce heterogeneity in $\sigma$, not in $\gamma$, because the key point in our analysis is the link between gains from trade and speed. It is important to understand that a higher value of $\gamma$ implies *lower* gains from trade. An investor with a high value of $\gamma$ is not eager to trade since she can simply wait for her type to mean revert. In particular, a high value of $\gamma$ would not capture the idea of fleeting trading opportunities. This idea is better captured by a high value of $\sigma$.

[7] It would be straightforward to add bargaining with market makers and bid-ask spreads, but this would not bring new insights compared to Duffie et al. (2005) and Lagos and Rocheteau (2009). For simplicity we therefore assume competitive trading conditional on being in contact with the market. A similar market mechanism is considered in the monetary economy of Rocheteau and Wright (2005) (which they label competitive equilibrium).

## 2.2 Trading Equilibrium

We show that the asset price remains constant during the trading game. The value functions are thus time independent and equation (1) becomes simply

$$rV_{\sigma\epsilon}(a) = u_{\sigma,\epsilon}(a) + \gamma \sum_{\epsilon'} \phi_{\epsilon'} \left[V_{\sigma\epsilon'}(a) - V_{\sigma\epsilon}(a)\right] + \rho \left[V_{\sigma\epsilon}(a^*_{\sigma,\epsilon}) - V_{\sigma\epsilon}(a) - p(a^*_{\sigma,\epsilon} - a)\right]. \tag{2}$$

Following Lagos and Rocheteau (2009), we define the adjusted holding utility as

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

Lagos and Rocheteau (2009) (see Lemma 1 there) show that $\bar{u}$ is the object that investors seek to maximize when deciding how much to trade. Note that since $\epsilon$ is i.i.d. with mean zero, we have $\mathbb{E}\left[u_{\sigma,\epsilon'}(a) \mid \epsilon\right] = \mu a$ for any $a$ and any $\epsilon$. This expected utility over $\epsilon'$ does not depend on $\sigma$ or $\epsilon$. This implies that

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

Recall that $G$ was the ex ante distribution of permanent types. Let $\tilde{G}(\sigma)$ be the measure of traders of type lower than $\sigma$ in the market. If all potential investors join the market, we simply have $\tilde{G} = G$. In the generic case, however, we have $\tilde{G} \leq G$ since some investors do not participate. Indeed, we shall see that in the multiple-venue model the distribution $\tilde{G}$ is typically discontinuous. We therefore present our results without putting any restrictions on the function $\tilde{G}$.

**Lemma 1.** *An equilibrium with constant price $p$ is characterized by the demand functions*

$$a^*(p;\sigma,\epsilon) = \arg\max_a \bar{u}(a;\sigma,\epsilon) - rpa. \tag{3}$$

*and the market-clearing condition*

$$\int_\sigma \sum_\epsilon \phi_\epsilon a^*(p;\sigma,\epsilon)\, d\tilde{G}(\sigma) = \bar{a}\tilde{G}(\bar{\sigma}), \tag{4}$$

*Proof.* See Proposition 1 in Lagos and Rocheteau (2009). The proposition only needs to be adapted to take into account heterogeneity in $\sigma$. *Q.E.D.*

There is a clear symmetry around $\bar{a} = 1/2$ since half the investors are of trading type $\epsilon = +1$ and half are of trading type $\epsilon = -1$. It is therefore sufficient to analyze a market where $\bar{a} \leq 1/2$. In this case, supply is short and low types always sell their entire holdings when they contact the market. Moreover, there is a marginal type $\hat{\sigma}$ that is indifferent between buying and not buying when $\epsilon = 1$. This marginal type is defined by

$$\hat{\sigma}\left(p, \rho\right) \equiv \left(1 + \frac{\gamma}{r + \rho}\right)\left(rp - \mu\right). \tag{5}$$

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

We can use these demand curves to rewrite the market-clearing condition. All negative trading types $\epsilon = -1$ want to hold $a = 0$, and they represent half of the traders. The trading types $\epsilon = +1$ want to hold one unit if $\sigma > \hat{\sigma}$ and nothing if $\sigma < \hat{\sigma}$. The demand for the asset is $1/2 \times (\tilde{G}\left(\bar{\sigma}\right) - \tilde{G}\left(\hat{\sigma}\right))$. The ex ante supply of the asset (per capita) is $\bar{a}$. The market clearing-condition is therefore

$$\frac{\tilde{G}\left(\bar{\sigma}\right) - \tilde{G}\left(\hat{\sigma}\right)}{2} = \bar{a}\tilde{G}\left(\bar{\sigma}\right). \tag{6}$$

Note that the asset holdings of types $\sigma < \hat{\sigma}$ are non-stationary since they never purchase the asset. A type $\sigma < \hat{\sigma}$ sells its holding $\bar{a}$ on first contact with the market and never holds the asset again. Over time the assets move from the low-$\sigma$ types to the high-$\sigma$ types and then keep circulating among high types in response to $\epsilon$ 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 $\rho\bar{a}\tilde{G}\left(\bar{\sigma}\right)$. The gross demand from high types is always $\rho\left(\tilde{G}\left(\bar{\sigma}\right) - \tilde{G}\left(\hat{\sigma}\right)\right)/2$. From equation (6), the market always clears.

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

$$\frac{\gamma}{2}\alpha_{\sigma,+}(1) = \frac{\gamma}{2}\alpha_{\sigma,-}(1) + \rho\alpha_{\sigma,+}(0). \tag{7}$$

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

$$\frac{\gamma}{2}\alpha_{\sigma,-}(0) = \rho\alpha_{\sigma,-}(1) + \frac{\gamma}{2}\alpha_{\sigma,+}(0). \tag{8}$$

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

$$\left(\frac{\gamma}{2} + \rho\right)\alpha_{\sigma,+}(0) = \frac{\gamma}{2}\alpha_{\sigma,-}(0) \tag{9}$$

$$\left(\frac{\gamma}{2} + \rho\right)\alpha_{\sigma,-}(1) = \frac{\gamma}{2}\alpha_{\sigma,+}(1) \tag{10}$$

Finally, the shares must add up to one, therefore

$$\sum_{\epsilon=\pm,a=0,1} \alpha_{\sigma,\epsilon}(a) = 1 \tag{11}$$

We summarize our results in the following lemma

**Lemma 2.** *The trading equilibrium is characterized by the price $p$ and marginal type $\hat{\sigma}$ defined in equations (5) and (6). The transition dynamics are as follows. The price remains constant while asset holdings shift from low $\sigma$-types to high $\sigma$-types. Low types $(\sigma < \hat{\sigma})$ sell their initial holdings $\bar{a}$ and never purchase the asset again. High types $\sigma \geq \hat{\sigma}$ buy when $\epsilon = 1$ and sell when $\epsilon = -1$. The distribution of holdings among high $\sigma$-types 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 $\alpha_{\sigma,+}(0) = \alpha_{\sigma,-}(1) = \frac{1}{4}\frac{\gamma}{\gamma+\rho}$. These allocation converge to the Walrasian allocation when $\rho \to \infty$.*

*Proof.* To see the steady state allocations, add (7) and (10) to get $\alpha_{\sigma,-}(1) = \alpha_{\sigma,+}(0)$. This immediately implies $\alpha_{\sigma,-}(0) = \alpha_{\sigma,+}(1)$. Using (7), we obtain $\alpha_{\sigma,+}(1) = \left(1 + 2\frac{\rho}{\gamma}\right)\alpha_{\sigma,-}(1)$. We can then solve for the shares of each type$\alpha_{\sigma,+}(1) = \frac{1}{4}\frac{\gamma+2\rho}{\gamma+\rho}$; and $\alpha_{\sigma,+}(0) = \frac{1}{4}\frac{\gamma}{\gamma+\rho}$. Notice also that the market clearing condition among high types is simply $\alpha_{\sigma,+}(1) + \alpha_{\sigma,-}(1) = 1/2$. *Q.E.D.*

## 2.3 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 market. We do it in two steps. We first compute the steady-state value functions for investors that keep on trading. We later compute the ex ante values, taking into account the transition dynamics.

Consider the steady-state value functions for any type $\sigma > \hat{\sigma}$. They solve the following system. For the types holding the assets, we have

$$\begin{aligned} rV_{\sigma,+}(1) &= \mu + \sigma + \frac{\gamma}{2}\left[V_{\sigma,-}(1) - V_{\sigma,+}(1)\right] && (12) \\ rV_{\sigma,-}(1) &= \mu - \sigma + \frac{\gamma}{2}\left[V_{\sigma,+}(1) - V_{\sigma,-}(1)\right] + \rho\left(p + V_{\sigma,-}(0) - V_{\sigma,-}(1)\right). && (13) \end{aligned}$$

For the types not holding the assets, we have

$$\begin{aligned} rV_{\sigma,-}(0) &= \frac{\gamma}{2}\left[V_{\sigma,+}(0) - V_{\sigma,-}(0)\right] && (14) \\ rV_{\sigma,+}(0) &= \frac{\gamma}{2}\left[V_{\sigma,-}(0) - V_{\sigma,+}(0)\right] + \rho\left(V_{\sigma,+}(1) - V_{\sigma,+}(0) - p\right). && (15) \end{aligned}$$

Define $I_{\sigma,\epsilon} \equiv V_{\sigma,\epsilon}(1) - V_{\sigma,\epsilon}(0)$ as the value of owning the asset for type $(\sigma, \epsilon)$. Then, taking differences of equations 12-15, we obtain

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

Note that the asset price $p$ is pinned down by the marginal (minimum type in each market). For now we keep it as a (market-specific) parameter. We can then solve $r\left(I_{\sigma,+} - I_{\sigma,-}\right) = 2\sigma - (\gamma + \rho)\left(I_{\sigma,+} - I_{\sigma,-}\right)$ and obtain the gains from trade for type $\sigma$ in market $\rho$:

$$I_{\sigma,+} - I_{\sigma,-} = \frac{2\sigma}{r + \gamma + \rho}.$$

Note that these gains from trade do not depend on the equilibrium price. Hence they do not depend on the allocation of types to the market. They only depend on the market speed $\rho$ and the individual type $\sigma$. Using the gains from trade $I_{\sigma,+} - I_{\sigma,-}$, we can reconstruct the functions $I_{\sigma,\epsilon}$ and finally the initial value functions. The no-trade outside option of any investor is

$$W_{out} = \frac{\mu \bar{a}}{r}. \qquad (16)$$

The following proposition characterizes the ex ante value functions, taking into account the transition dynamics leading up to the steady state allocations.

**Proposition 1.** *The ex ante value for type $\sigma$ of participating in a market with speed $\rho$ and price $p$, $W$, is the sum of the value of ownership and the value of trading:*

$$W\left(\sigma, \hat{\sigma}, s\right) - W_{out} = \frac{s\bar{a}\hat{\sigma}}{r} + \frac{s}{2r}\max\left(0; \sigma - \hat{\sigma}\right), \tag{17}$$

*where effective speed $s$ defined by*

$$s\left(\rho\right) \equiv \frac{\rho}{r + \gamma + \rho}, \tag{18}$$

*and the marginal type $\hat{\sigma}\left(p, \rho\right)$, defined in equation (5), is increasing in $p$ and decreasing in $\rho$.*

The intuition is that $W$ is composed of two parts. The value of ownership is $\frac{\mu\bar{a}+s\bar{a}\hat{\sigma}}{r}$, which is independent of $\sigma$. It is the value that can be achieved by all types $\sigma < \hat{\sigma}$ with the "sell and leave" strategy. The second part, $\frac{s}{2r}\max\left(0; \sigma - \hat{\sigma}\right)$, is the value of trading repeatedly and it depends on the type $\sigma$. This part of the value function is supermodular in $(s, \sigma)$.

# 3 Competition and Affiliation

In this section we analyze competition among a given set of trading venues and the resulting allocation of investors across these venues. We characterize the pricing decisions and equilibrium profits of trading venues and the affiliations choices of investors. Importantly, we analyze how price protection in the trading game affects these equilibrium outcomes. In other words, we analyze how trading regulations affect the ex ante competition among exchanges. In this section we take the set of venues as given. In the next section we endogenize entry and speed.

In all cases we start with a mass one of investors, aggregate supply $\bar{a}$, and an ex ante distribution of types $G$. The participation decision of type $\sigma$ is described by

$$\mathcal{P} : [0, \bar{\sigma}] \longrightarrow \left\{0, 1, 2\right\},$$

where $\mathcal{P}(\sigma) = 0$ means staying out, 1 means joining market 1, and 2 means joining market 2. Trading venues compete à la Bertrand. 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, so formally $q_0 = 0$ and $W = W_{out}$). Recall that we denote by $\tilde{G}_i(.)$ the distribution of types that join market $i$, so $\tilde{G}_i(\bar{\sigma})$ is the total number of investors that join market $i$. This is the key equilibrium object. Let us now formally define an equilibrium of the affiliation game.

**Definition 1.** An equilibrium of the affiliation game is a set of participation decisions $\mathcal{P}$ by investors and pricing decisions $q$ by trading venues such that

- Participation decisions are optimal: For all $\sigma$ and all $i$, $\mathcal{P}(\sigma) = i$ implies $W(\sigma, \hat{\sigma}_i, s) - q_i \geq W(\sigma, \hat{\sigma}_j, s) - q_j$ for all $j \neq i$; reciprocally, when $W(\sigma, \hat{\sigma}_i, s) - q_i > W(\sigma, \hat{\sigma}_j, s) - q_j$ for all $j \neq i$, then we must have $\mathcal{P}(\sigma) = i$.
- Venues maximize profits: $q_i = \arg\max q_i \tilde{G}_i(\bar{\sigma})$.
- The investor market clears: $\sum_{i=0,1,2} \tilde{G}_i(\sigma) = G(\sigma)$ for all $\sigma \in [0, \bar{\sigma}]$.
- Subsequent asset prices and marginal types satisfy equations (5) and (6).

The remainder of this section considers several versions of the affiliation game: with one or two venues and with or without trading regulations.

$\alpha$,

## 3.1 One Speed

With one speed the marginal trading type must be indifferent between joining the market and not joining the market. So we must have $W(\hat{\sigma}, \hat{\sigma}, s) - W_{out} = q$ and therefore

$$q = \frac{s\bar{a}\hat{\sigma}}{r}. \tag{19}$$

All types below $\hat{\sigma}$ are indifferent between joining and staying out. Let $\delta$ be the mass of investors that join, sell, and leave.[8] Market clearing requires

$$\delta = \left(\frac{1}{2\bar{a}} - 1\right)(1 - G(\hat{\sigma}))$$

This condition holds at an interior solution as long as $\delta < G(\hat{\sigma})$, or in other words, as long as

$$\frac{G(\hat{\sigma})}{1 - G(\hat{\sigma})} > \frac{1}{2\bar{a}} - 1.$$

In the remainder of the paper we assume that either $\bar{a}$ is close enough to $1/2$ or that there is a sufficient mass of low type investors to ensure the existence of interior solutions.

Total profits for the exchange are given by $\pi = q(1 - G(\hat{\sigma}) + \delta)$ which we can write using market clearing as

$$\pi = q\frac{1 - G(\hat{\sigma})}{2\bar{a}}.$$

Note that if $\bar{a} = 1/2$ we get $\delta = 0$, the simplest case to analyze. When $\bar{a}$ is less than $1/2$, we simply need to remember that $\delta$ investors sell and become inactive. The equilibrium is depicted in the top panel of Figure 4.

### Consolidated Market (monopoly)

A consolidated market center with exogenous speed $s$ behaves like a classic monopolist. We index this market structure by $m$. Using equation (19), the program of the monopoly is

$$\max_{q} q\frac{1 - G(\hat{\sigma})}{2\bar{a}}$$

The first-order condition (FOC) for profit maximization is[9]

[8] There can also be a corner solution with full participation, characterized by the market clearing condition $G(\sigma^{\mathsf{T}}_{\min}) = 1 - 2\bar{a}$. All investor pay the participation fee $q_{\min}$ which is also the total profit of the trading venue. Then, $G(\sigma^{\mathsf{T}}_{\min})$ sell and drop out, while the remaining $1 - G(\sigma^{\mathsf{T}}_{\min})$ 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 $q \leq q_{\min} = \frac{s}{r}\bar{a}\sigma^{\mathsf{T}}_{\min}$.

[9] 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 yields quasi-concavity of $\pi = \sigma(1 - G(\sigma))$.

$$1 - G\left(\hat{\sigma}_m\right) = g\left(\hat{\sigma}_m\right)\hat{\sigma}_m. \tag{20}$$

This is a standard result. The monopoly restricts participation to maximize its profits. Note that the choice of $\hat{\sigma}_m$ is independent of the speed in the market. The fee $q_m$ increases one to one with $s$.

#### Fragmented Markets (Bertrand duopoly)

In the fragmented case, exchanges compete in fees à la Bertrand. In equilibrium, fees and profits are both zero. All investors participate and the distribution of investors across trading venues is immaterial. The solution is

$$q_{Bertrand} = 0.$$

In the presence of fixed costs, this would not be an equilibrium. Without differentiation by speed there is a natural monopoly.

## 3.2 Segmented Venues

Formally, suppose there are two venues, 1 and 2, with speeds $\rho_1$ and $\rho_2$ and participation fees $q_1$ and $q_2$, respectively. We define venue 2 as the fast market, so $\rho_2 > \rho_1$. A critical issue is the segmentation of trades and the possibility of different prices. We consider two types of regulations.

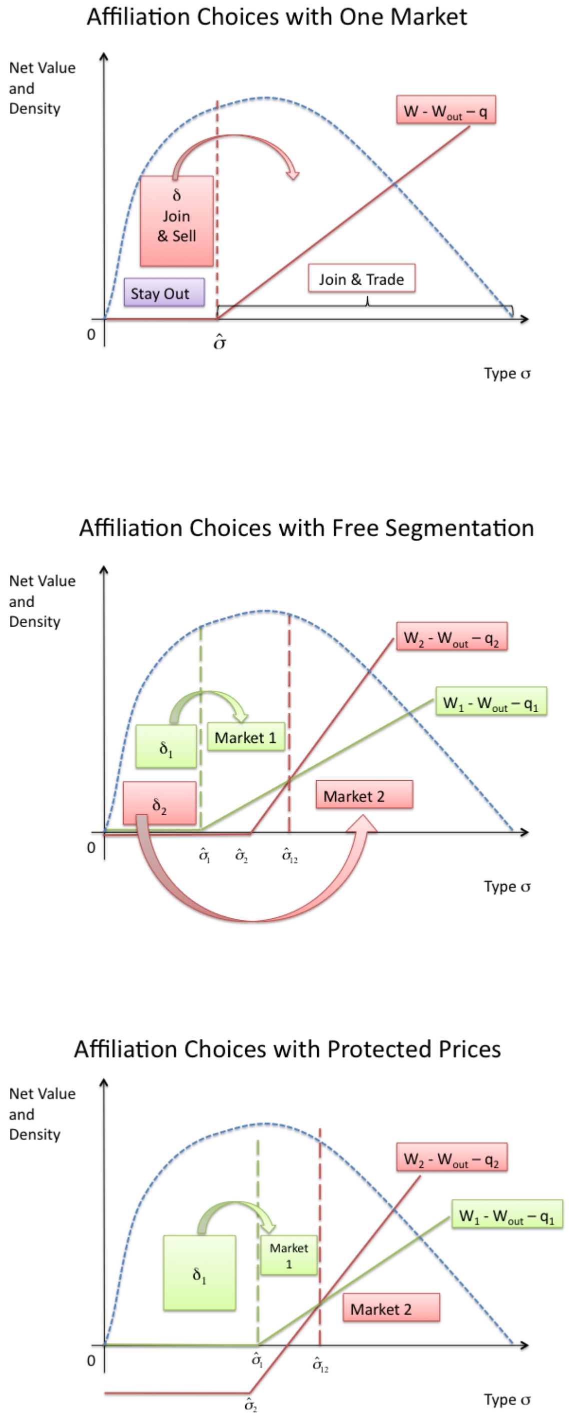
**Definition 2.** We say that there is *segmentation* if a venue refuses to execute trades coming from investors of another venue. Otherwise we say that there is *price protection.*[10]

Under free segmentation an investor joins a market and cannot trade with an investor in the other market. The trading venues are effectively segmented and equilibrium asset prices can be different. Under price protection asset prices must be the same in both venues.

Consider first the case where there is free segmentation. Prices can then be different in the two venues because exchange 2 can refuse to execute the trade of an

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

Figure 4: Investors' Market Affiliation Choice


Affiliation Choices with One Market

Affiliation Choices with Free Segmentation

Affiliation Choices with Protected Prices investor from exchange 1. The key issue is to understand the affiliation choices of investors. We proceed by backward induction. Investors anticipate that each market will be characterized by its speed and its price, which together define the marginal trading type $\hat{\sigma}$. Investors can then estimate their value functions $W$, defined in equation (17). The net value from joining market $i = 1, 2$ is $W(\sigma, \hat{\sigma}_i, s_i) - W_{out} - q_i$. These value functions are depicted in the middle panel of Figure 4.

Let $\hat{\sigma}_1$ be the marginal type that is indifferent between joining market 1 and staying out. It must satisfy equation (19); therefore we have

$$q_1 = \frac{\bar{a} s_1 \hat{\sigma}_1}{r}. \tag{21}$$

It is useful to keep in mind that the value functions are not supermodular for low types. In addition, we know that each market must attract a mass $\delta$ of types that join and sell their assets. Because these types must be indifferent between joining and staying out, we must have $W(\hat{\sigma}_i, \hat{\sigma}_i, s_i) - W_{out} - q_i = 0$ in both markets. Otherwise all the low types would strictly prefer one market to another. The above condition guarantees this for market 1. For market 2 we must also have

$$q_2 = \frac{\bar{a} s_2 \hat{\sigma}_2}{r}. \tag{22}$$

Note an important point: $\hat{\sigma}_2$ is defined as the marginal trader in market 2, that is, the type that is indifferent between trading repeatedly and dropping out after selling. It is clear from Figure 4 (proven below) that $\hat{\sigma}_2$ does not, in fact, join market 2. Rather, $\hat{\sigma}_2$ joins market 1.

With two markets, we must define a new marginal type, $\hat{\sigma}_{12}$, that is indifferent between joining market 1 and market 2. By definition, this type must be such that $W(\hat{\sigma}_{12}, \hat{\sigma}_2, s_2) - q_2 = W(\hat{\sigma}_{12}, \hat{\sigma}_1, s_1) - q_1$. This implies $\frac{s_1 \bar{a} \hat{\sigma}_1}{r} + \frac{s_1}{2r}(\hat{\sigma}_{12} - \hat{\sigma}_1) - q_1 = \frac{s_2 \bar{a} \hat{\sigma}_2}{r} + \frac{s_2}{2r}(\hat{\sigma}_{12} - \hat{\sigma}_2) - q_2$ and therefore, using equations (21) and (22), we obtain

$$\hat{\sigma}_{12} = \frac{r}{\bar{a}} \frac{q_2 - q_1}{s_2 - s_1}. \tag{23}$$

Note that $\hat{\sigma}_1 < \hat{\sigma}_2 < \hat{\sigma}_{12}$. The set of types that join market 2 cannot be an interval. It is composed of all the types above $\hat{\sigma}_{12}$ and some types below $\hat{\sigma}_1$. The affiliation is depicted in the middle panel of Figure 4.

Market clearing in market 2 requires $(1 - G(\hat{\sigma}_{12}) + \delta_2)\bar{a} = \frac{1 - G(\hat{\sigma}_{12})}{2}$. Total profits for the fast exchange under free segmentation are $\pi_2^{seg} = q_2 \left(1 - G\left(\hat{\sigma}_{12}\right) + \delta_2\right) = q_2 \frac{1-G(\hat{\sigma}_{12})}{2\bar{a}}$. Market clearing for the slow exchange requires $\left(G\left(\hat{\sigma}_{12}\right) - G\left(\hat{\sigma}_1\right) + \delta_1\right)\bar{a} = \frac{G(\hat{\sigma}_{12})-G(\hat{\sigma}_1)}{2}$. Total profits for the slow exchange are $\pi_1^{seg} = q_1 \frac{G(\hat{\sigma}_{12})-G(\hat{\sigma}_1)}{2\bar{a}}$. The affiliation of investors to markets 1 and 2 are given by the marginal types 19 and 23. Exchanges 1 and 2 simultaneously solve

$$\begin{aligned} \max_{q_1} \pi_1^{seg} =& \frac{q_1}{2\bar{a}} \left(G\left(\hat{\sigma}_{12}\right) - G\left(\hat{\sigma}_1\right)\right) \\ \max_{q_2} \pi_2^{seg} =& \frac{q_2}{2\bar{a}} \left(1 - G\left(\hat{\sigma}_{12}\right)\right). \end{aligned} \tag{24}$$

Taking first-order conditions from the previous system, we obtain the following lemma:

**Lemma 3.** *Under free segmentation the allocation* $\left(\hat{\sigma}_1^{seg}, \hat{\sigma}_{12}^{seg}\right)$ *among trading venues solves the following system*

$$1 - G\left(\hat{\sigma}_{12}\right) = g\left(\hat{\sigma}_{12}\right)\left(\hat{\sigma}_{12} + \hat{\sigma}_1 \frac{s_1}{s_2 - s_1}\right), \tag{25}$$

$$G\left(\hat{\sigma}_{12}\right) - G\left(\hat{\sigma}_1\right) = \left(g\left(\hat{\sigma}_1\right) + \frac{s_1}{s_2 - s_1} g\left(\hat{\sigma}_{12}\right)\right)\hat{\sigma}_1. \tag{26}$$

## 3.3 Protected Prices

Now consider the case where is there is a single price but two venues with different speeds. The asset price is $p$ in both markets. Market 1 is still characterized by the indifference condition (21) for the marginal type $\hat{\sigma}_1$. However, this condition does not hold for market 2 because low types can join market 1 and then sell their assets to investors in market 2. Instead, we have the condition that the asset price is the same in both markets. From equation (5), this implies the constraint

$$\left(1 + \frac{\gamma}{r + \rho_1}\right)\hat{\sigma}_2 = \left(1 + \frac{\gamma}{r + \rho_2}\right)\hat{\sigma}_1. \tag{27}$$

This means that $\hat{\sigma}_2 < \hat{\sigma}_1$. The indifference condition for $\hat{\sigma}_{12}$ is still $W\left(\hat{\sigma}_{12}, \hat{\sigma}_2, s_2\right) - q_2 = W\left(\hat{\sigma}_{12}, \hat{\sigma}_1, s_1\right) - q_1$. We show in the Online Appendix that this leads to

$$\hat{\sigma}_{12} = \frac{2r}{s_2 - s_1}\left(q_2 - \frac{z}{2\bar{a}} q_1\right), \tag{28}$$

where

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

The structure of the value functions is still as depicted in the bottom panel of Figure 4. There is now only one market-clearing condition. As a result, the sell and leave traders join market 1, where they can sell at a higher price because they can sell to investors in market 2. We then have $\delta_2 = 0$ and the market clearing condition is

$$\left(1 - G\left(\hat{\sigma}_1\right) + \delta_1\right)\bar{a} = \frac{1 - G\left(\hat{\sigma}_1\right)}{2}.$$

The following lemma summarizes the protected price equilibrium.

**Lemma 4.** *Under price protection the allocation* $\left(\hat{\sigma}_1^{prot}, \hat{\sigma}_{12}^{prot}\right)$ *among trading venues solves the following system*

$$\begin{aligned} 1 - G\left(\hat{\sigma}_{12}\right) &= g\left(\hat{\sigma}_{12}\right)\left(\hat{\sigma}_{12} + z\frac{s_1}{s_2 - s_1}\hat{\sigma}_1\right) \\ G\left(\hat{\sigma}_{12}\right) - \frac{G\left(\hat{\sigma}_1\right)}{2\bar{a}} &= \left(\frac{g\left(\hat{\sigma}_1\right)}{2\bar{a}} + z\frac{s_1}{s_2 - s_1}g\left(\hat{\sigma}_{12}\right)\right)\hat{\sigma}_1 + 1 - \frac{1}{2\bar{a}}. \end{aligned}$$

Price protection has two consequences: It increases the profits of the slower exchange and it decreases price competition and participation for given speeds and given exchanges.

We can now compare the outcome of the various market structures. To derive analytical results we assume that the ex ante distribution of types $G$ is exponential.

**Assumption A1**. $G\left(\sigma\right) = 1 - e^{-\frac{\sigma}{\nu}}$.

We can now state the following proposition.

**Proposition 2.** *Competition among exchanges increases participation. With or without price protection, participation in the fast venue is higher than total participation with a monopoly, i.e.* $\hat{\sigma}_{12} < \hat{\sigma}_m$. *Total participation is even higher since* $\hat{\sigma}_1 < \hat{\sigma}_{12}$. *Under A1 price protection increases the profits of the slow venue and decreases total active participation; that is,* $\pi_1^{prot} \geq \pi_1^{seg}$ *and* $\hat{\sigma}_1^{prot} \geq \hat{\sigma}_1^{seg}$. *Price protection does not affect the fee* $q_2 = \frac{\nu}{2r}\left(s_2 - s_1\right)$ *and has an ambiguous impact on participation in the fast venue.*

The intuition for the first half of the proposition is simply that price competition increases participation. A result that is perhaps less obvious is that participation in just the fast venue is already higher than total participation with a monopoly. The intuition for the second half of the proposition is as follows. Price protection is a subsidy to the slow market because its investors are allowed to sell their assets to investors in the fast market. This creates a larger demand for the slow market. When one considers its profits $q_1 \left(1 - G\left(\hat{\sigma}_1\right) + \delta_1\right)$, the presence of this demand makes encourages the slow market to increase its price. This is why $\hat{\sigma}_1^{prot} \geq \hat{\sigma}_1^{seg}$. Protection also softens the price elasticity of the marginal type $\hat{\sigma}_{12}$, which again is good for the slow venue. Thus profits of the slow venue increase under protection for two reasons: more demand and less price elasticity.[11]

The impact on participation in the high-speed market is small, in practice, and positive for the parameter values that we consider, as discussed in the Online Appendix: We typically find $\hat{\sigma}_{12}^{prot} \leq \hat{\sigma}_{12}^{seg}$.

Proposition 2 plays an important role in our paper. The results regarding profits are important in understanding the impact of price protection on entry and therefore on the equilibrium market structure. The results regarding participation are important in understanding the welfare implications of various regulations. We explore these issues in the next section.

### 3.4 Trading Fees and Multi-Market Participation

We study two extensions of the model in the Online Appendix: We characterize competition in trading fees instead of membership fees and we allow investors to join both markets.

In our benchmark model, investors pay membership fees and then trade freely. Alternatively, exchanges could charge a fee per trade (either when the trade is submitted or when it is executed). We show in the Online Appendix that trading fees do not change equilibrium allocations; that is, $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\sigma}_{1,2}$ remain the same as with membership fees. Trading fees do change the profits of the exchanges. In theory, there are two effects. First, trading fees are inefficient since there is no marginal cost of trading. In our model with linear preferences, the inefficiency does not play out because the high types always trade the maximum amounts. But in a more general model (with concave utility), trading volume and welfare would decrease. The second

[11] We checked numerically the robustness of the result $\pi_1^{prot} \geq \pi_1^{seg}$ to alternative assumptions about the underlying distribution of $\sigma$.

effect of trading fees is to allow some price discrimination. The one-time traders pay the fee only once, while the permanent traders pay the fee many times. In equilibrium this allows the exchange to extract more surplus from the investors. Because the first type of inefficiency is not present in our model, we find that profits are typically higher with trading fees.

Having analyzed trading fees, we chose to write our benchmark model with membership fees because trading fees fail to capture the basic idea of affiliation.[12]

We have also analyzed the possibility that some traders may choose to pay both membership fees and trade in both markets. To analyze this case, we first need to characterize the optimal trading strategies of traders who can trade in two markets. The key issue is whether multi-market traders send both buy and sell orders to both markets. If they do, asset allocations and prices $p_1$ and $p_2$ are the same as with a single affiliation because these traders submit the same numbers of buys and sells to both markets. The key condition to check if therefore whether multi-market traders prefer to wait for a good deal rather than sell at a low price in the slow market or buy at a high price in the fast market. This possibility is clearly interesting, especially in its implications on asset prices and arbitrage. In the context of our model, however, we show in the Online Appendix that multi-market traders do not play a quantitatively important role because only investors with extremely large $\sigma$ would in fact choose to join two markets.

[12] With fees paid at execution, it is optimal for investors to submit orders to both markets, wait for the first one to execute, and then cancel the second order. In this case the notion of affiliation to an exchange is lost. Exchanges could then forbid cancellation, but we would need to model exactly how this is done. With fees paid at initiation, the fast exchange has an advantage, but we still need to specify the cancellation policy, since a trader may submit an order and then experience a change in type and prefer to cancel the trade. With membership fees the allocations are the same, the affiliation choices are better defined, and we do not need to specify ad hoc cancellation policies. Therefore we prefer to work with membership fees. Because membership fees relate to market speed in the model, one natural counterpart in reality is co-location costs. Some investors or brokers, after paying a typically monthly fee, can place their trading engine physically close to the exchange matching engine to reduce transmission delays. Trading fees in real markets are less clearly linked to speed advantages than co-location charges.

### 3.5 Equilibrium Asset Prices

From equation (5) we know that the equilibrium asset price in market $i$ is given by

$$p_i = \frac{\mu}{r} + \underbrace{\frac{\hat{\sigma}_i}{r}}_{\text{Participation}} \underbrace{\left(\frac{r + \gamma s_i}{r + \gamma}\right)}_{\text{Speed}}. \tag{29}$$

The key differences between our equilibrium price and the benchmark case in Duffie et al. (2005) is that here both participation decisions among heterogeneous traders and liquidity frictions (driven by the market speed) are endogenously determined. For example, under price protection, $\hat{\sigma}^{prot}$ is given by Lemma 4. Under free segmentation, there are two prices: The asset price in venue $i$ is as in equation (29), where $(\hat{\sigma}_1^{seg}, \hat{\sigma}_2^{seg})$ are given by equations (21) and (22). Consequently, regulations, the market structure, and speed and affiliation choices all affect asset prices. This simple framework then offers a rich set of empirical predictions on asset prices, both at the domestic level and internationally. These relations are explored thoroughly in Pagnotta (2013).

# 4 Endogenous Speed and Entry

In this section we complete the description of the equilibrium market structure by analyzing the entry decisions of trading venues, as well as their optimal investment in speed.

### 4.1 Price Protection and Entry

We develop here the relation between trading regulation and entry for given speeds. There are two potential entrants, with effective speeds $s_1$ and $s_2$, with the convention that $s_1 < s_2$. The entry cost $\kappa$ is the same for both exchanges. Market $i$'s net profit is then given by $\pi_i^{\mathsf{T}} - \kappa$, where $\mathsf{T} \in \{seg; prot\}$ denotes trading regulations.[13] For a

[13] Evidence suggests that entry costs have decreased significantly over time. This is natural since some of these setup costs relate to the development of knowledge and specific computer algorithms, which can be costly to develop but cheaper to subsequently replicate. Entry costs can vary greatly across economies, however, and sometimes relate to the vertical integration aspect of the securities exchange industry. One such example occurs in Brazil, where the incumbent exchange BM&F Bovespa also controls the single national clearinghouse. By denying clearing access to entrants such

Table 1: Entry Game

| Markets 1 ↓ and 2 → | In | Out |
|---|---|---|
| In | $(\pi_1^{\mathsf{T}} - \kappa, \pi_2^{\mathsf{T}} - \kappa)$ | $(\pi_1^m - \kappa, 0)$ |
| Out | $(0, \pi_2^m - \kappa)$ | $(0, 0)$ |

given speed, asset supply $\overline{a} \leq 1/2$, and regulatory framework, the profit functions $\pi$ are as in Section 3. A given venue $i$ finds it optimal to enter whenever net profits are non-negative.

We model entry as a simultaneous game. The payoffs of the entry game are shown in Table 1. From our previous analysis, we know that (i) for a given trading regulation $\mathsf{T}$, $\pi_1^{\mathsf{T}} < \pi_2^{\mathsf{T}}$ simply because 2 is faster and (ii) $\pi_1^{seg} < \pi_1^{prot}$ from Proposition 2. Consequently, we have the following proposition.

**Proposition 3.** *Price protection at the trading stage helps sustain entry at the initial stage.*

As shown in Figure 5, price protection expands the ex ante number of markets for economies with intermediate entry costs (between $\pi_1^{seg}$ and $\pi_1^{prot}$). The expected level of fragmentation hence depends on price regulation.

Depending on parameter values, the entry game may have more than one Nash equilibrium in pure strategies. To simplify our presentation, we assume hereafter that our economies satisfy the inequality $\pi_1^m < \min\left\{\pi_2^{seg}, \pi_2^{prot}\right\}$. Thus only the fast exchange enters whenever $\kappa > \pi_1^{prot}$. We characterize the cases with multiple equilibria in the proof of Proposition 3 in the Online Appendix.

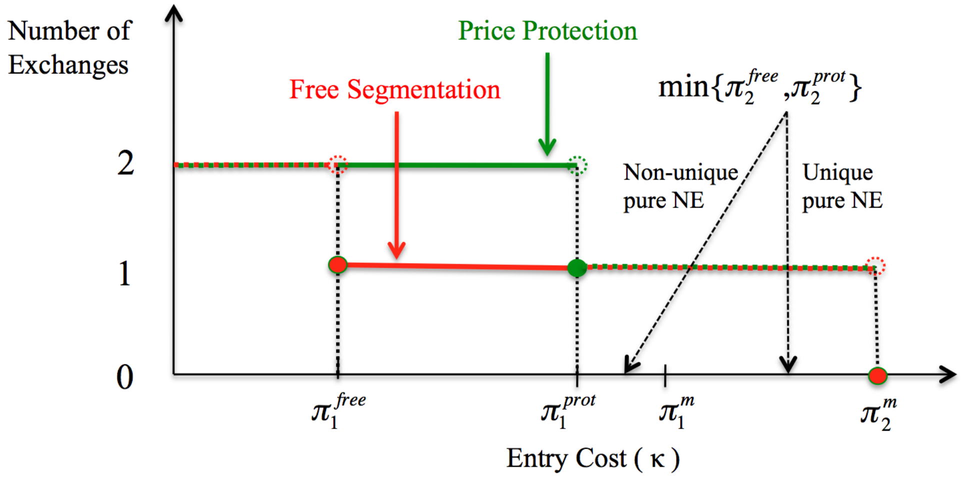
## 4.2 Speed Choices

In this section we analyze speed choices, taking the number of active markets as given. For simplicity, we concentrate on the case where $\bar{a} = 1/2$. In this limiting case trading regulation does not affect markets' profit functions and thus trading regulations become immaterial. When convenient, we assume the following cost to derive analytical results.

**Assumption A2.** *The cost of achieving contact rate $\rho$ is given by $c \max\left\{\rho - \underline{\rho}; 0\right\}$, where $c > 0$ is the constant marginal cost of speed beyond the default level $\underline{\rho}$.*

as BATS and DirectEdge, the incumbent forces new competitors to develop their own clearinghouses.

Figure 5: Entry Cost, Regulation and Equilibrium Fragmentation



The graph shows the equilibrium number of exchanges, as a function of entry costs $\kappa$. Price protection affects the equilibrium number of exchanges that enter the market when entry costs are between the expected profits of the slow venue under segmentation, $\pi_1^{seg}$, and under price protection, $\pi_1^{prot}$. When there are two Nash equilibriums, the outcomes are that either the fast or slow venue decides to enter, and the other venue stays out.

Under Assumption A2, the total cost of entering and reaching the effective speed $s$ is

$$C(s) = c \max\left\{ (r+\gamma)\frac{s}{1-s} - \underline{\rho}; 0 \right\}. \tag{30}$$

These costs are convex in effective speed $s$. We first analyze the case of a monopolist.

## Consolidated Market

Given the monopolist's speed, denoted $s_M$, the marginal type is such that

$$q_M = \frac{s_M}{2r}\hat{\sigma}_M. \tag{31}$$

The program of the monopolist is then

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

We can now characterize the consolidated market equilibrium.

**Proposition 4.** ***Monopoly***. *The equilibrium with consolidated markets and endogenous speed has the following properties: (i) Participation is the same as with exogenous speed: $\hat{\sigma}_M = \hat{\sigma}_m$; (ii) effective speed is given by*

$$2rC'(s_M) = (1 - G(\hat{\sigma}_M))\,\hat{\sigma}_M; \tag{32}$$

*and (iii) under A1-A2 optimal effective speed is given by*

$$s_M = 1 - (2rc(\gamma + r)e)^{1/2}\,\nu^{-1/2}. \tag{33}$$

The monopolist determines market participation based on the distribution of investor types only. Thus investment is speeds are participation-neutral in single exchange economies. Note than in any interior solution, optimal speed does not depend on the default speed level. Naturally, investments in speed increase with investor heterogeneity $\nu$. When the distribution of permanent types $G$ has fatter right tails, the average investor gains from trade increase. Interestingly, the contact rate $\rho_M$ is concave in the frequency of preference shocks $\gamma$: It first increases with $\gamma$ and then decreases and has a global maximum at $\gamma = \frac{\nu}{8cer} - r$. On the one hand, when the frequency of preference shocks increases, investors want to reallocate their assets more frequently, which increases demand for speed. The marginal value of each trade decreases, however, since the desired holding period shrinks. Since speed is costly, there is a maximum speed that can be supported in any market equilibrium.

### Fragmented Market

When trading is fragmented, exchanges have an incentive to differentiate their intermediation services by offering different speeds, since Bertrand competition with a fixed speed drives profits down to zero. We simplify the analysis of this case by assuming that market 1's speed is exogenously given ($s_1 = \frac{\underline{\rho}}{r+\gamma+\underline{\rho}}$) while market 2 chooses an effective speed $s_2$ thats costs is $C(s_2)$.[14] After market 2's speed is chosen,

[14] If given the option, the slow market would intuitively try to reduce speed as much as possible to relax price competition (e.g., Shaked and Sutton (1982)). One can interpret OTC (off-exchange) stock trading volume, currently representing approximately a fourth of the U.S. total, as representing a such group of slow venues. This group includes dark pools, internalization pools, OTC dealers, and crossing networks.

there is simultaneous fee competition, as in Section 3. In the speed choice stage, market 2 solves

$$\max_{s_2} \left(1 - G\left(\hat{\sigma}_{12}\right)\right) q_2 - C\left(s_2\right).$$

The following proposition characterizes the equilibrium.

**Proposition 5.** ***Duopoly.*** *The equilibrium with fragmented markets and endogenous speed has the following properties: (i) Participation is determined by the marginal types $\hat{\sigma}_1$ and $\hat{\sigma}_{12}$, as in Lemma 4; (ii) participation in the fast venue alone is higher than in the monopolist case; (iii) speed in market 2 is determined by equation*

$$2rC'\left(s_2\right) = \left(1 - G\left(\hat{\sigma}_{12}\right)\right)\left\{\hat{\sigma}_{12} + s_1 \frac{\partial \hat{\sigma}_1}{\partial s_2}\right\}; \tag{34}$$

*and (iv) under A1, the duopoly chooses a higher speed than the monopoly.*

The incentives of exchanges to differentiate their services thus increase trading speed. The intuition is as follows. There are two forces at play: scale and differentiation. On the one hand, a monopolist earns higher profits and mechanically wants to invest more in speed. In the limit of Bertrand competition, profits are zero irrespective of speed and there is no incentive to invest in speed. On the other hand, the incentive to differentiate pushes toward higher speed in a duopoly. We study the welfare consequences in Section 5.

# 5 Welfare and Efficient Solution

## 5.1 Welfare Functions

We study the welfare gains of a given market equilibrium with respect to the no-trade benchmark $\mathcal{W}$:

$$\mathcal{W} \equiv \underbrace{\sum_i \int_\sigma \left(W\left(\sigma, \hat{\sigma}_i, s_i\right) - W_{out}\right) dG\left(\sigma\right)}_{\text{Partic. gains \& Allocation efficiency}} - \underbrace{\sum_i \left(\kappa + C\left(s_i\right)\right)}_{\text{Entry+Speed Investment}}.$$

The welfare gains are intuitive. They reflect, first, the sum of investors' expected participation gains. Note that the market's effective speed is part of the function because it affects allocative efficiency. Second, the welfare gains reflect the cost of a

Table 2: Cases of analysis

| | Consolidated Market | Competition |
|---|---|---|
| No Speed Choice | $\mathcal{W}_m$ | $\mathcal{W}_{Bertrand}$ |
| Endogenous Speed | $\mathcal{W}_M$ | $\mathcal{W}_{comp}$ |

given market structure: entry costs and investments in speed. The following lemma characterizes the welfare functions.

**Lemma 5.** *Social welfare in a single market equals*

$$\mathcal{W} = \frac{s}{2r}\int_{\hat{\sigma}}^{\bar{\sigma}} \sigma dG(\sigma) - C(s) - \kappa.$$

*With two trading venues, social welfare is given by*

$$\mathcal{W} = \frac{s_1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_{12}} \sigma dG(\sigma) + \frac{s_2}{2r}\int_{\hat{\sigma}_{12}}^{\bar{\sigma}} \sigma dG(\sigma) - \sum_{i=1,2} C(s_i) - 2\kappa. \tag{35}$$

To simplify the exposition, in this section we consider only the case $\overline{a} = 1/2$, where price regulation is immaterial (we denote social welfare in this case as $\mathcal{W}_{comp}$ but it is the same as $\mathcal{W}_{\mathsf{T}}$, $\mathsf{T} \in \{seg, prot\}$). We analyze the welfare consequences of price protection in Section 6. In the remainder of this section we compare the social gains of different market organizations. We assume that every single venue equilibrium of the entry game involves speed investment.[15] Table 2 summarizes the relevant cases.

## One Speed

As a benchmark, we discuss in the context of our paper the social gains of market organization when investments that improve trading speeds are not available. This is the case considered in the existing literature. Welfare in the monopoly case is given by

$$\mathcal{W}_m = \frac{s}{2r}\int_{\hat{\sigma}_m}^{\overline{\sigma}} \sigma dG(\sigma).$$

In the fragmented case exchanges compete in fees à la Bertrand, where fees and profits are both zero in equilibrium. All investors participate and the distribution of investors

[15] In the proof of Proposition 3 we characterize the cases in which the outcome of the entry game has a fixed-speed monopolist.

across trading venues is immaterial. Social welfare in this case is given by

$$\mathcal{W}_{Bertrand} = \frac{s}{2r} E(\sigma).$$

For any given effective speed $s$, welfare is higher than under monopoly. This is the classic case for inter-market competition when liquidity externalities are moderate (Economides (1996)).

## 5.2 Welfare, Speed, and Competition

### Does speed increase welfare?

We discuss here the welfare consequences of the advent of innovations in technologies that permit faster trading.

**Proposition 6.** *When trading is consolidated, social welfare is always higher with endogenous speed. With fragmented markets, under A1, there exists a unique default speed $\underline{s}_0$ such that welfare increases with endogenous speed if and only if $\underline{s} < \underline{s}_0$.*

The intuition for Proposition 6 is as follows. With a monopoly we know from Proposition 4 that speed does not affect market participation. However, the monopolist has an incentive to invest in speed to extract higher fees from investors with types $\sigma > \hat{\sigma}_M$. Because the monopolist bears the investment costs entirely, there are no negative externalities. Consequently, taxing technology investments is never optimal in this environment.

Under the duopoly, speed allows venues to differentiate and relax Bertrand competition. Whether social welfare increases with technology investments thus depends on the tradeoff between investor participation levels and trading efficiency. When the default effective speed is low, the gains from trading efficiency are large and dominate the negative impact on participation. The opposite happens when the default effective speed is high. In this case, taxing technology investments can increase welfare.

### Does competition increase welfare?

In this section we endogenize entry and ask whether market competition increases welfare. Section 4.2 shows that competition affects investor participation and speed investments. In turn, the outcome of the entry game in Section 4.1 determines the number of active venues. The net social gains of competition are given by $\mathcal{W}_{comp} - \mathcal{W}_M - \kappa$. Let $\overline{\kappa}$ be the entry cost value that makes these gains equal to zero (see Figure 6). The social benefits of higher speeds and higher participation may, in principle, be offset by inefficient cost duplication. We can establish the following proposition.

**Proposition 7.** *Under A1, consolidation increases welfare only when entry costs satisfy $\overline{\kappa} < \kappa \leq \pi_1^{comp}$. Otherwise fragmentation always (weakly) increases welfare.*

The benefits of higher speeds and higher participation are offset by inefficient cost duplication in economies with intermediate entry costs. Note from equation 24 that the slow venue profits can be expressed as

$$\pi_1^{comp} = \frac{s_1}{2r}\hat{\sigma}_1 \left(G\left(\hat{\sigma}_{12}\right) - G\left(\hat{\sigma}_1\right)\right).$$

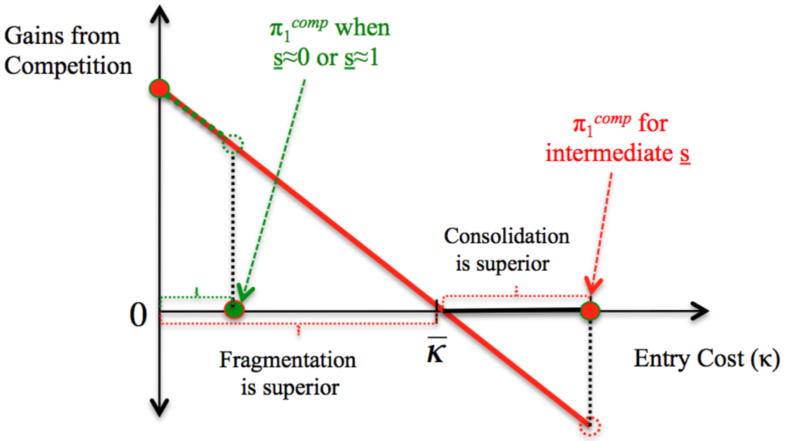
Naturally, $\pi_1^{comp}$ approaches zero when the default speed speed $\underline{s}$ is close to zero. When $\underline{s}$ approaches one, on the other hand, exchange differentiation becomes increasingly difficult and equilibrium profits approach zero as well (the Bertrand outcome with $\hat{\sigma}_1 = 0$). Positive values of $\pi_1^{comp}$ occur for intermediate $\underline{s}$.

In the proof of Proposition 7 we show that, when the marginal cost of technology approaches zero, $\overline{\kappa}$ approaches $\frac{\nu}{2r}(1 - 1/e))$. This limiting value is intuitive. The term $(1 - 1/e))$ represents the participation loss under the monopoly when $\underline{s} \to 1$. The term $\frac{\nu}{2r}$ represents the ex ante participation value of investors of types $\sigma < \nu$ that decide not to participate under monopoly (using equation 17, we have $W(\sigma, \nu, 1) - W_{out} = \frac{\nu}{2r}$ when $\overline{a} = 1/2$).

## 5.3 Constrained Efficiency

How would a planner organize trading in our economy? We first show that the planner chooses to operate only one venue. Without financing constraints, this result is fairly obvious. The setup costs are fixed that there is no marginal cost of adding traders to an exchange. The unconstrained solution is then clearly to open one fast exchange with full participation financed by lump-sum taxes on all agents. This, however, is not a very relevant benchmark. Following a long tradition in public finance, we therefore solve a constrained planner's program where we rule out external subsidies. In other words, we require trading venues to (at least) break even.

Figure 6: Competition, Entry Costs and Welfare



The graph shows the welfare gains from competition $\mathcal{W}_{comp} - \mathcal{W}_M - \kappa$, as a function of entry costs $\kappa$. Gains are zero for entry cost $\overline{\kappa}$. When $\pi_1^{comp}$ is low relative to $\overline{\kappa}$, competition is never socially sub-optimal. When $\pi_1^{comp}$ is high relative to $\overline{\kappa}$, there is a range of entry costs $[\overline{\kappa}, \pi_1^{comp}]$ where consolidation is socially preferred. This occurs for intermediate values of default speed $\underline{s}$.

The more interesting result is then that the planner chooses a single venue even when it has to break even and cannot levy lump sum taxes. The planner faces the same cost structure as the private sector: a setup cost $\kappa$ for each venue, a default effective speed $\underline{s}$ available at no cost, and a speed $s_2 > \underline{s}$ available at cost $C(s_2)$. With financing constraints one might expect the planner to create two trading venues. It could potentially relax the break-even constraints by charging a high price for the fast venue while maintaining participation in the slower, but cheaper, venue. Surprisingly, however, we find that the planner chooses not to do so.

**Proposition 8.** *The planner operates a single venue.*

To understand the intuition, it is better to think of $\hat{\sigma}_1$ and $\hat{\sigma}_{12}$ as control variables instead of $q_1$ and $q_2$. We show in the appendix that the Lagrangian of the planner's problem is

$$\mathcal{L} = \underline{s} \int_{\hat{\sigma}_1}^{\overline{\sigma}} \sigma dG(\sigma) + (s - \underline{s}) \int_{\hat{\sigma}_{12}}^{\overline{\sigma}} \sigma dG(\sigma) - 2rC(s) + \lambda \left\{ \left( (s - \underline{s}) \hat{\sigma}_{12} + \underline{s} \sigma_1 \right) (1 - G(\hat{\sigma}_{12})) - 2rC(s) \right\}$$

where $\lambda$ is the multiplier of the budget constraint of the fast exchange, and we have replaced $q_2 = (s - \underline{s}) \hat{\sigma}_{12} + \underline{s} \hat{\sigma}_1$. The welfare cost of raising $\sigma_1$ is $\underline{s} \hat{\sigma}_1 g(\sigma_1)$, and the financing gain is $\lambda \underline{s} \left(1 - G\left(\hat{\sigma}_{12}\right)\right)$. It is simple to show that the ratio of gains to costs is always higher for $\hat{\sigma}_1$ than for $\hat{\sigma}_{12}$. This implies that the planner chooses to increase $\hat{\sigma}_1$ until it reaches $\hat{\sigma}_{12}$. In other words, the slow market is always inactive. Note that the planner chooses a single venue for investors, even when there are no concerns of cost duplication (the result holds for $\kappa = 0$). This result also extends to the case where prices in the venues can be consolidated or when the planner can choose $s_1$.

In the remainder of this section, we therefore focus, without loss of generality, on the case where the planner operates a single venue. Its program is simply

$$\max_{s,q} \frac{s}{2r} \int_{\hat{\sigma}}^{\overline{\sigma}} \sigma dG\left(\sigma\right) - C\left(s\right) \tag{36}$$

$$\text{s.t. } q\left(1 - G\left(\hat{\sigma}\right)\right) \geq C\left(s\right), \tag{37}$$

where the marginal type $\hat{\sigma}$ is defined as in equation (31). The following proposition compares the constrained efficient solution to the market outcome.

**Proposition 9.** *In the constrained efficient allocation, participation is always higher, but trading speed can be higher or lower than under monopoly.*

We provide in the proof of Proposition 9 an example where $s_M > s*$. Intuitively, the monopolist chooses an inefficiently high speed when the distribution of permanent types has a fat right tail. In this case the monopolist targets investors with high private gains from trade, offering a high-speed–high-price package. The planner may prefer to include the "middle class" of investors even if that means a lower speed because of the break-even constraint.

We know from Proposition 5 that the fast venue chooses a higher speed than the monopolist. Thus it follows from Proposition 9 that the fast venue can also choose a suboptimally high speed. Our numerical analysis shows that investor participation in the duopoly equilibrium can be inefficiently low or inefficiently high.

# 6 Numerical Illustration

In this section we illustrate the implications of the model numerically under assumptions A1 and A2. We compare market outcomes to those of a Walrasian market, which represents a frictionless competitive market with $\kappa_w = 0$ and $c_w = 0$ (which implies $\rho_w = \infty$ and $s_w = 1$). The Walrasian outcomes are as follows.

Table 3: Baseline Parameters

| $\gamma$ | $\underline{\rho}$ | $c$ | $\nu$ | $\bar{a}$ | $\mu$ | $r$ | $\kappa$ |
|---|---|---|---|---|---|---|---|
| 608 | 1070 | 0.0002 | $e$ | 0.45 | 1 | 0.05/252 | 0 |

**Lemma 6.** *The Walrasian equilibrium outcomes are as follows: (i) Investor participation equals $2\overline{a}$, (ii) $p_w = \frac{1}{r}\left[\mu - \nu \log\left(2\overline{a}\right)\right]$, (iii) the instantaneous transaction rate equals $\tau_w = \frac{\overline{a}\gamma}{2}$, and (iv) $\mathcal{W}_w = \frac{\overline{a}\nu}{r}\left(1 - \log\left(2\overline{a}\right)\right)$.*

We select parameters for a numerical illustration of the implications of the model for a representative stock with a daily volume of 100, 000 shares and a mean trade size of 1, 000 shares. Table 3 contains the baseline parameter values. We match default trading delays to the ones prevalent around the implementation of decimalization in U.S. equity markets (see Angel et al. (2011)) so that the default speed $\underline{\rho}$ is relatively low, given current standards (average roundtrip trade is approximately 20 seconds). The model-implied daily preference shock rate is then 608. This is a high value for most individual investors. We think of our investors as either investment buy-side institutions or intermediary firms ("brokers") representing a large group of end users. We analyze a range of values of $c$ that generate endogenous speeds consistent with observed recent values (Figures 7 and 8).

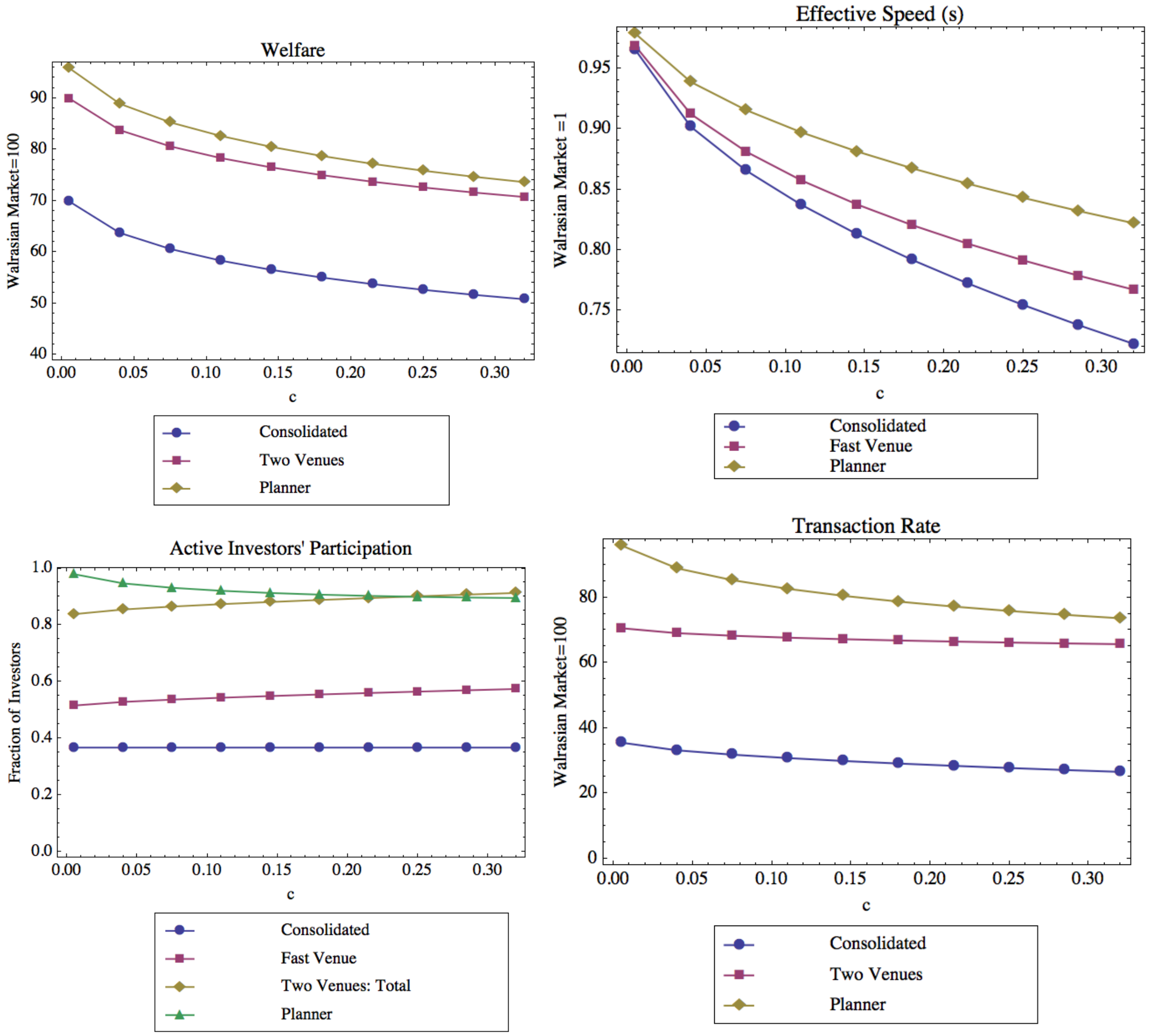
## 6.1 Regulation-Free Equilibrium

Figure 7 shows the main equilibrium outcome when regulation plays no role (case with $\overline{a} = 1/2$). In all cases the horizontal axis represents the marginal cost of speed $c$. The top left panel shows gross welfare in the planner, differentiated duopoly, and monopoly cases. The gross gain from competition is large in this economy: The entry costs that would make competition socially undesirable are approximately 20% of the total gains from trade. Naturally, in all cases welfare decreases with the marginal cost $c$ due to poorer equilibrium allocation efficiency.

As discussed in Section 5, participation with consolidated trading is always inefficiently low and independent of technology choices, as the bottom left panel of Figure 7 illustrates. The equilibrium analysis also shows that participation can be inefficiently low or inefficiently high in a fragmented environment. Participation under duopoly is inefficiently low when the cost of speed is low, since it is easy for exchange to differentiate their services and thus relax fee competition significantly. On the other hand,

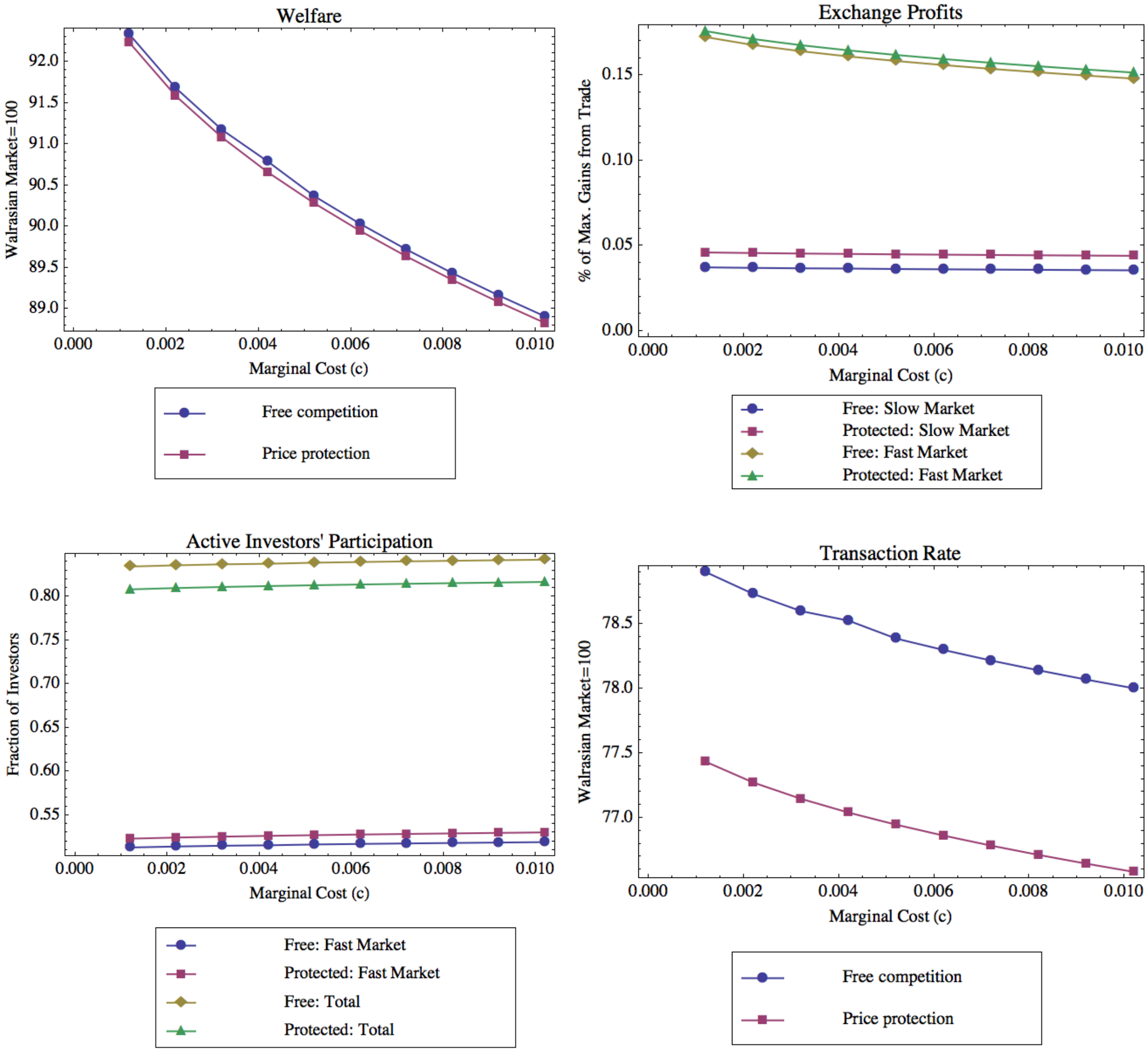
Figure 7: Equilibrium Outcomes and Technology cost

Equilibrium outcomes when asset supply $\overline{a}$ equals 0.5. In this case price regulation is immaterial. The labels are as follows. "Consolidated" correspond to a single monopolist venue. "Two-venues" correspond to a speed-differentiated duopoly. "Planner" correspond to the constrained-efficient solution in section 5. In all cases entry costs $\kappa$ are zero. Parameter choices are described in section 6.



## Figure 8: Equilibrium Outcomes, Technology Cost and Price Regulation

Equilibrium outcomes when asset supply $\overline{a}$ equals 0.45. In this case price regulation matters. The labels are as follows. "Free Competition" corresponds to the case of segmented markets, with two asset prices. "Price Protection" corresponds to the case with a single asset price in all venues. "Free: Total" denotes total investor participation under free competition. In all cases entry costs $\kappa$ are zero. Parameter choices are described in section 6.



when the marginal cost of technology increases, the competition outcome increasingly resembles the fragmented equilibrium without speed choice (Bertrand equilibrium). In fact, we can observe that when $c$ is high enough total participation crosses the optimal level from below and becomes socially excessive. In this case a high level of differentiation is difficult to achieve and the fast exchange is not able to afford socially desirable speed investments. A similar argument can be made when default speed is very high: The scope for market differentiation is also limited and total participation approaches the maximum sustainable level $2\overline{a}$. For similar reasons, we also find that participation can be inefficiently high when the shock arrival rate $\gamma$ is very high, or investor heterogeneity (driven by $v$) is very low.

## 6.2 Is price protection socially desirable?

Figure 8 shows the main equilibrium outcomes when under price protection (case with $\overline{a} < 1/2$). Price protection affects not only investor participation and speed choices but, as shown in Proposition 3, also the number of equilibrium exchanges. The latter effect can be dominant in economies with intermediate entry costs. When a given trading regulation $\mathsf{T} \in \{seg, prot\}$ is set at time $I$, the net social effects can be computed by analyzing its effect on the entry game and the posterior speed–fee competition and affiliation decisions. For a given entry cost $\kappa$, the net social gains of price protection are $\mathcal{W}_{prot}(\kappa) - \mathcal{W}_{free}(\kappa)$. Accordingly, the gains from protection are

$$\begin{cases} \mathcal{W}_{prot} - \mathcal{W}_{free} & \text{if} \quad \kappa \leq \pi_1^{free} \\ \mathcal{W}_{prot} - \mathcal{W}_{M} - \kappa & \text{if} \quad \pi_1^{free} < \kappa \leq \pi_1^{prot} \\ 0 & \text{if} \quad \kappa > \pi_1^{prot}. \end{cases}$$

We find that when price protection affects entry, it has a first-order effect on welfare. This is due to higher participation and allocative efficiency gains when the entry outcome changes from a single- to a dual-venue equilibrium. Whether social welfare increases depends on the tradeoff between these gains and higher entry costs.

However, in economies with low entry costs, price protection does not affect entry and has a negative effect on welfare. This is the case displayed in Figure 8, where entry costs are zero. In such case there is a loss in total market participation that also reduces transaction volumes. We can also observe on the top right panel of Figure 8 that the slow venue realizes higher profits in equilibrium (see Proposition 2).

# 7 Concluding Remarks

We provide a positive and normative analysis of trading speed in financial markets. On the positive side, our model provides an explanation for the joint evolution of trading regulations, fragmentation, and speed. On the normative side, our model clarifies the circumstances under which competition, fragmentation and speed improve or reduce welfare. Our approach to liquidity is distinct from the usual liquidity externality based on increasing returns in the number of traders.

The most important caveat to our analysis is that our model ignores asymmetric information. It is important to point out, however, that our approach is the logical first step, since the tradeoffs and economic forces we have identified would be present in any model, with or without asymmetric information. In particular, speed-sensitive gains from trade are required to consider investment in speed. With free entry, if the average investor does not care about speed, then there would be no investment in speed. Nothing prevents the formation of a relatively slow and cheap exchange. If uninformed traders choose to join fast exchanges, it must be that they value speed. Otherwise they would all join the slow exchange, depriving the fast exchange of liquidity. The idea that speed is provided exclusively to satisfy a fraction of informed traders is therefore inconsistent with free entry. What information would do, then, is change the value of speed for some investors. It is possible that some participants may use speed to take advantage of other investors (e.g., Jovanovic and Menkveld (2012) and Biais et al. (2011)). It is also possible that speed would allow uninformed traders to hedge. Hence we certainly do not claim that asymmetric information is irrelevant, but we do claim that the building blocks of our model are required to analyze speed, fragmentation, and welfare with or without asymmetric information.

# References

Amihud, Y., Lauterbach, B., and Mendelson, H. (2003). The Value of Trading Consolidation : Evidence from the Exercise of Warrants. *Journal of Financial and Quantitative Analysis*, 38(4):829 –846. 8

Amihud, Y. and Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics*, 17(2):223–249. 9

Angel, J. J., Harris, L. E., and Spatt, C. S. (2011). Equity Trading in the 21st Century. *The Quarterly Journal of Finance*, 01(01):1. 2, 36

Biais, B. (1993). Price Formation and Equilibrium Liquidity in Fragmented and Centralized Markets. *Journal of Finance*, 48(1):157–185. 8

Biais, B., Foucault, T., and Moinas, S. (2011). Equilibrium High Frequency Trading. *Working Paper, HEC Paris*. 40

Colliard, J.-E. and Foucault, T. (2012). Trading Fees and Efficiency in Limit Order Markets . *Working Paper, HEC Paris*. 8

Duffie, D., Garleanu, N., and Pedersen, L. H. (2005). Over-the-Counter Markets. *Econometrica*, 73(6):1815–1847. 9, 11, 26

Duffie, D., Garleanu, N., and Pedersen, L. H. (2007). Valuation in Over-the-Counter Markets. *Review of Financial Studies*, 20(6):1865–1900. 6

Economides, N. (1996). The economics of networks. *International Journal of Industrial Organization*, 14(6):673–699. 32

Foucault, T. and Parlour, C. A. (2004). Competition for Listings. *Rand Journal of Economics*, 35(2):329–355. 8

Gabszewicz, J. and Thisse, J.-F. (1979). Price Competition, Quality and Income Disparities. *Journal of Economic Theory*, 20:340–359. 9

Garbade, K. D. and Silber, W. L. (1977). Technology, Communication and the Performance of Financial Markets: 1840-1975. *Journal of Finance*, 33(3):819–832. 8

Glosten, L. R. (1994). Is the electronic open limit order book inevitable? *Journal of Finance*, 49(4):1127–1161. 8

Harris, L. E. (2003). *Trading and Exchanges*. Oxford University Press. 8

Hendershott, T. and Mendelson, H. (2000). Crossing Networks and Dealer Markets: Competition and Performance. *Journal of Finance*, 55(5):2071–2115. 8

Huang, J. and Wang, J. (2010). Market liquidity, asset prices, and welfare. *Journal of Financial Economics*, 95(1):107–127. 9

Jovanovic, B. and Menkveld, A. J. (2012). Middlemen in Limit-Order Markets. *Working Paper, New York University*. 40

Lagos, R. and Rocheteau, G. (2009). Liquidity in Asset Markets With Search Frictions. *Econometrica*, 77(2):403–426. 9, 11, 12

Madhavan, A. (1995). Consolidation, Fragmentation, and the Disclosure of Trading Information. *Review of Financial Studies*, 8(3):579–603. 8

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

Menkveld, A. J. (2011). High Frequency Trading and The New-Market Makers. 3

Muscarella, C. J. and Piwowar, M. S. (2001). Market microstructure and securities values : Evidence from the Paris Bourse. *Journal of Financial Markets*, 4:209–229. 44

O'Hara, M. and Ye, M. (2011). Is market fragmentation harming market quality? *Journal of Financial Economics*, 100(3):459–474. 8

Pagano, M. (1989). Trading volume and Asset Liquidity. *Quarterly Journal of Economics*, 104(2):255–274. 8

Pagnotta, E. S. (2013). Asset Pricing Frictions in Fragmented Markets. *NYU Stern*. 9, 26

Parlour, C. A. and Seppi, D. J. (2003). Liquidity-based competition for order flow. *Review of Financial Studies*, 16(2):329–355. 8

Rocheteau, G. and Wright, R. (2005). Money in search equilibrium, in competitive equilibrium, and in competitive search equilibrium. *Econometrica*, 73(1):175–202. 11

Rust, J. and Hall, G. (2003). Middlemen versus Market Makers: A Theory of Competitive Exchange. *Journal of Political Economy*, 111(2):353–403. 8

Santos, T. and Scheinkman, J. A. (2001). Competition Among Exchanges. *Quarterly Journal of Economics*, 116(3):225–1061. 8

Securities and Exchange Commission (2010). *Concept Release on Equity Market Structure*. Number 34. 3

Shaked, A. and Sutton, J. (1982). Relaxing Price Competition Through Product Differentiation. *Review of Economic Studies*, 49(1):3–13. 9, 29

Shaked, A. and Sutton, J. (1983). Natural Oligopolies. *Econometrica*, 51(5):1469–1483. 9

Vayanos, D. and Wang, T. (2007). Search and endogenous concentration of liquidity in asset markets. *Journal of Economic Theory*, 136(1):66–104. 9

Weill, P.-O. (2007). Leaning Against the Wind. *Review of Economic Studies*, 74:1329–1354. 9

# Appendix: Discussion of International Experiences

## Regulations and Investor Protection

There are essentially two approaches to investor protection: the trade-through model and the principles-based model (see Table 4).

**Trade-through Model.** Under this approach market centers are connected to one another and prevent trading through better prices available elsewhere. This requires complex and costly connections as well as strong monitoring activity from market regulators. Investors can opt out and use a smart routing system to create the linkages. Clearly, price is the primary criterion for best execution. Prices are quoted gross of trading fees (the SEC places a cap on fees). In the U.S. only the top of the book is protected: When a big trading order arrives at a given marketplace, only the amount of shares represented by the depth of the book at the National Best Bid and Offer is protected. As an example, suppose that the NASDAQ and the NYSE are the only market centers and that an investor submits a market order to buy 100,000 shares of a given stock to the NASDAQ. Currently the ask price at the NASDAQ is higher than the ask price at the NYSE (where the ask depth is 10,000 shares). Then the NASDAQ can either match the price at the NYSE or the first execution occurs at the NYSE for 10,000 shares. The remaining 90,000 shares "walk up" the book at the NASDAQ.

**Principles-based Model.** Since criteria other than price are included in the best execution policy, such as the type of investor behind the trade, this approach allows for more discretion and less transparency in the assessment of the results. 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." Some of the criteria to be taken into account are the place of listing, price, liquidity, execution probability, and execution speed. This system in Japan does not apply to professional investors. Both in Europe and Japan, sell-side best execution policies are not obliged to consider every venue. Monitoring of execution quality is generally left to clients, which can be a problem in countries where investors have inadequate knowledge of financial markets. The claimed advantages of the principles-based approach lie in a much simpler set of linkages between markets and promoting innovation by not forcing uniformity.

## Investment in Speeds

### Historical Perspective

Our model captures not only very recent investments to achieve ultra-low latencies but also any increase in transaction frequency enabled by changes in market organization. At one time, all European stock markets (except London) conducted periodic trading by auctions, one to three times a day. Progressively, but not simultaneously, markets moved to continuous trading, which represents a massive increase

Table 4: Investor Protection in Selected Economies

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

Source: www.fidessa.com

* Currently under revision

** Revision of 2009 version

in trading frequencies by comparison to auctions. Our results can then be adapted to such a period of differentiated competition. Interestingly, Muscarella and Piwowar (2001) provide evidence that the transition from call auctions to continuous trading increased asset prices in the Paris Bourse.

### Modern Exchanges

In traditional exchanges such as the NYSE, floor brokers enjoyed advantages in trading speed compared to off-floor investors. Although access to the floor conveyed additional advantages, the differential cost of participating in the exchange floor can be seen as a sort of speed premium in terms of our model. However, nowadays all major exchanges work as electronic platforms that thousands of investors and brokerage firms can access directly.

Although our discussion in the main body of the paper focuses on European[16] and U.S. experiences, our analysis and results relate to multiple recent international experiences. In particular, long used to operating as monopolies, bourses in Asia and Latin America now face the threat of competition from alternative trading platforms. Table 5 illustrates some of the recent investments (with hand-collected data) made in low-latency technologies by major exchanges around the world. This is largely due to the fact that several national regulation agencies have felt encouraged by the examples in the U.S., Canada, and Europe to remove barriers to entry.

### Inter-Market Connectivity

An alternative interpretation of the model has investors choosing providers of "connectivity" between market centers. In this interpretation, exchanges represent Walrasian "end nodes." This alternative interpretation sheds light on important recent investments observed in developed economies. The following are some prominent examples.

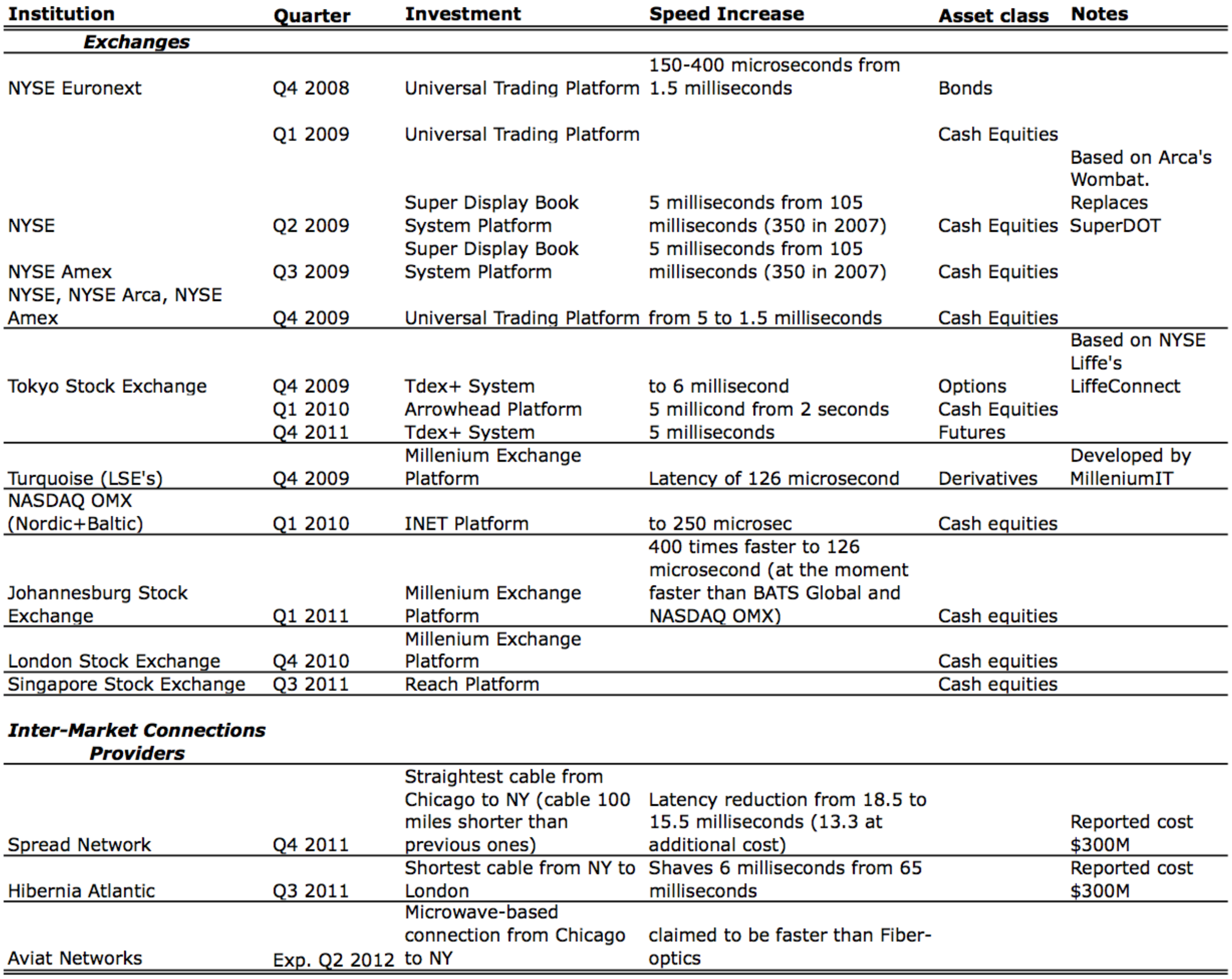
[16] Spain and Italy remain exceptions within Europe since regulators have not encouraged trading venue competition in those markets.

- **Spread Networks** invested approximately $300 million in a a new fiber optic cable that links Chicago and New York through the straightest possible route, saving about 100 miles with respect to existing ones. This allows the company to shave 6 milliseconds off the delay, for a total delay of 15 milliseconds, to attract exchanges, market data sellers, brokers, electronic communication networks, high-frequency trading firms, and alternative trading systems.

- **Hibernian Express** laid a new transatlantic fiber optic cable between New York and London that follows the most common flight path for airlines. Investors such as hedge funds, currency dealers, and exotic proprietary trading firms are queuing up for the switch-on in 2013 and they are expected to pay 50 times as much to link up via the Hibernian Express than they do via existing transatlantic cables (shaving 6 milliseconds off the delay).

- **Aviat Networks**. The speed of light in air is approximately 50 percent faster than the speed of light in fiber optic cables. Exploiting this insight, Aviat Network is currently investing in the creation of a $300 million microwave-based network of towers connecting Chicago and New York. Microwave provides the ability to set up direct line-of-sight paths between transmission points, whereas fiber cables must be routed under streets and around obstacles, adding latency.

Table 5: Speed Investments around the World



| Institution | Quarter | Investment | Speed Increase | Asset class | Notes |
|---|---|---|---|---|---|
| ***Exchanges*** | | | | | |
| NYSE Euronext | Q4 2008 | Universal Trading Platform | 150-400 microseconds from 1.5 milliseconds | Bonds | |
| | Q1 2009 | Universal Trading Platform | | Cash Equities | |
| NYSE | Q2 2009 | Super Display Book System Platform | 5 milliseconds from 105 milliseconds (350 in 2007) | Cash Equities | Based on Arca's Wombat. Replaces SuperDOT |
| NYSE Amex | Q3 2009 | Super Display Book System Platform | 5 milliseconds from 105 milliseconds (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 | Based on NYSE Liffe's LiffeConnect |
| | Q1 2010 | Arrowhead Platform | 5 millicond 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 | Developed by MilleniumIT |
| 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 (at the moment faster than BATS Global and NASDAQ OMX) | Cash equities | |
| London Stock Exchange | Q4 2010 | Millenium Exchange Platform | | Cash equities | |
| Singapore Stock Exchange | Q3 2011 | Reach Platform | | Cash equities | |
| ***Inter-Market Connections Providers*** | | | | | |
| Spread Network | Q4 2011 | Straightest cable from Chicago to NY (cable 100 miles shorter than previous ones) | Latency reduction from 18.5 to 15.5 milliseconds (13.3 at additional cost) | | Reported cost $300M |
| Hibernia Atlantic | Q3 2011 | Shortest cable from NY to London | Shaves 6 milliseconds from 65 milliseconds | | Reported cost $300M |
| Aviat Networks | Exp. Q2 2012 | Microwave-based connection from Chicago to NY | claimed to be faster than Fiber-optics | | |

# Competing on Speed-Online Appendix

Emiliano Pagnotta and Thomas Philippon

New York University Stern School of Business

This Appendix comprises proofs of propositions and lemmas in the main paper, and two model extensions: Trading fees and investor multi-market affiliation.

## Proofs

**Proof of Proposition 1.**

Define $I_{\sigma,\epsilon} \equiv V_{\sigma,\epsilon}(1) - V_{\sigma,\epsilon}(0)$ as the value of owning the asset for type $(\sigma,\epsilon)$. Then, taking differences of equations 12-15 we get

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

We can then solve $r\left(I_{\sigma,+} - I_{\sigma,-}\right) = 2\sigma - (\gamma+\rho)\left(I_{\sigma,+} - I_{\sigma,-}\right)$ and obtain the gains from trade for type $\sigma$ in market $\rho$:

$$I_{\sigma,+} - I_{\sigma,-} = \frac{2\sigma}{r+\gamma+\rho}.$$

Using the gains from trade $I_{\sigma,+} - I_{\sigma,-}$, we can reconstruct the functions $I_{\sigma,\epsilon}$

$$\begin{aligned} I_{\sigma,-} &= \frac{\mu+\rho p}{r+\rho} - \frac{\sigma}{r+\gamma+\rho} \\ I_{\sigma,+} &= \frac{\mu+\rho p}{r+\rho} + \frac{\sigma}{r+\gamma+\rho} \end{aligned}$$

and the average values

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

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

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

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

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

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

From the definition of $\hat{\sigma}$ we also now that

$$\frac{\rho}{r+\rho}(rp-\mu) = s(\rho)\hat{\sigma}$$

with $s(\rho) \equiv \frac{\rho}{r+\gamma+\rho}$, therefore

$$\hat{W} = \frac{\mu\bar{a}}{r} + s\frac{\bar{a}}{r}\hat{\sigma}$$

the marginal type $\hat{\sigma}(p,\rho)$ is defined in (5), is increasing in $p$ and decreasing in $\rho$. The key point is that $\hat{W}$ does not depend on the type $\sigma$, but only on the price and speed of the market. Of course we also have

$$\hat{W} = \bar{a}\bar{V}_{\hat{\sigma}}(1)$$

Let us now consider the steady state types, $\sigma > \hat{\sigma}$. Their average endowment is $\bar{a}$. There are two interpretations. Either they all have $\bar{a}$ or they have a probability $\bar{a}$ to have one unit. Since all agents are risk neutral, the two interpretations are equivalent.

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

Using the expression above, we get

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

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

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

*Q.E.D.*

**Proof of Lemma 4**

First notice that $W\left(\hat{\sigma}_{12}, \hat{\sigma}_2, s_2\right) - q_2 = W\left(\hat{\sigma}_{12}, \hat{\sigma}_1, s_1\right) - q_1$ can be written as:

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

Since $q_1 = \frac{\bar{a} s_1 \hat{\sigma}_1}{r}$, we get $\frac{s_2 - s_1}{2r}\hat{\sigma}_{12} = q_2 - \frac{\bar{a} s_2 \hat{\sigma}_2}{r} + \frac{s_2 \hat{\sigma}_2 - s_1 \hat{\sigma}_1}{2r}$. Using $\hat{\sigma}_2 = m\hat{\sigma}_1$, we get $\frac{s_2 - s_1}{2r}\hat{\sigma}_{12} = q_2 - q_1\left(\frac{1}{2\bar{a}} - \frac{s_2}{s_1} m\left(\frac{1}{2\bar{a}} - 1\right)\right)$ where $m \equiv \frac{1 + \frac{\gamma}{r + \rho_2}}{1 + \frac{\gamma}{r + \rho_1}}$. Since $\frac{s_2}{s_1} m = \frac{\rho_2}{\rho_1} \frac{r + \rho_1}{r + \rho_2}$, we get

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

where

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

Note that $z \leq 1$, $z \approx 1$ when $a \approx 0.5$, and $z \approx 2a$ when $r/\rho$ is small (the realistic case). The profits of market 1 are

$$\pi_1^{prot} = q_1\left(G\left(\hat{\sigma}_{12}\right) - G\left(\hat{\sigma}_1\right) + \delta_1\right)$$

We know introduce notations to simplify the equation

$$\begin{aligned} \alpha &\equiv 2\bar{a} \\ k &\equiv \frac{s_1}{s_2 - s_1} \end{aligned}$$

In the protected price equilibrium, firms therefore maximize

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

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

$$\begin{aligned} 1 - G\left(\hat{\sigma}_{12}\right) &= g\left(\hat{\sigma}_{12}\right)\left(\hat{\sigma}_{12} + zk\hat{\sigma}_1\right) \\ 1 - \alpha + \alpha G\left(\hat{\sigma}_{12}\right) - G\left(\hat{\sigma}_1\right) &= \left(g\left(\hat{\sigma}_1\right) + \alpha zkg\left(\hat{\sigma}_{12}\right)\right)\hat{\sigma}_1 \end{aligned}$$

*Q.E.D.*

**Proof of Proposition 2.** Define

$$\nu\left(\hat{\sigma}\right) \equiv \frac{1 - G\left(\hat{\sigma}\right)}{g\left(\hat{\sigma}\right)}$$

Let us compare the three market structures. The monopoly allocation $\hat{\sigma}_m$ is the solution to

$$\hat{\sigma}_m = \nu\left(\hat{\sigma}_m\right)$$

Rearranging the first order conditions, the free segmentation allocation $\left(\hat{\sigma}_1^{seg}, \hat{\sigma}_{12}^{seg}\right)$ is the solution to

$$\begin{aligned}\hat{\sigma}_{12} &= \nu\left(\hat{\sigma}_{12}\right)-k\hat{\sigma}_1 \\ \hat{\sigma}_1\left(\frac{g\left(\hat{\sigma}_1\right)}{g\left(\hat{\sigma}_{12}\right)}+k\right) &= \frac{g\left(\hat{\sigma}_1\right)}{g\left(\hat{\sigma}_{12}\right)}\nu\left(\hat{\sigma}_1\right)-\nu\left(\hat{\sigma}_{12}\right)\end{aligned}$$

The price protection allocation $\left(\hat{\sigma}_1^{prot},\hat{\sigma}_{12}^{prot}\right)$ is the solution to

$$\begin{aligned}\hat{\sigma}_{12} &= \nu\left(\hat{\sigma}_{12}\right)-z\left(\alpha\right)k\hat{\sigma}_1 \\ \hat{\sigma}_1\left(\frac{g\left(\hat{\sigma}_1\right)}{g\left(\hat{\sigma}_{12}\right)}+\alpha z\left(\alpha\right)k\right) &= \frac{g\left(\hat{\sigma}_1\right)}{g\left(\hat{\sigma}_{12}\right)}\nu\left(\hat{\sigma}_1\right)-\alpha\nu\left(\hat{\sigma}_{12}\right)\end{aligned}$$

where we highlight in red the differences to help the comparison. Notice first $\hat{\sigma}_{12}<\hat{\sigma}_m$ irrespective of whether prices are free or protected.

**Exponential Distribution.**

Under A1, we have $G\left(\sigma\right)=1-e^{-\sigma/\nu}$ and therefore $\nu\left(\hat{\sigma}\right)=\nu$ and the system is

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

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

$$\left(1+zk\right)x = 1-\Delta \tag{38}$$

$$e^{\Delta}-\alpha = \left(e^{\Delta}+\alpha zk\right)x \tag{39}$$

*Impact of protection on $\hat{\sigma}_1$*

The second equation of the system is

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

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

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

since $\frac{e^{\Delta}}{\alpha}>1$ we have $\frac{1}{1+kz}-\frac{1}{\frac{e^{\Delta}}{\alpha}+kz}>0$. Similarly $\frac{1}{\alpha}-\frac{1}{\alpha+\frac{e^{\Delta}}{kz}}>0$. So $\frac{\alpha(1+kz)}{e^{\Delta}+\alpha zk}$ is increasing in $\alpha$. Therefore the equilibrium condition $e^{\Delta}-\alpha=\left(e^{\Delta}+\alpha kz\right)x$ implies a schedule $x$ increasing in $\Delta$ and decreasing in $\alpha$. The first equilibrium condition $\left(1+zk\right)x=1-\Delta$

gives a schedule $x$ decreasing in $\Delta$ and decreasing in $\alpha$. Straightforward analysis then shows that $x$ must be decreasing in $\alpha$. The free price structure corresponds to $\alpha = 1$, while the protected price structure corresponds to $\alpha = 2a < 1$. Therefore, since $\hat{\sigma}_1 = \nu x$, $\hat{\sigma}_1$ must be higher under price protection.

_Impact of protection on $\hat{\sigma}_{12}$_

The analysis of $\hat{\sigma}_{12}$ is ambiguous. It is clear that when $k \to 0$ we have $\hat{\sigma}_{12} \to \nu$, which is the monopoly solution. Define $y = \frac{\hat{\sigma}_{12}}{\nu} = x + \Delta$, and get the system

$$\begin{aligned}(1 + kz)\, y &= 1 + kz\Delta \\ 1 - y &= kz \frac{e^{\Delta} - \alpha}{e^{\Delta} + \alpha kz}\end{aligned}$$

The first curve is $y$ increasing in $\Delta$ and decreasing in $\alpha$. The second curve can be written gives $y = 1 - kz + \frac{kz\alpha(kz+1)}{e^{\Delta} + \alpha kz}$, which shows $y$ decreasing in $\Delta$. With respect to $\alpha$, however, it is not clear. In the realistic case where $\frac{r}{\rho_1}$ is small, we have $z(\alpha) = \alpha$ so

$$\begin{aligned}(1 + k\alpha)\, y &= 1 + k\alpha\Delta \\ 1 - y &= k\alpha \frac{e^{\Delta} - \alpha}{e^{\Delta} + k\alpha^2}\end{aligned}$$

We study the case where $\alpha$ is close to one. The free price solution is

$$\begin{aligned}(1 + k)\, \bar{y} &= 1 + \bar{\Delta} k \\ 1 - \bar{y} &= k \frac{e^{\bar{\Delta}} - 1}{e^{\bar{\Delta}} + k}\end{aligned}$$

and we look for small deviations: $\alpha = 1 - \epsilon$, $\Delta = \bar{\Delta} + \hat{\Delta}$, $y = \bar{y} + \hat{y}$. The first equation is simply

$$\begin{aligned}(1 + k)\, \hat{y} - k\bar{y}\epsilon &= k\left(\hat{\Delta} - \bar{\Delta}\epsilon\right) \\ (1 + k)\, \hat{y} &= k\hat{\Delta} + k\left(\bar{y} - \bar{\Delta}\right)\epsilon\end{aligned}$$

The second one gives

$$\begin{aligned}1 - \bar{y} - \hat{y} &= \frac{k}{e^{\bar{\Delta}} + k}\left(e^{\bar{\Delta}} - 1 + \hat{\Delta} e^{\bar{\Delta}} + \left(2 - e^{\bar{\Delta}}\right)\epsilon - \frac{e^{\bar{\Delta}} - 1}{e^{\bar{\Delta}} + k}\left(e^{\bar{\Delta}}\hat{\Delta} - 2k\epsilon\right)\right) \\ -\left(e^{\bar{\Delta}} + k\right)^2 \hat{y} &= ke^{\bar{\Delta}}\left((1 + k)\,\hat{\Delta} + \left(2 - e^{\bar{\Delta}} + k\right)\epsilon\right)\end{aligned}$$

From the first schedule we get $k\hat{\Delta} = (1 + k)\,\hat{y} - k\left(\bar{y} - \bar{\Delta}\right)\epsilon$. The second schedule then becomes

$$-\left(\left(e^{\bar{\Delta}} + k\right)^2 + e^{\bar{\Delta}}(1 + k)^2\right)\hat{y} = ke^{\bar{\Delta}}\left(2 + k - e^{\bar{\Delta}} - (1 + k)\left(\bar{y} - \bar{\Delta}\right)\right)\epsilon$$

The evolution of $y$ therefore depends on the sign of $\chi = 2 + k - e^{\bar{\Delta}} - (1+k)\left(\bar{y} - \bar{\Delta}\right)$. From the equilibrium condition at $\alpha = 1$, we get $\bar{y} = \frac{1+\bar{\Delta}k}{1+k}$, and the $\Delta$ under free prices solves

$$\left(\bar{\Delta} + k\right) e^{\bar{\Delta}} = 1 + k\left(2 - \bar{\Delta}\right)$$

In the special case $k = 0$, we get $\bar{y} = 1$ and $\bar{\Delta}e^{\bar{\Delta}} = 1$ implies $\bar{\Delta} = 0.5671$ then $\chi = 1 - e^{\bar{\Delta}} + \bar{\Delta} = -0.1961 < 0$. In this case $\hat{y}$ increases with $\epsilon$: $\sigma_{12}$ is higher under price protection. However, as long as $k$ is not too small ($k > 0.185$), we have $2 + k - e^{\bar{\Delta}} - (1+k)\left(\bar{y} - \bar{\Delta}\right) > 0$ and $\hat{y}$ decreasing with $\epsilon$: $\sigma_{12}$ is lower, and participation in the fast market is higher under price protection.

### Comparing Profits

It is convenient to define a system that nests price protection and free competition as special cases. Fist, define the scaled controls

$$t_1 \equiv \frac{2r}{\alpha s_1} q_1,$$
$$t_2 \equiv \frac{2r}{s_1} q_2.$$

Next the scaled profits by $F_i \equiv \frac{2r}{s_1}\pi_i$. With these notations, the profit functions are

$$F_1(t_1, t_2, \alpha) = t_1\left(1 - \alpha + \alpha G(\hat{\sigma}_{12}) - G(t_1)\right)$$
$$F_2(t_1, t_2, \alpha) = t_2\left(1 - G(\hat{\sigma}_{12})\right)$$

and we have

$$\hat{\sigma}_{12} = k\left(t_2 - z(\alpha)\, t_1\right)$$
$$\hat{\sigma}_1 = t_1$$

The general system is the one with protected prices with $\alpha < 1$ and $z(\alpha) = 1 - \frac{1+\frac{r}{\rho_1}}{1+\frac{r}{\rho_2}}(1-\alpha)$. The free segmentation case corresponds to $\alpha = 1$ and $z = 1$. We can always return to the system in $\sigma$ using $t_2 = \frac{\hat{\sigma}_{12}}{k} + z\hat{\sigma}_1$ and $t_1 = \hat{\sigma}_1$.

Let us now derive the FOCs. Using $\frac{\partial \pi_1^{prot}}{\partial t_1} = 0$ and $\frac{\partial \pi_2^{prot}}{\partial t_2} = 0$ we get

$$1 - \alpha + \alpha G(\hat{\sigma}_{12}) - G(\hat{\sigma}_1) = t_1\left(\alpha z(\alpha)\, k g(\hat{\sigma}_{12}) + g(\hat{\sigma}_1)\right)$$
$$1 - G(\hat{\sigma}_{12}) = t_2 k g(\hat{\sigma}_{12})$$

With exponential distributions we have that $t_2$ is constant: $t_2 = \frac{\nu}{k}$. Note that this implies $q_2 \frac{2r}{s_1} = \frac{\nu}{k}$ so $q_2 = \frac{\nu}{2r}(s_2 - s_1)$. The fees of the fast venue are proportional to the difference in effective speed.

To understand the impact of price protection of profits, take the total differential

$$\frac{dF_1}{d\alpha} = \frac{\partial F_1}{\partial t_1}\frac{dt_1}{d\alpha} + \frac{\partial F_1}{\partial t_2}\frac{dt_2}{d\alpha} + \frac{\partial F_1}{\partial \alpha}$$

Optimality implies $\frac{\partial F_1}{\partial t_1} = 0$, and we have just seen that $\frac{dt_2}{d\alpha} = 0$. Therefore $\frac{dF_1}{d\alpha} = \frac{\partial F_1}{\partial \alpha}$ and

$$\frac{\partial F_1}{\partial \alpha} = t_1 \left( -1 + G\left(\hat{\sigma}_{12}\right) + \alpha g\left(\hat{\sigma}_{12}\right) \frac{\partial \hat{\sigma}_{12}}{\partial \alpha} \right) = t_1 \left( -1 + G\left(\hat{\sigma}_{12}\right) - \alpha g\left(\hat{\sigma}_{12}\right) k t_1 z'\left(\alpha\right) \right)$$

Since $z'\left(\alpha\right) > 0$, we see that $\frac{\partial F_1}{\partial \alpha} < 0$: price protection increases the profits of the slow market. The economic intuition is simple. The term $-1 + G\left(\hat{\sigma}_{12}\right)$ corresponds to the "sell and leave" investors who come to the slow venue under protection. The term with $z'$ corresponds to the softer price effect on the marginal type $\hat{\sigma}_{12}$. Q.E.D.

**Proof of Proposition 3.** The relationship between entry costs $\kappa$ and profits determines the number of active venues in equilibrium. Let $\overline{\pi}_i \equiv \max\left\{\pi_i^{prot}, \pi_i^{seg}\right\}$ and $\underline{\pi}_i \equiv \min\left\{\pi_i^{prot}, \pi_i^{seg}\right\}$. We analyze below the existence of NE in pure strategies of the normal-form game shown in figure 5.

- *Two-venues equilibriums.* Suppose $\kappa \leq \overline{\pi}_1$, By Proposition 2, we have that $\underline{\pi}_1 = \pi_1^{seg}$. It is immediate then that entry is always optimal for the slow venue when $\kappa \leq \pi_1^{seg}$ and that, for any $\pi_1^{seg} < \kappa \leq \pi_1^{prot}$, we have $\pi_1^{seg} - \kappa < 0$ and $\pi_1^{prot} - \kappa \geq 0$. A duopoly is never sustainable whenever $\kappa > \pi_1^{prot}$.
- *Single-venue equilibriums.* Suppose $\pi_1^{prot} < \kappa \leq \pi_2^m$.
    - Case 1: $\pi_2^m \geq \kappa > \pi_1^m$. The only NE has the slow venue out and the fast venue entering, with payoff $\pi_2^m$.
    - Case 2: $\overline{\pi}_1 \leq \kappa \leq \underline{\pi}_2$. In this case there is a single NE where only the fast venue enters.
    - Case 3: $\overline{\pi}_2 < \kappa < \pi_1^m$. There are two NE where only one venue enters, either the slow of fast one.
    - Case 4: $\underline{\pi}_2 < \kappa \leq \min\left\{\overline{\pi}_2, \pi_1^m\right\}$. When $\pi_2^{\intercal} = \overline{\pi}_2$, there is a single NE where only the fast venue enters. When $\pi_2^{\intercal} = \underline{\pi}_2$, there are two NE where only one venue enters, either the slow of fast one.
- *No-entry equilibrium.* Whenever $\kappa > \pi_2^m$ the only NE has both venues out. *Q.E.D.*

**Proof of Proposition 4.** Part (i) is straightforward. The interior solution FOC for speed is

$$-g\left(\hat{\sigma}_M\right) \frac{\partial \hat{\sigma}_M}{\partial s_M} q = C'\left(s_M\right) \tag{40}$$

Combining 40 with the FOC for $q$, and using 31 to compute the partial derivative delivers (ii). Using A2 we have that the LHS of 32 is given by

$$2rC'(s_M) = \frac{2rc(\gamma+r)}{(1-s)^2}$$

Using A1 we have that the RHS of 32equals $\nu/e$. Combining these expressions yields 33. Using 18 in 33 we have that the optimal speed $\rho_M$ is given by

$$\rho_M = \frac{(\gamma+r)^{1/2}\left(\nu^{1/2} - (2rce(\gamma+r))^{1/2}\right)}{(2rce)^{1/2}}$$

When $c$ is sufficiently low, it is easy to see that the value of $\frac{\partial \rho_M}{\partial \gamma}$ is positive (negative) for small (large) values of $\gamma$ and achieves a maximum at $\frac{\nu}{8cer} - r$. We also have

$$\frac{\partial^2 \rho_M}{\partial \gamma^2} = -\frac{\nu^{1/2}}{4(2rce)^{1/2}(\gamma+r)^{3/2}}$$

which is negative for all $\gamma > 0$. *Q.E.D.*

**Proof of Proposition 5.** Parts (i) and (ii) are straightforward. To prove part (iii) we proceed in 3 steps.

Step 1: Necessary condition

Under A1 and with $\alpha = 1$, we have

$$\hat{\sigma}_{12} = \nu - \frac{s_1}{s_2 - s_1}\hat{\sigma}_1$$

and

$$q_2 = \frac{\nu}{2r}(s_2 - s_1)$$

The profits of the fast venue are $\pi_2 = q_2(1 - G(\hat{\sigma}_{12}))$ therefore

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

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

$$2rC'(s_2) = \nu(1 - G(\hat{\sigma}_{12})) - \nu(s_2 - s_1)g(\hat{\sigma}_{12})\frac{\partial \hat{\sigma}_{12}}{\partial s_2} \qquad (41)$$

The consolidated solution is $2rC'(\overline{s}_2) = \nu e^{-1}$. With two active venues we have

$$\frac{\partial \hat{\sigma}_{12}}{\partial s_2} = \frac{k}{s_2 - s_1}\hat{\sigma}_1 - k\frac{\partial \hat{\sigma}_1}{\partial s_2}$$

Then,

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

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

$$2rC'\left(s_2\right)=\nu e^{kx-1}\left(1-kx+s_1\frac{\partial x}{\partial s_2}\right) \tag{42}$$

Since $C'$ is an increasing function, market 2 chooses a higher speed whenever the RHS of 42 is greater than $\nu e^{-1}$. That is,

$$e^{kx}\left(1-kx+s_1\frac{\partial x}{\partial s_2}\right)-1>0 \tag{43}$$

Step 2: Finding $\frac{\partial x}{\partial s_2}$
Differentiating the system 38-39 we have

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

After appropriate substitutions we get

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

Re-arranging,

$$s_1\frac{\partial x}{\partial s_2}=\frac{k^2x\left(1+e^{\Delta}\left(1-x\right)\right)}{e^{\Delta}\left(1+\Delta\right)+k\left(1+e^{\Delta}\right)} \tag{44}$$

Step 3: Verifying inequality 43
Substituting 44 in 43 let

$$S\left(k\right)\equiv e^{kx}\left(1-kx+\frac{k^2x\left(1+e^{\Delta}\left(1-x\right)\right)}{e^{\Delta}\left(1+\Delta\right)+k\left(1+e^{\Delta}\right)}\right)-1 \tag{45}$$

Re-arranging we have

$$S(k)=e^{kx}\left(\frac{e^{\Delta}\left(1+\Delta\right)+k\left(1+e^{\Delta}\right)-kxe^{\Delta}\left(1+\Delta-kx\right)}{e^{\Delta}\left(1+\Delta\right)+k\left(1+e^{\Delta}\right)}\right)-1 \tag{46}$$

To satisfy the inequality we need $S(k) > 0$ for all $k > 0$ and $S(0) = 0$ (corresponding to the monopolist case where $s_1 = 0$). Let $x(k)$ and $\Delta(k)$ denote the solutions to the system 38-39 for a given $k \geq 0$. Since $x(k)$ and $\Delta(k)$ are continuous functions, $S(k)$ is continuous. Using 38-39 one can see that

$$\lim_{k\to\infty} x(k) = 0 \tag{47}$$
$$\lim_{k\to\infty} \Delta(k) = \underline{\Delta}$$

where $\underline{\Delta}$ is defined by $e^{\underline{\Delta}} + \underline{\Delta} = 2$. Notice that $\lim_{k\to\infty} x(k)\,k = 1 - \underline{\Delta}$. Similarly,

$$\lim_{k\to 0} x(k) = 1 - \overline{\Delta} \tag{48}$$
$$\lim_{k\to 0} \Delta(k) = \overline{\Delta}$$

where where $\overline{\Delta}$ is defined by $e^{\overline{\Delta}}\overline{\Delta} = 1$. Taking limits of 45 we find

$$\lim_{k\to 0} S(k) = e^0 - 1 = 0$$
$$\lim_{k\to\infty} S(k) = e^{1-\underline{\Delta}} - 1 > 0$$

A sufficient condition for $S(k) > 0$ for all $k > 0$ is to show that the term between brackets in 46 is greater than one. This is the case whenever

$$e^{\Delta}(1 + \Delta + k) + k + e^{\Delta}k\left[(1 - x) + (xk)^2 - x\Delta\right] > e^{\Delta}(1 + \Delta + k) + k \tag{49}$$

Note from 38 that $1 - x = kx + \Delta$. Then,

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

We conclude that $S(k) > 0$ for all $k > 0$. *Q.E.D.*

**Proof of Lemma 5.** The welfare formula reflects the joint welfare of four groups: exchange owners, dropout investors, active traders in the slow market and active traders in the fast market. Transfers from investors to exchange owners do not represent net social gains and hence are not reflected in 35. When dropout investors join market $i$, their before-fees gains are independent of their types and equal to

$$W(\sigma, \hat{\sigma}_i, s_i) - W_{out} = \frac{1}{r} s_i \bar{a} \hat{\sigma}_i$$

The total mass of these investors equals $\left(\frac{1}{2\bar{a}} - 1\right)(1 - G(\hat{\sigma}_1))$, and under free competition a mass equal to $\left(\frac{1}{2\bar{a}} - 1\right)(1 - G(\hat{\sigma}_{12}^{seg}))$ joins the fast market. Thus, the social gains for this group in the free and protected cases are given by

$$\frac{\bar{a}}{r}\left(\frac{1}{2\bar{a}}-1\right)\left[\left(G\left(\hat{\sigma}_{12}\right)-G\left(\hat{\sigma}_{1}\right)\right)s_{1}\hat{\sigma}_{1}^{free}+\left(1-G\left(\hat{\sigma}_{12}\right)\right)s_{2}^{seg}\hat{\sigma}_{12}^{seg}\right], \tag{50}$$

$$\frac{\bar{a}}{r}\left(\frac{1}{2\bar{a}}-1\right)\left[\left(1-G\left(\hat{\sigma}_{1}\right)\right)s_{1}\hat{\sigma}_{1}^{prot}\right] \tag{51}$$

Using Proposition 1, the welfare of active investors in the slow market is given by

$$\int_{\hat{\sigma}_{1}^{R}}^{\hat{\sigma}_{12}^{\top}}\left[\frac{s_{1}}{r}\left(\bar{a}-\frac{1}{2}\right)\hat{\sigma}_{1}^{R}+\frac{s_{1}}{2r}\sigma\right]dG\left(\sigma\right) \tag{52}$$

The welfare of active investors in the fast market under free and protected prices are given by

$$\int_{\hat{\sigma}_{12}^{seg}}^{\bar{\sigma}}\left[\frac{s_{2}^{seg}}{r}\left(\bar{a}-\frac{1}{2}\right)\hat{\sigma}_{2}^{free}+\frac{s_{2}^{seg}}{2r}\sigma\right]dG\left(\sigma\right), \tag{53}$$

$$\int_{\hat{\sigma}_{12}^{prot}}^{\bar{\sigma}}\left[\frac{s_{1}}{r}\left(\bar{a}-\frac{1}{2}\right)\hat{\sigma}_{1}^{prot}+\frac{s_{2}^{prot}}{2r}\sigma\right]dG\left(\sigma\right) \tag{54}$$

Adding up 50, 52 and 53 yields gross social welfare under free competition. Similarly, adding up 51, 52 and 54 yields gross social welfare under price protection. Expression 35 is obtained by subtracting speed investment costs. The single speed market equation is a particular case. *Q.E.D.*

**Proof of Proposition 6.** We proceed in three steps.

Step 1: Finding net social value of speed investment. Endogenous speed choice increases social welfare anytime that $\mathcal{W}_{seg}-\mathcal{W}_{Bertrand}>0$. Under A1 we then require that

$$\left(s_{2}-s_{1}\right)\int_{\hat{\sigma}_{12}^{free}}^{\bar{\sigma}}\sigma e^{\frac{-\sigma}{\nu}}d\sigma-s_{1}\int_{\hat{\sigma}_{1}^{free}}^{\bar{\sigma}}\sigma e^{\frac{-\sigma}{\nu}}d\sigma-2r\nu C\left(s_{2}\right)>0$$

Computing the integrals, we have

$$\left(s_{2}-s_{1}\right)\left(\hat{\sigma}_{12}^{free}+\nu\right)e^{\frac{-\hat{\sigma}_{12}^{free}}{\nu}}-s_{1}\left(\hat{\sigma}_{1}^{free}+\nu\right)e^{\frac{-\sigma}{\nu}}-\nu-2r\nu C\left(s_{2}\right)>0$$

Dividing by $\nu s_1$ and using $x\equiv\frac{\hat{\sigma}_1}{\nu},\Delta\equiv\frac{\hat{\sigma}_{12}-\hat{\sigma}_1}{\nu},k\equiv s_1/\left(s_2-s_1\right)$, we can write the LHS of the above inequality as follows

$$R\left(k\right)\equiv\frac{1}{k}\left(\Delta+x+1\right)e^{-\left(\Delta+x\right)}-\left(x+1\right)e^{-x}-1-2rC\left(s_{2}\right) \tag{55}$$

The net social value of speed investments is positive anytime that $R(k)>0$.

Step 2: $R'(k)<0$

Differentiating 55 we have that $R'(k)<0$ if and only if

$$-e^{-\Delta}\left[\left(\Delta+x\right)\left(1+\Delta'+x'\right)+1\right]+xx'<0$$

Re-arranging

$$(\Delta + x)\left(\Delta' + 1\right) + \left(1 + \Delta x'\right) > xx'\left(e^{\Delta} - 1\right)$$

Differentiating 38-39 we find that

$$x'(k) = \frac{-x\left(e^{\Delta}(1-x)+1\right)}{e^{\Delta}\left(1+\Delta\right)+k\left(1+e^{\Delta}\right)} \tag{56}$$

$$\Delta'(k) = \frac{-x\left(e^{\Delta}-1\right)}{e^{\Delta}\left(1+\Delta\right)+k\left(1+e^{\Delta}\right)} \tag{57}$$

Thus, $x(k)$ and $\Delta(k)$ are decreasing functions. The RHS is thus negative. The sign of the LHS depends on the expressions $(\Delta' + 1)$ and $(1 + \Delta x')$. Using 56 and 57 we have

$$\Delta' + 1 = \frac{e^{\Delta}\left(1+\Delta+k-x\right)+k+x}{e^{\Delta}\left(1+\Delta\right)+k\left(1+e^{\Delta}\right)} > 0$$

and

$$1 + \Delta x' = \frac{e^{\Delta}\left(1+k+\Delta\left(1-x(1-x)\right)\right)+k-\Delta x}{e^{\Delta}\left(1+\Delta\right)+k\left(1+e^{\Delta}\right)} > 0$$

Thus, $R'(k) < 0$

Step 3. Verifying the claim

Note that the inequality is always satisfied when $s_1 \to 0$ $(k \to 0)$ since $R(k) \to \infty$. When $s_1 \to 1$, for any solution $s_2 > s_1$ where the fast venue is active, we know from the proof of proposition 5 that $\lim_{k\to\infty} x(k) = 0$. Then,

$$\lim_{k\to\infty} R(k) = -2 - 2rC\left(s_2\right) < 0$$

Consequently, we found that $\lim_{k\to\infty} R(k) < 0$ and $\lim_{k\to 0} R(k) \to \infty$. Since $R$ is a continuous function, by the intermediate value theorem there is a number $\overline{k} > 0$ such that $R(\overline{k}) = 0$. Since $R$ is monotonically decreasing, $\overline{k}$ is unique. The, for any $s_2 > s_1$, $\overline{s}_1$ is given by $\overline{s}_1 = \frac{s_2\overline{k}}{1+\overline{k}}$. *Q.E.D.*

**Proof of Proposition 7.** Using Lemma 5, the gains of competition are given by

$$\begin{aligned}\mathcal{W}_{comp} - \mathcal{W}_M - \kappa = \frac{s_1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_{12}} \sigma dG\left(\sigma\right) + \frac{s_2}{2r}\int_{\hat{\sigma}_{12}}^{\overline{\sigma}} \sigma dG\left(\sigma\right) - \frac{s_M}{2r}\int_{\hat{\sigma}_M}^{\overline{\sigma}} \sigma dG\left(\sigma\right) \\ - C\left(s_2\right) + C\left(s_M\right) - \kappa\end{aligned}$$

Using A1, computing the integrals, and re-arranging, we have that the gains from competition are equal to zero if and only if

$$\begin{aligned}\frac{s_1}{2r}e^{-\frac{\hat{\sigma}_1}{\nu}}\left(\hat{\sigma}_1+\nu\right) + \frac{\left(s_2-s_1\right)}{2r}e^{-\frac{\hat{\sigma}_{12}}{\nu}}\left(\hat{\sigma}_{12}+\nu\right) - \frac{\nu s_M}{re} = \\ \left(\overline{\kappa} + C\left(s_2\right) - C\left(s_M\right)\right)\end{aligned} \tag{58}$$

which yields $\overline{\kappa}$. Note that using 24 under A1 we have

$$\begin{aligned}
\pi_1^{comp} &= \frac{s_1}{2r}\hat{\sigma}_1\left(e^{-\frac{\hat{\sigma}_1}{\nu}} - e^{-\frac{\hat{\sigma}_{12}}{\nu}}\right)\\
\pi_2^{comp} &= \frac{1}{2r}e^{-\frac{\hat{\sigma}_{12}}{\nu}}\left(\hat{\sigma}_1 s_1 + \hat{\sigma}_{12}\left(s_2 - s_1\right)\right) - C(s_2)\\
\pi_M &= \frac{\nu s_M}{2re} - C\left(s_M\right)
\end{aligned}$$

Re-arranging 58, and using the profit functions above, we obtain the following expression

$$\pi_1^{comp} - \overline{\kappa} = -\pi_2^{comp} - \frac{\nu}{\hat{\sigma}_1}\pi_1^{comp} + \pi_M - \frac{\nu}{2r}s_2 e^{-\frac{\hat{\sigma}_{12}}{\nu}} \tag{59}$$

We from the proof of 5 that when $\underline{s} \to 1$, $\hat{\sigma}_1 \to 0$ and $\hat{\sigma}_{12} \to \nu\underline{\Delta}$. Thus, $\pi_1^{comp}$ and $\pi_2^{comp}$ converge to zero and we have from 59 that $\overline{\kappa}$ approaches $\frac{\nu}{2r}(1 - 1/e)$. *Q.E.D.*

**Proof of Proposition 8.** In general, its objective function is

$$\max_{s_2,q_1,q_2} \frac{\underline{s}}{2r}\int_{\sigma_1}^{\sigma_{12}} \sigma dG\left(\sigma\right) + \frac{s_2}{2r}\int_{\sigma_{12}}^{\overline{\sigma}} \sigma dG\left(\sigma\right) - C\left(s_2\right)$$

and the marginal types are given by 21 and 23, so we have

$$\begin{aligned}
q_1 &= s_1\frac{\sigma_1}{2r},\\
q_2 &= \left(s_2 - s_1\right)\frac{\sigma_{12}}{2r} + q_1.
\end{aligned}$$

The break-even constraint is $q_2\left(1 - G\left(\hat{\sigma}_{12}\right)\right) \geq C\left(s_2\right)$, so the Lagrangian (scaled by $2r$) is

$$\mathcal{L} = \underline{s}\int_{\sigma_1}^{\overline{\sigma}} \sigma dG\left(\sigma\right) + \left(s - \underline{s}\right)\int_{\sigma_{12}}^{\overline{\sigma}} \sigma dG\left(\sigma\right) - 2rC\left(s\right) + \lambda\left\{\left(\left(s - \underline{s}\right)\sigma_{12} + \underline{s}\sigma_1\right)\left(1 - G\left(\sigma_{12}\right)\right) - 2rC\left(s\right)\right\}$$

and the FOCs for affiliations are

$$\begin{aligned}
\sigma_1^* g\left(\sigma_1^*\right) &= \lambda\left(1 - G\left(\sigma_{12}^*\right)\right),\\
\sigma_{12}^* g\left(\sigma_{12}^*\right) &= \frac{\lambda}{1+\lambda}\left(1 - G\left(\sigma_{12}^*\right) - \frac{\underline{s}}{s - \underline{s}}g\left(\sigma_{12}^*\right)\sigma_1^*\right).
\end{aligned}$$

Optimal speed satisfies

$$2r\frac{\partial C}{\partial \rho}\left(s^*\right) = \frac{1}{1+\lambda}\int_{\sigma_{12}^*}^{\overline{\sigma}} \sigma dG\left(\sigma\right) + \frac{\lambda}{1+\lambda}\left(1 - G\left(\sigma_{12}^*\right)\right)\sigma_{12}^*,$$

and the break-even constraint is simply $2rC\left(s^*\right) = \left(1 - G\left(\sigma_{12}\right)\right)\left(\left(s - \underline{s}\right)\sigma_{12}^* + \underline{s}\sigma_1^*\right)$. From the first two FOCs it is immediate that $\sigma_1^* g\left(\sigma_1^*\right) > \sigma_{12}^* g\left(\sigma_{12}^*\right)$. From the second-order conditions we know that $\sigma g\left(\sigma\right)$ is increasing in $\sigma$ (at the optimum values). Therefore $\sigma_1^* > \sigma_{12}^*$, which is inconsistent with our assumption that market 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 $(G(\hat{\sigma}_{12}) - G(\hat{\sigma}_1))\, q_1 + (1 - G(\hat{\sigma}_{12}))\, q_2 > C(s_2)$ and the Lagrangian is

$$\mathcal{L} = \underline{s}\int_{\sigma_1}^{\overline{\sigma}} \sigma dG(\sigma) + (s - \underline{s})\int_{\sigma_{12}}^{\overline{\sigma}} \sigma dG(\sigma) - 2rC(s) + \lambda\left((1 - G(\sigma_1))\,\underline{s}\sigma_1 + (1 - G(\sigma_{12}))\,(s - \underline{s})\,\sigma_{12} - 2rC(s)\right)$$

and the FOCs for affiliations are

$$1 - G(\sigma_1^*) = g(\sigma_1^*)\,\frac{1+\lambda}{\lambda}\sigma_1^*$$
$$1 - G(\sigma_{12}^*) = g(\sigma_2^*)\,\frac{1+\lambda}{\lambda}\sigma_{12}^*$$

Optimal speed satisfies the same equation as before. In this case, we see that $\sigma_1^* = \sigma_{12}^*$, market 1 is still inactive. *Q.E.D.*

**Proof of Proposition 9.** With one venue, the Lagrangian is

$$\mathcal{L} = s\int_{\hat{\sigma}}^{\overline{\sigma}} \sigma dG(\sigma) - 2rC(s) + \lambda\left(s\hat{\sigma}\,(1 - G(\hat{\sigma})) - 2rC(s)\right)$$

From the previous section, it is immediate that

$$1 - G(\sigma^*) = g(\sigma^*)\,\frac{1+\lambda}{\lambda}\sigma_1^*$$

Since the monopoly solution is $\frac{1-G(\sigma_M)}{g(\sigma_M)} = \sigma_M$, it is clear that $\sigma_M > \sigma^*$.

Regarding speed, the planner chooses

$$2r\frac{\partial C}{\partial \rho}(s^*) = \frac{1}{1+\lambda}\int_{\sigma^*}^{\overline{\sigma}} \sigma dG(\sigma) + \frac{\lambda}{1+\lambda}(1 - G(\sigma^*))\,\sigma^*,$$

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 $s^* > s_M$, as expected. However, when the break-even constraint binds, the comparison is ambiguous. We now provide a simple example to show that it is indeed possible for the monopoly to over-invest in speed. *Q.E.D.*

**Proof of Lemma 6.** In a frictionless competitive market we have maximum investor participation. Thus, the marginal type is given by

$$\frac{G(\hat{\sigma}_w)}{1 - G(\hat{\sigma}_w)} = \frac{1}{2\overline{a}} - 1$$

Using A1 we obtain $\hat{\sigma}_w = -\nu\log(2\overline{a})$, which combined with $s_w = 1$ and 29 yields $p_w = \frac{1}{r}\left[\mu - \nu\log(2\overline{a})\right]$. With $s_w = 1$ the instantaneous transaction rate becomes

$$\tau_w = \frac{\gamma}{4}(1 - G(\hat{\sigma}_w)) = \frac{\gamma}{4}\left(e^{-\frac{\hat{\sigma}_w}{\nu}}\right) = \frac{\gamma\overline{a}}{2}$$

By Lemma 5, social welfare is given by

$$\begin{aligned}\mathcal{W}_w &= \frac{\overline{a}}{r}\int_{\hat{\sigma}_w}^{\overline{\sigma}} \sigma dG\left(\sigma\right) \\ &= \frac{\overline{a}}{r\nu}\int_{\hat{\sigma}_w}^{\infty} \sigma e^{-\frac{\hat{\sigma}_w}{\nu}} d\sigma = \frac{\overline{a}}{r}\nu\left(1-\log\left(2\overline{a}\right)\right)\end{aligned}$$

Note that when $\overline{a}=1/2$, Walrasian social welfare is simply given by $\frac{\nu}{2r}$. *Q.E.D.*

# Appendix to Section 2: Trading Fees

In this section we derive the equilibrium when exchanges charge a trading fee $\phi$ per unit of trading. We consider two types of timing assumptions: fees at execution and fees at initiation.

## Execution Fees

The fee is paid when the trade is executed. A seller effectively receives only $p-\phi$ while a buyer effectively pays $p+\phi$. For ease of exposition, we highlight in red the trading fee.

### Bellman Equations

Consider the steady state value functions for any type $\sigma > \hat{\sigma}$ (of course the marginal type $\hat{\sigma}$ is affected by the trading fee, as explained below). As before, define the value of ownership as $I_\sigma^\epsilon \equiv V_{\sigma,\epsilon}\left(1\right)-V_{\sigma,\epsilon}\left(0\right)$. We have

$$\begin{aligned} rV_{\sigma,+}\left(0\right) &= \frac{\gamma}{2}\left[V_{\sigma,-}\left(0\right)-V_{\sigma,+}\left(0\right)\right]+\rho\left(I_\sigma^+-p-\phi\right) \\ rV_{\sigma,-}\left(1\right) &= \mu-\sigma+\frac{\gamma}{2}\left[V_{\sigma,+}\left(1\right)-V_{\sigma,-}\left(1\right)\right]+\rho\left(p-\phi-I_\sigma^-\right)\end{aligned}$$

The equations for $V_{\sigma,-}\left(0\right)$ and $V_{\sigma,+}\left(1\right)$ are unchanged since these types do not trade. Therefore

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

The equilibrium gains from trade for type $\sigma$ in market $\rho$ become:

$$I_\sigma^+-I_\sigma^- = 2\frac{\sigma+\rho\phi}{r+\gamma+\rho}. \tag{60}$$

Then we can solve

$$\begin{aligned} I_{\sigma}^{-} &= \frac{\mu+\rho p}{r+\rho} - \frac{\sigma+\rho\phi}{r+\gamma+\rho} \\ I_{\sigma}^{+} &= \frac{\mu+\rho p}{r+\rho} + \frac{\sigma+\rho\phi}{r+\gamma+\rho} \end{aligned}$$

and the average values (across $\epsilon$) is

$$\begin{aligned} \bar{V}_{\sigma}(0) &= \frac{\rho}{2r}\left(I_{\sigma}^{+} - p - \phi\right) \\ \bar{V}_{\sigma}(1) &= \frac{\mu}{r} + \frac{\rho}{2r}\left(p - \phi - I_{\sigma}^{-}\right) \end{aligned} \tag{61}$$

### Marginal Type

We define the marginal type $\hat{\sigma}$ as the type who is indifferent between buying and not buying when $\epsilon = +1$. The key Bellman equation is that of $V_{\sigma,+}(0)$. The marginal type is then defined by $I_{\hat{\sigma}}^{+} = p + \phi$, therefore:

$$\frac{\rho}{r+\rho}(rp-\mu) = s\left(\hat{\sigma} - (r+\gamma)\phi\right)$$

Let us now compute the ex ante value functions. Let us first consider types $\sigma < \hat{\sigma}$. As before, they join the market to sell but the key difference with the case of no transaction cost is that they only sell when their type is low. When their type is high they strictly prefer to wait. That did not happen without trading fees, since in that case $I_{\hat{\sigma}}^{+} = p$ implied $V_{\hat{\sigma},+,1} - V_{\hat{\sigma},+,0} = p$ so type $(\hat{\sigma},+)$ was indifferent to buying starting from $a = 0$ *and* to selling starting from $a = 1$. With trading fees, we have $I_{\hat{\sigma}}^{+} = p + \phi$. so type type $(\hat{\sigma},+)$ is indifferent to buying, but starkly prefers to keep the asset instead of selling it at price $p - \phi$.

The ex ante value function $\hat{W}$ solves

$$\hat{W} = \bar{a}\bar{V}_{\hat{\sigma}}(1) + (1-\bar{a})\bar{V}_{\hat{\sigma}}(0) \tag{62}$$

which, since $\bar{V}_{\hat{\sigma}}(0) = 0$, leads to

$$\hat{W}(\phi) = \frac{\mu\bar{a}}{r} + \frac{\bar{a}s}{r}\left(\hat{\sigma} - (r+\gamma)\phi\right)$$

### Ex ante Value Functions for Steady State Traders

Let us now consider the steady state types, $\sigma > \hat{\sigma}$. We take the probabilistic allocation interpretation:

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

Using the Bellman equations, we get

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

Therefore we have

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

### Exchange's Profits

Consider now the profit maximization problem of an exchange. With one speed, the indifference condition for the marginal type $\hat{W} - W_{out} = q$ implies

$$\frac{\bar{a}s}{r}\left(\hat{\sigma} - (r+\gamma)\phi\right) = q$$

Note immediately that, using the marginal type, this implies

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

The price depends only on $q$, not on $\phi$. In particular, when $q = 0$, we have $p = \mu/r$ irrespective of the speed of the market. Market clearing still requires $\delta = \left(\frac{1}{2\bar{a}} - 1\right)(1 - G(\hat{\sigma}))$. Total profits for the exchange are

$$\pi^{TOT} = q\frac{1 - G(\hat{\sigma})}{2\bar{a}} + \pi^\phi$$

where $\pi^\phi$ denote the value of trading fees. These fees come from temporary and permanent traders. There are $\delta$ investors who only trade when their type is low. Let $\pi^\epsilon$ be the value of trading fees from type $\epsilon$. We have

$$\begin{aligned} r\pi^+ &= \frac{\gamma}{2}\left(\pi^- - \pi^+\right) \\ r\pi^- &= \frac{\gamma}{2}\left(\pi^+ - \pi^-\right) + \rho\left(\phi - \pi^-\right) \end{aligned}$$

Therefore $\pi^+ = \frac{\rho}{r}\left(\phi - \pi^-\right) - \pi^-$, and $\left(r + \gamma + \left(1 + \frac{\gamma}{2r}\right)\rho\right)\pi^- = \left(1 + \frac{\gamma}{2r}\right)\rho\phi$, and $\pi^+ + \pi^- = \frac{\rho\phi}{r}\frac{r+\gamma}{r+\gamma+\left(1+\frac{\gamma}{2r}\right)\rho}$. The NPV is $\pi_\delta^\phi = \frac{\delta\bar{a}}{2}\left(\pi^+ + \pi^-\right)$ therefore

$$\pi_\delta^\phi = \frac{\delta\bar{a}\rho\phi}{2r + \frac{2r+\gamma}{r+\gamma}\rho}$$

Note that if $\gamma \to \infty$, we have $\pi_{\delta}^{\phi} = \frac{\delta \bar{a} \rho \phi}{2r+\rho}$.

For the permanent investors, we can see the value of the fees from $W$ as $(1-G)\,\frac{\rho\phi}{2r}\left(1 - \frac{\rho}{r+\gamma+\rho}\right)$. Therefore the NPV of trading fees is

$$\pi^{\phi} = (1-G)\,\frac{\rho\phi}{2r}\left(1 - \frac{\rho}{r+\gamma+\rho} + \frac{\frac{1}{2} - \bar{a}}{1 + \frac{2r+\gamma}{r+\gamma}\frac{\rho}{2r}}\right)$$

and total profits are

$$\pi^{TOT} = (1 - G(\hat{\sigma}))\left(\frac{q}{2\bar{a}} + \frac{\rho\phi}{2r}\left(1 - \frac{\rho}{r+\gamma+\rho} + \frac{\frac{1}{2} - \bar{a}}{1 + \frac{2r+\gamma}{r+\gamma}\frac{\rho}{2r}}\right)\right)$$

Since

$$s\hat{\sigma} = \frac{rq}{\bar{a}} + \frac{r+\gamma}{r+\gamma+\rho}\rho\phi$$

we have

$$\begin{aligned}\frac{2r\pi^{TOT}}{1 - G(\hat{\sigma})} &= \frac{rq}{\bar{a}} + \rho\phi\left(1 - \frac{\rho}{r+\gamma+\rho} + \frac{\frac{1}{2} - \bar{a}}{1 + \frac{2r+\gamma}{r+\gamma}\frac{\rho}{2r}}\right) \\ &= s\hat{\sigma} + \frac{\frac{1}{2} - \bar{a}}{1 + \frac{2r+\gamma}{r+\gamma}\frac{\rho}{2r}}\rho\phi\end{aligned}$$

For a given $\hat{\sigma}$, profits are increasing in trading fees as long as there are infra-marginal investors, i.e., as long as $\bar{a} < 0.5$.

### Monopolist

Consider the case $q = 0$. Then we have

$$\hat{\sigma} = (r+\gamma)\,\phi$$

and

$$\pi^{TOT} = (1 - G(\hat{\sigma}))\,\hat{\sigma}\chi$$

where we define $\chi \equiv \frac{\rho}{2r}\left(\frac{1}{r+\gamma+\rho} + \frac{\frac{1}{2} - \bar{a}}{r+\gamma+\rho+\gamma\frac{\rho}{2r}}\right)$. Since $\hat{\sigma} = \arg\max \pi^{TOT}$, we have

**Lemma 7.** *The monopolist chooses exactly the same value for $\hat{\sigma}$ with trading fees or with membership fees.*

**Competition in fees**

Consider a duopoly competing in fees. The value of joining market 1 is

$$W\left(\sigma, s_1, \phi_1\right) = \frac{\mu \bar{a}}{r} + \frac{\bar{a} s_1}{r}\left(\hat{\sigma}_1 - \left(r+\gamma\right)\phi_1\right) + \frac{s_1}{2r}\left(\sigma - \hat{\sigma}_1\right)$$

Let $\hat{\sigma}_1$ be the marginal type who is indifferent between joining market 1 and staying out:

$$\phi_1 = \frac{\hat{\sigma}_1}{r+\gamma}$$

Hence the value is simply

$$W\left(\sigma, s_1, \phi_1\right) = \frac{\mu \bar{a}}{r} + \frac{s_1}{2r}\left(\sigma - \hat{\sigma}_1\right)$$

For market 2, we must also have

$$\phi_2 = \frac{\hat{\sigma}_2}{r+\gamma}$$

Notice that $\hat{\sigma}_2$ does not in fact join market 2. Rather, $\hat{\sigma}_2$ joins market 1. With two markets, we must define a new marginal type, $\hat{\sigma}_{12}$, who is indifferent between joining market 1 and market 2. Therefore $s_1\left(\hat{\sigma}_{12} - \hat{\sigma}_1\right) = s_2\left(\hat{\sigma}_{12} - \hat{\sigma}_2\right)$ or

$$\hat{\sigma}_{12} = \frac{s_2 \hat{\sigma}_2 - s_1 \hat{\sigma}_1}{s_2 - s_1}$$

Note that $\hat{\sigma}_1 < \hat{\sigma}_2 < \hat{\sigma}_{12}$.

Market clearing in market 2 requires $\left(1 - G\left(\hat{\sigma}_{12}\right) + \delta_2\right)\bar{a} = \frac{1-G\left(\hat{\sigma}_{12}\right)}{2}$ or

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

Total profits for the fast exchange under free segmentation are

$$\begin{aligned}
\pi_2^{seg} &= \left(1 - G\left(\hat{\sigma}_{12}\right)\right)\frac{\rho_2 \phi_2}{2r}\frac{r+\gamma}{r+\gamma+\rho_2} + \frac{\delta_2 \bar{a} \rho_2 \phi_2}{2r + \frac{2r+\gamma}{r+\gamma}\rho_2} \\
&= \left(1 - G\left(\hat{\sigma}_{12}\right)\right)\phi_2\left(r+\gamma\right)\frac{\rho_2}{2r}\left(\frac{1}{r+\gamma+\rho_2} + \frac{\frac{1}{2} - \bar{a}}{1 + \frac{2r+\gamma}{r+\gamma}\frac{\rho_2}{2r}}\right)
\end{aligned}$$

hence

$$\pi_2^{seg} = \chi_2\left(1 - G\left(\hat{\sigma}_{12}\right)\right)\hat{\sigma}_2$$

where

$$\chi_2 \equiv \frac{\rho_2}{2r}\left(\frac{1}{r+\gamma+\rho_2}+\frac{\frac{1}{2}-\bar{a}}{1+\frac{2r+\gamma}{r+\gamma}\frac{\rho_2}{2r}}\right)$$

Market clearing for the slow exchange requires $\left(G\left(\hat{\sigma}_{12}\right)-G\left(\hat{\sigma}_1\right)+\delta_1\right)\bar{a}=\frac{G(\hat{\sigma}_{12})-G(\hat{\sigma}_1)}{2}$. Total profits for the slow exchange are

$$\pi_1^{seg}=\chi_1\left(G\left(\hat{\sigma}_{12}\right)-G\left(\hat{\sigma}_1\right)\right)\hat{\sigma}_1$$

The affiliation of investors to markets 1 and 2 are given by the marginal types

$$\begin{aligned}\max_{\hat{\sigma}_1}\pi_1^{seg} &= \chi_1\left(G\left(\hat{\sigma}_{12}\right)-G\left(\hat{\sigma}_1\right)\right)\hat{\sigma}_1\\ \max_{\hat{\sigma}_2}\pi_2^{seg} &= \chi_2\left(1-G\left(\hat{\sigma}_{12}\right)\right)\hat{\sigma}_2\end{aligned}$$

subject to

$$\hat{\sigma}_{12}=\frac{s_2\hat{\sigma}_2-s_1\hat{\sigma}_1}{s_2-s_1}$$

Since $\frac{d(G(\hat{\sigma}_{12})-G(\hat{\sigma}_1))\hat{\sigma}_1}{d\hat{\sigma}_1}=G\left(\hat{\sigma}_{12}\right)-G\left(\hat{\sigma}_1\right)+\hat{\sigma}_1\left(-\frac{s_1}{s_2-s_1}g\left(\hat{\sigma}_{12}\right)-g\left(\hat{\sigma}_1\right)\right)$ we have

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

Similarly, we have $1-G\left(\hat{\sigma}_{12}\right)-g\left(\hat{\sigma}_{12}\right)\hat{\sigma}_2\frac{s_2}{s_2-s_1}=0$ which we can write as

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

These are the same FOCs as in Lemma 2.

## Initiation Fees

Anyone initiating an order pays $\phi$.

### Bellman Equations

Consider the steady state value functions for any type $\sigma>\hat{\sigma}$. The equations for $V_{\sigma,-}(0)$ and $V_{\sigma,+}(1)$ are

$$\begin{aligned}rV_{\sigma,-}(0) &= \frac{\gamma}{2}\left[V_{\sigma,+}(0)-V_{\sigma,-}(0)-\phi\right]\\ rV_{\sigma,+}(1) &= \mu+\sigma+\frac{\gamma}{2}\left[V_{\sigma,-}(1)-V_{\sigma,+}(1)-\phi\right]\end{aligned}$$

Note that these assume that each time the type changes, a fee is paid. As before, define the value of ownership as $I_\sigma^\epsilon \equiv V_{\sigma,\epsilon}(1) - V_{\sigma,\epsilon}(0)$. We have

$$\begin{aligned} rV_{\sigma,+}(0) &= \frac{\gamma}{2}\left[V_{\sigma,-}(0) - V_{\sigma,+}(0)\right] + \rho\left(I_\sigma^+ - p\right) \\ rV_{\sigma,-}(1) &= \mu - \sigma + \frac{\gamma}{2}\left[V_{\sigma,+}(1) - V_{\sigma,-}(1)\right] + \rho\left(p - I_\sigma^-\right) \end{aligned}$$

Therefore

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

The equilibrium gains from trade for type $\sigma$ in market $\rho$ become:

$$I_\sigma^+ - I_\sigma^- = \frac{2\sigma - \gamma\phi}{r + \gamma + \rho}.$$

Then we can solve

$$\begin{aligned} I_\sigma^- &= \frac{\mu + \rho p}{r + \rho} - \frac{\sigma}{r + \gamma + \rho} + \frac{\gamma}{2}\frac{\phi}{r + \gamma + \rho} \\ I_\sigma^+ &= \frac{\mu + \rho p}{r + \rho} + \frac{\sigma}{r + \gamma + \rho} - \frac{\gamma}{2}\frac{\phi}{r + \gamma + \rho} \end{aligned}$$

and the average values (across $\epsilon$) is

$$\begin{aligned} \bar{V}_\sigma(0) &= \frac{\rho}{2r}\left(I_\sigma^+ - p\right) - \frac{\gamma}{4r}\phi \\ \bar{V}_\sigma(1) &= \frac{\mu}{r} + \frac{\rho}{2r}\left(p - I_\sigma^-\right) - \frac{\gamma}{4r}\phi \end{aligned}$$

### Marginal Type

We define the marginal type $\hat{\sigma}$ as the type who is indifferent between submitting an order and not submitting when $\epsilon$ switches from $-1$ to $+1$. The key Bellman equation is that of

$$V_{\sigma,+}(0) - V_{\sigma,-}(0) = \phi$$

Since

$$(r + \gamma)\left(V_{\sigma,+}(0) - V_{\sigma,-}(0)\right) = \rho\left(I_\sigma^+ - p\right) + \frac{\gamma}{2}\phi$$

we get

$$\rho\left(I_{\hat{\sigma}}^+ - p\right) = \left(r + \frac{\gamma}{2}\right)\phi$$

or

$$\rho\left(\frac{\mu - rp}{r+\rho} + \frac{\hat{\sigma}}{r+\gamma+\rho}\right) = \left(r + \frac{\gamma}{2}\frac{r+\gamma+2\rho}{r+\gamma+\rho}\right)\phi$$

or

$$\rho\left(p - I_{\sigma}^{-}\right) = \rho\frac{2\hat{\sigma} - \gamma\phi}{r+\gamma+\rho} - \left(r + \frac{\gamma}{2}\right)\phi = \rho\frac{2\hat{\sigma} - \left(\gamma + \frac{r}{\rho} + \frac{\gamma}{2\rho}\right)\phi}{r+\gamma+\rho}$$

Let us now compute the ex ante value functions. Let us first consider types $\sigma < \hat{\sigma}$. As before, they join the market to sell but the key difference with the case of no transaction cost is that they only sell when their type is low. When their type is high they strictly prefer to wait.

The ex ante value function $\hat{W}$ solves

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

which, since $\bar{V}_{\hat{\sigma}}(0) = 0$, leads to

$$\hat{W}(\phi) = \frac{\mu\bar{a}}{r} + \frac{\rho\bar{a}}{2r}\left(p - I_{\sigma}^{-}\right) - \frac{\gamma\bar{a}}{4r}\phi = \frac{\mu\bar{a}}{r} + \frac{\bar{a}}{2r}\left(\frac{2\rho\hat{\sigma} - \left(\gamma\rho + r + \frac{\gamma}{2}\right)\phi}{r+\gamma+\rho} - \frac{\gamma}{2}\phi\right)$$

This equation is of the usual type:

$$\hat{W}(\phi) = \frac{\mu\bar{a}}{r} + \frac{\bar{a}s}{r}\left(\hat{\sigma} - \Gamma\phi\right)$$

for some (complicated in this case) $\Gamma$ that depends on $r$, $\gamma$ and $\rho$.

### Ex ante Value Functions for Steady State Traders

As usual, we have

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

### Monopolist

Once again, the profits are of the type

$$\pi^{TOT} = \left(1 - G(\hat{\sigma})\right)\hat{\sigma}\chi$$

for some constant $\chi$ that depends on $r$, $\gamma$ and $\rho$.

Therefore *the monopolist chooses exactly the same value of $\hat{\sigma}$ with trading fees of with membership fees.*

### Competition in fees

The calculations are the same as before, with slightly different constant $\chi_1$ and $\chi_2$. So the equilibrium allocations are the same: $\hat{\sigma}_1$, $\hat{\sigma}_2$ and $\hat{\sigma}_{1,2}$ are the same.

# Appendix to Section 2: Multi-market Traders

Let us now discuss the possibility that some traders might choose to pay both membership fees and trade in both markets. To analyze this case, we need first to characterize the optimal trading strategies in case a trader actually can trade in two markets. Let market 1 be the slow market, with the low price $p_1 < p_2$, let us call the types that trade in both markets the multi-market traders (MMTs). We consider the case $\bar{a} = 1/2$ for simplicity.

### Bellman equations

Suppose MMTs always send orders to both markets, and always trade when they get the chance. This happens if and only if $I_{\sigma,+}^{MM} > p_2$ and $p_1 > I_{\sigma,-}^{MM}$. In this case, the value functions are

$$\begin{aligned} rV_{\sigma,+}^{MM}(0) &= \frac{\gamma}{2}\left[V_{\sigma,-}^{MM}(0) - V_{\sigma,+}^{MM}(0)\right] + \rho_1\left(I_{\sigma,+}^{MM} - p_1\right) + \rho_2\left(I_{\sigma,+}^{MM} - p_2\right) \\ rV_{\sigma,-}^{MM}(1) &= \mu - \sigma + \frac{\gamma}{2}\left[V_{\sigma,+}^{MM}(1) - V_{\sigma,-}^{MM}(1)\right] + \rho_1\left(p_1 - I_{\sigma,-}^{MM}\right) + \rho_2\left(p_2 - I_{\sigma,-}^{MM}\right) \end{aligned}$$

and

$$\begin{aligned} rV_{\sigma,-}^{MM}(0) &= \frac{\gamma}{2}\left[V_{\sigma,+}^{MM}(0) - V_{\sigma,-}^{MM}(0)\right] \\ rV_{\sigma,+}^{MM}(1) &= \mu + \sigma + \frac{\gamma}{2}\left[V_{\sigma,-}^{MM}(1) - V_{\sigma,+}^{MM}(1)\right] \end{aligned}$$

The key issue is whether MMTs send both buy and sell orders in both markets. This happens if and only if $I_{\sigma,+}^{MM} > p_2$ and $p_1 > I_{\sigma,-}^{MM}$, where the values of ownership are defined as before, and the Bellman equations for MMT are equivalent to one market with an average price $p^T = \frac{\rho_1 p_1 + \rho_2 p_2}{\rho_1 + \rho_2}$ and a total speed $\rho^\intercal = \rho_1 + \rho_2$. In particular the gains from trade are given by

$$I_{\sigma,+}^{MM} - I_{\sigma,-}^{MM} = \frac{2\sigma}{r + \gamma + \rho_1 + \rho_2}.$$

There are two important points to understand. First, when MMTs always trade in both markets, the equilibrium is the same as without MMTs because the MMTs submit the same numbers of buys and sells in both markets. Asset allocations across markets do not change, $p_1$ and $p_2$ remain the same.

The second key point is that we must check that MMTs actually want to buy at the high price and sell at the low price, rather than wait for a better deal. In other words, we must check that $I^{MM}_{\sigma,+} > p_2$ and $p_1 > I^{MM}_{\sigma,-}$. These conditions are equivalent to $\sigma > \sigma^{MM}_{buy}$ and $\sigma > \sigma^{MM}_{sell}$, where we define two marginal types

$$\frac{\sigma^{MM}_{buy}}{r+\gamma+\rho_1+\rho_2} = p_2 - \frac{\mu+\rho_1 p_1+\rho_2 p_2}{r+\rho_1+\rho_2}$$

and

$$\frac{\sigma^{MM}_{sell}}{r+\gamma+\rho_1+\rho_2} = \frac{\mu+\rho_1 p_1+\rho_2 p_2}{r+\rho_1+\rho_2} - p_1$$

Note in particular that we immediately obtain an upper bound for price dispersion:

$$p_2 - p_1 < I^{MM}_{\sigma,+} - I^{MM}_{\sigma,-} = \frac{2\sigma^{MM}}{r+\gamma+\rho_1+\rho_2}$$

This implies the following Lemma.

**Lemma 8.** *The price difference cannot be higher than the gains from trade of the lowest MMTs.*

Finally, we can solve for the marginal MMT, i.e. the type who is just indifferent between trading only in market 2 and trading in both markets. For this type, we must have

$$\bar{W}_{MM}\left(\hat{\sigma}_{MM}\right) - W_2\left(\hat{\sigma}_{MM}\right) = q_1$$

which, using the equilibrium conditions, we can write as

$$\hat{\sigma}_{MM} \equiv \frac{\frac{\rho_1}{r+\rho_1}}{\frac{\rho_1+\rho_2}{r+\gamma+\rho_1+\rho_2} - \frac{\rho_2}{r+\gamma+\rho_2}}\left(rp_1 - \mu\right)$$

By definition, all types above $\hat{\sigma}_{MM}$ would like to become MMTs.

The possibility of MMTs is clearly interesting, especially for its implications on asset prices. For the purpose of our model, however, they do not play a quantitatively important role. To see why, we use our benchmark calibration to see where $\hat{\sigma}_{MM}$, $\sigma^{MM}_{buy}$ and $\sigma^{MM}_{buy}$ are. We find

| | $\hat{\sigma}_1$ | $\hat{\sigma}_2$ | $\hat{\sigma}_{12}$ | $\hat{\sigma}_{MM}$ | $\sigma^{MM}_{buy}$ | $\sigma^{MM}_{sell}$ |
|---|---|---|---|---|---|---|
| Value | 0.49 | 0.96 | 1.81 | 913 | $3.45 \times 10^6$ | $1.36 \times 10^8$ |
| CDF | 0.166 | 0.298 | 0.488 | 1 | 1 | 1 |

Therefore, in our model, the equilibrium does not change if we allow for MM trading.

# Appendix to Section 5: Counter-example

Consider a binary distribution. High $\sigma^H = \bar{\sigma}$ with population share $n$. Low sigma $\sigma^L = \alpha\bar{\sigma}$ with $\alpha < 1$ and population share $1 - n$. Cost function $2rC = \frac{c}{2}s^2$. The marginal price is $q^i = \rho\sigma^i$. The monopoly has two choices:

- Set price to $\rho\alpha\bar{\sigma}$, get everyone to participate, then $\pi = \rho\alpha\bar{\sigma} - c(s)$.
- Set high price $\rho\bar{\sigma}$, only high types participate, then $\pi = \rho n\bar{\sigma} - c(s)$.

The monopoly chooses high speed low participation if and only if $n > \alpha$. The speed choice is $\max(n, \alpha)\bar{\sigma}/c$.

The Planner has two main choices. If all participate $W = \rho\bar{\sigma}((1-n)\alpha + n) - c(s)$. Then it depends on whether the break-even constraint binds. If it does not, then the planner chooses a higher speed than any monopoly: $s^* = \frac{\bar{\sigma}((1-n)\alpha+n)}{c}$. The break-even constraint binds if $s^*\alpha\bar{\sigma} < c(s^*)$, which is equivalent to $cs > 2\alpha\bar{\sigma} \Leftrightarrow (1-n)\alpha + n > 2\alpha \Leftrightarrow \alpha < n(1-\alpha)$. The planner can still choose full participation, but at limit price $c(s) = s\alpha\bar{\sigma} \Leftrightarrow s = \frac{2\alpha\bar{\sigma}}{c}$. Then welfare is $W = s\bar{\sigma}n(1-\alpha) = \frac{2}{c}(\bar{\sigma})^2 n\alpha(1-\alpha)$.

The other choice for the planner is that only high type participate. This is same program as monopoly. Speed choice is $n\bar{\sigma}/c$. Welfare is $\frac{1}{2c}(n\bar{\sigma})^2$. The Planner chooses low speed high participation if and only if $\frac{2}{c}(\bar{\sigma})^2 n\alpha(1-\alpha) > \frac{1}{2c}(n\bar{\sigma})^2$ or $4\alpha(1-\alpha) > n$.

To summarize, for the planner to choose lower speed than monopoly, we need: (i) $n > \alpha$ so monopoly goes for high speed low participation; (ii) $4\alpha(1-\alpha) > n$ so planner chooses high participation; (iii) $\alpha < n(1-\alpha)$ so break-even violated; and (iv) $n\bar{\sigma}/c > \frac{2\alpha\bar{\sigma}}{c} \Leftrightarrow n > 2\alpha$ so monopoly speed indeed higher. It is easy to see that (i) is not binding. So we have the three following conditions

1. $4\alpha(1-\alpha) > n$
2. $\alpha < \frac{n}{1+n}$
3. $n > 2\alpha$

Take $n = 1/4$ then we need $\alpha < 1/8$ for third, second is not binding, and it is easy to find a solution for the first. QED.