Competing on Speed*

Emiliano S. Pagnotta† and Thomas Philippon‡

September 2016

Abstract

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

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

*We are particularly grateful to Joel Hasbrouck for his insights and to Pierre-Oliver Weill, Guillaume Rocheteau, Albert Menkveld, Jonathan Brogaard, and Eric Aldrich (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, the London School of Economics, the Toulouse School of Economics, Rochester University, the Tinbergen Institute, the University of Illinois at Chicago, Imperial College Business School, the University of Lugano, the University of Amsterdam, IESE Business School, the University of San Andrés, the Federal Reserve Bank of New York, the Bank of England, and seminar participant at the following conferences: the 2011 Society of Economic Dynamics, the Fourth Annual Conference on Money, Banking and Asset Markets at University of Wisconsin-Madison, the Third Annual Conference of Advances in Macro Finance Tepper-LAEF, the Western Finance Association (Las Vegas), the Finance Theory Group Meeting (Harvard Business School), the Cowles Foundation GE Conference (Yale University), the Econometric Society NA Meetings (Evanston), the 2012 Financial Management Association Napa Conference, the Fourth Annual Hedge Fund Euronext Conference, the 2013 Conference of the Paul Woolley Centre for the Study of Capital Market Disfunctionality (London School of Economics), and the 2014 CAFIN Conference. We are grateful to market participants at Citibank, Société Générale and the TABB Group for their feedback. We acknowledge the support of the Smith Richardson Foundation.

†Imperial College Business School. Email: e.pagnotta@imperial.ac.uk (corresponding author)

‡New York University Stern School of Business, National Bureau of Economic Research, and the Centre for Economic Policy Research. Email: tphilipp@stern.nyu.edu

1 Introduction

The securities exchange industry has changed deeply over the past decade. Entry of new venues has led to fragmentation of trading, particularly in the United States and in Europe. Trading speed has increased a lot in some markets (equities and standardized derivatives in particular), but much trading still relies on human inputs. As a result, we now observe significant heterogeneity in trading across venues and asset classes. These evolutions have triggered heated debates in academic and policy circles. Why do venues compete on speed? Is there a connection between speed and fragmentation? What are the welfare consequences of these changes? What are the appropriate regulations? To shed light on these issues, we propose a model of the market for markets, i.e., we analyze competition among trading venues offering differentiated trading services. For simplicity, we refer broadly to the quality of these services as *speed*, by which we mean a feature that reduces the time between the occurrence of a desire to trade and the execution of the trade.[1]

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

Given the structure of the demand for speed, we can then analyze the supply of trading services as a sequential game. Venues first decide whether to enter or not (entry game), then invest in trading technologies (speed choices), and finally compete on fees to attract investors (affiliation game). In the equilibrium of the affiliation game, we show that faster venues charge higher fees and attract speed-sensitive investors, and that competition leads to lower fees and greater investor participation (Proposition 2).

We then turn to speed choices and entry. The key point is that choosing different speeds allows the venues to offer vertically differentiated products, as in Shaked and Sutton (1982; 1983). We find that speed choices are generally inefficient because differentiation relaxes price competition

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

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

and because venues do not internalize the welfare gains of infra-marginal investors (Propositions 3 and 4). In this context, we show that a regulator would find it optimal to impose a minimum speed requirement, but not a maximum speed limit (Proposition 5). Finally, in the entry game, we highlight the tension between business stealing, competition and product diversity. As long as the cost of speed is not too high, we show that entry by a second venue always enhances welfare (Proposition 7). However, excess entry is possible in a general oligopoly and we study one such extension.

An important contribution of our paper is the analysis of trade price regulations in a fragmented multi-venue market. We consider two polar cases. In the *segmented* case, a venue only executes the orders of its own investors and trades occur at different prices in different venues. In the *integrated* case, there is a unique price and venues offer different "gates of entry" to a single asset market. This is our stylized way to capture a price protection rule such as that in Reg NMS (see Section 2), and we thus label this case *price protection*.[3] These regulations affect the gains from trade, and therefore have an impact on all the stages of the model: affiliation, speed, and entry. In the affiliation game, protection acts as a subsidy to the slow venue because its investors enjoy interacting with investors from the fast venue who are eager to trade. The slow venue therefore charges higher fees and enjoys higher profits under protection. This then encourages entry and fragmentation, and it may lead to greater investment in speed. We find that the welfare consequences of price protection depend crucially on its impact on entry decisions. For a range of intermediate entry costs only one venue can enter profitably in the segmented market equilibrium. In such cases, the implicit subsidy embedded in price protection can allow entry by a slower venue, stimulate competition, and result in higher participation and welfare (Proposition 6).[4] To the best of our knowledge this is the first formal analysis of this issue. The predictions of the model also seem broadly consistent with the recent U.S. experience. After the implementation of Reg NMS, new market centers proliferated and trading speed increased rapidly.

To summarize, our theoretical results rely on three key assumptions: (i) agents anticipate random trading needs; (ii) venues make costly investments to allow for easier/faster trading; and (iii) these investments allow exchanges to cater to different investors by offering different speeds/convenience levels.[5] In the last section of the paper, we provide a comprehensive calibration of the model for

[3] When we conduct policy experiments, we consider the case in which integration is mandated by regulation using a price protection rule. For simplicity, we do not consider intermediate cases with imperfect arbitrage between markets.

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

[5] Our analysis applies to various asset classes and investors. At the slow end of the spectrum, retail investors have access to several types of brokerage accounts. Brokers invest in information technology to allow investors to trade more easily. Brokers compete in fees as well as in the speed and the convenience of their trading systems. Some of the costs are fixed (as in opening an account) and some are on a per-trade basis. Our benchmark model emphasizes fixed costs. We analyze trading costs in an extension of the model. Speed can also refer to the frequency at which price quotes are refreshed. For U.S. equities, anyone with access to the internet can obtain free quotes with a few minutes' delay. One must pay for a subscription to receive faster updates. Biais, Hombert, and Weill (2014) propose an interpretation in which traders have continuous access to the market but are uncertain about the preferences of their institutions. In that case, speed is related to the flow of information before trades happen. All

three asset classes associated with different trading speeds: corporate bonds, individual equities, and equity index futures. Our main benchmark for welfare comparisons is a planner subject to the same technological constraints as our private venues. The calibration allows us to analyze the welfare consequences of various regulations. We find that the welfare consequences of price protection depend mostly on its impact on entry. Although lower technological costs can dramatically increase trading speed and volume, the associated welfare gains are small. However, welfare gains from enforcing a minimum speed can be significant. We can decompose welfare losses from lack of investor participation, misallocation of investors among venues, and misallocation of assets among investors. Interestingly, we find that, because venues differentiate, welfare losses can be significant even when the fast venues trade at extreme speeds. As a result, a secular reduction in speed costs does not make the economy converge to the frictionless outcome. Finally, we present several extensions of the main framework in various appendices: trading fees instead of participation fees, multi-venue investor affiliation, an arbitrary number of venues, as well as alternative characterizations of the (constrained) first best.

Relation to the Literature. Our paper relates to several strands of the literature in economics and finance. Early theoretical analyses of fragmentation include those of Mendelson (1987) and Pagano (1989). These static models focus on the tradeoff between liquidity externalities, market power, and trading costs.[6] This tension was of the first order of importance when different market places were not as integrated as they are nowadays. Venues can differentiate in areas other than speed. For example, Santos and Scheinkman (2001) study competition in margin requirements, and Foucault and Parlour (2004) and Chao et al. (2016) study competition in listing and make-take fees, respectively. These papers consider static frameworks and do not analyze speed differentiation. By contrast, we develop a dynamic model where speed plays an explicit role. We also provide the first equilibrium analysis of price protection.

Our trading model builds on the recent literature that models dynamic trading with friction, spurred by Duffie, Garleanu, and Pedersen (2005), and is closest to that of to that of Lagos and Rocheteau (2009, LR09 hereafter).[7] We follow these models in that investors valuation change randomly. We do not, however, encompass all trading mechanisms. In contrast to Duffie et al. (2005), we do not model decentralized OTC trades through random search, a specific matching function, and a bar-

these interpretations are broadly consistent with the structure of our model. Banks and brokers also offer trading platforms to their institutional clients. These are typically faster and more expensive. Venues can be interpreted as exchanges, trading platforms, or dealer networks. One can also interpret OTC (off-exchange) stock trading as a group of slow venues. This group includes dark pools, internalization pools, OTC dealers, and crossing networks. It currently represents between one-fourth and one-third of the U.S. stock trading volume.

[6] The literature that analyzes fragmented trading contains several additional themes. Biais (1993), Glosten (1994), Hendershott and Mendelson (2000), Parlour and Seppi (2003), and Rust and Hall (2003) study competition between markets with different trading rules. More recently, Colliard and Foucault (2012) study the effect of trading fees in a context where an exchange competes with an OTC dealer. Chowdhry and Nanda (1991), Madhavan (1995) and Baruch, Karolyi, and Lemmon (2007), in turn, analyze information transmission with multiple venues. For a textbook analysis, see Chapter 26 of Harris (2003).

[7] Weill (2007) uses a related framework to analyze market making in exchanges. Vayanos and Wang (2007) and Weill (2008) study the concentration of liquidity across assets instead of venues. Many additional contributions are surveyed by Lagos, Rocheteau, and Wright (2015).

gaining game. We do not model a limit order market as do Biais, Hombert, and Weill (2014). For tractability and cleanness of analysis, we adopt instead a (frictional) Walrasian clearing protocol, as do LR09 and Gârleanu (2009). We seek to capture different trading frictions that affect the search for liquidity in a stylized fashion, by introducing a random delay in the execution of a trade. A distinctive feature of our model is that the distribution of such delays arises endogenously and is explicitly affected by the nature of the competition among venues. To the best of our knowledge, this paper offers the first model of a market for markets, with joint determination of trading and market structure. This allows us to jointly study the welfare consequences of entry, technological investments, participation, and various regulations. We are also the first to propose a complete calibration in such environment. The asset pricing implications are studied by Pagnotta (2015).

Our work complements the recent literature that analyzes HFT (e.g., Ait-Sahalia et al. (2013), Budish, Cramton, and Shim (2016), Foucault et al. (2016), Biais, Foucault, and Moinas (2015)). The literature models the speed-related advantages that some traders have over others by introducing a form of (typically short-lived) asymmetric information. We do not analyze asymmetric information but we provide a "macro" building block where positive and normative issues related to investors with different speed capacities can be analyzed.

In the industrial organization literature, Mussa and Rosen (1978), Gabszewicz and Thisse (1979), and Shaked and Sutton (1982; 1983) pioneered the analysis of vertically differentiated oligopolies. Our framework enriches the classical environment by having agents consuming a differentiated product first (trading services), and a homogeneous product second (the asset itself). Consequently, we can endogenize the value of product 'quality' (trading delays here) through a micro-founded trading game. This approach allows us to study, among other things, how regulations affecting the trade of the downstream product shape the market design.

The remainder of the paper is organized as follows. Section 2 presents key stylized facts about the global trading landscape. Section 3 presents our benchmark trading model and derives the value functions of investors. Section 4 provides a formal definition of the market structure equilibrium and characterizes the regulation problem. Section 5 analyzes competition among trading venues with and without price protection. Section 6 analyzes trading venues' investment in speed. Section 7 analyzes entry. Section 8 develops a quantitative version of the model and discusses the impact of several market interventions. Section 9 concludes the paper.

2 Securities Trading: Motivating Facts and Trends

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

Trading speed. Garbade and Silber (1977) study two early developments of market infrastructure: the telegraph connecting New York with other American market centers in the 1840s, and

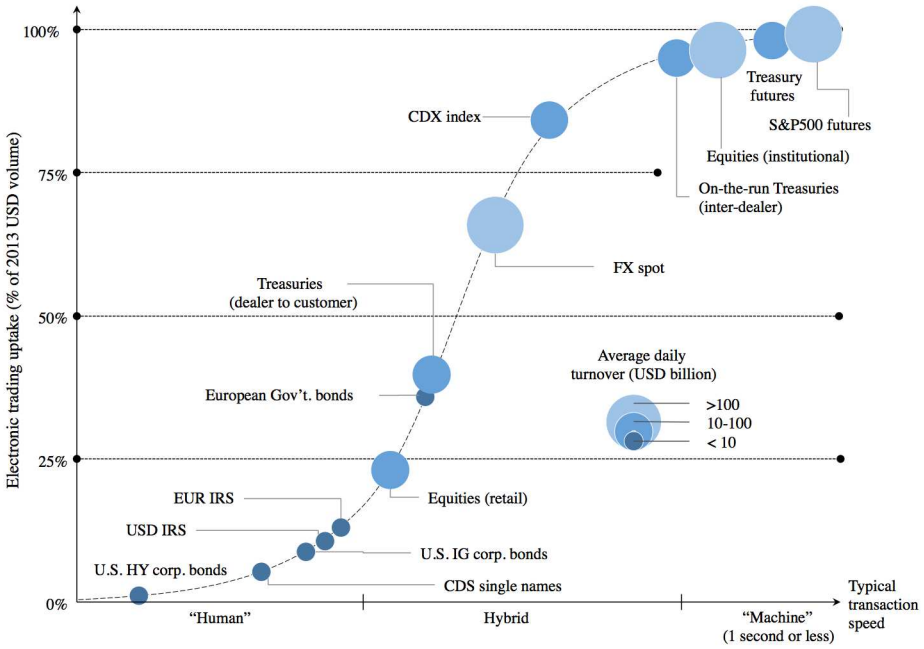

Figure 1. Assets classes and trading speeds. Source: TABB Group and various sources.

a trans-Atlantic cable connecting New York and London in 1866.[8] Early in the 1900s, all European stock markets (except London) conducted periodic auctions, once or several times a day. The progressive – although not simultaneous – adoption of continuous trading represented a massive increase in trading frequencies.[9] The diffusion of personal computers in the 1980s enabled electronic trading and the development of information systems, such Bloomberg terminals. The crash of 1987 and subsequent regulation reforms pushed exchanges to adopt automatic execution systems (like the Small Order Execution System) that did not rely on traditional floor brokers (Lewis (2014)). These historical examples highlight the interactions between technology, competition, and market structure that are at the heart of our model.

Figure 1 summarizes the current trading landscape. Over the past 10 years, market centers have made costly investments in trading infrastructure to reduce order execution and communication latencies (table A1 lists several recent examples). This process has gone beyond equities and futures to reach options, bonds and currencies. We want to emphasize two important stylized facts. First, trading speeds vary greatly across markets and much trading still relies on human input. For instance, electronic trading covered only 21% of the corporate bond market in 2014, while voice trading covered the remaining 79%. High-frequency trading (HFT) is not the norm in most markets. Second, it is important to distinguish the speed of quote updating from the speed of trades.

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

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

TABLE I
VENUE SPEED CHOICES IN ASSET MARKETS: EXAMPLES

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

According to a recent SEC study (see Figure A2), more than half of fully executed orders (and 60% of partial trades) take place between 5 seconds and 10 minutes. The blazing fast speed advertised by many trading venues corresponds to quote revisions, not trades, and it is trading speed that matters in our model.

Fragmentation and differentiation. The second major feature of the current trading landscape is fragmentation (see Figure A1). Traditional markets such as the London Stock Exchange (right panel) have lost market share to faster entrants such as Chi-X. The left panel shows an even more dramatic evolution: The fraction of NYSE-listed stocks traded on the NYSE decreased from 80% in 2004 to just over 20% in 2009. Most of the lost trading volume has been captured by new entrants (e.g., Direct Edge and BATS).[10] Overall, the phenomenon is so important that market participants keep track of *fragmentation indexes* across asset classes and countries.[11]

Besides the secular increase in average trading speed, we observe a lot of *differentiation* in every major asset class. Our model emphasizes the speed-related choices that investor must make and table I presents some examples. In equity markets, investors might need to choose between two exchanges, such as the NASDAQ vs. the NYSE in the U.S., or ASX vs. Chi-X in Australia.[12] A second, broader, interpretation is that investors sort themselves between "exchanges" and a range of "alternative trading venues." According to the SEC classification, U.S. investors can opt to direct their orders to registered exchanges, and a range of Alternative Trading Systems (ATS), that include Electronic Communication Networks (ECN), dark pools, broker–dealer Internalizers, and Crossing Networks.[13] Over-the-counter venues have made technical progress, but, as a group, organized

[10] We focus here on the European and U.S. experiences, but our analysis and results apply to other recent international cases. The links between entry, competition and speed investments is apparent in regions like Asia, Australia and Latin America, where traditional trading venues face the threat of alternative trading platforms. Indeed, Table A1 shows that investment in speed is a global phenomenon.

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

[12] Boehmer (2005) documents the trade-off between execution speed and costs in U.S. markets before Reg NMS. He finds that, analogously to venue 2 in our model, the NASDAQ is more expensive than the NYSE, but it is also faster. More recent data shows that the NASDAQ was still significantly faster than the NYSE at the time of Reg NMS implementation in 2007 (Angel et al., 2011).

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

exchanges typically offer investors the fastest communication and trading responses. To summarize, we can group broker-dealers/crossing networks and floor-driven exchanges as slow venues, and (lit) electronic exchanges as fast venues.[14] We discuss additional asset classes in Appendix A.2.

Regulation of Entry and Price Protection. Market regulators have not been passive witnesses to these evolutions. U.S. policy makers have encouraged fragmentation to reduce the market power of trading venues, prominently with the Regulation of Exchange and Alternative Trading Systems (Reg ATS) and Regulation National Market System (Reg NMS).[15] Encouraged by this experience, other economies started promoting competition between market centers. In Europe, for example, the Markets in Financial Instruments Directive (MiFID) transformed the trading landscape. Large-cap stocks that previously traded in one or two venues are now traded in almost 50 venues, including internalization pools and over-the-counter (OTC) venues.

Concerns about adverse effects of *price fragmentation*, in turn, motivated regulators to promote rules for investor protection. There are essentially two approaches to the regulation of order execution prices. Under the *trade-through model* market centers are connected to one another and they prevent trading through better prices available elsewhere. Price is thus the primary criterion for best execution. This requires complex and costly connections as well as strong monitoring activity by market regulators. In the U.S., Rule 611 of Reg NMS essentially requires that venues execute their trades at the national best bid and offer quotes, thereby consolidating prices from scattered trading.[16] Under the *principles-based model* criteria other than prices can be included in the best execution policy, such as investor type, listing exchange, liquidity, execution probability

Systematic Internalizers.

[14] Yet another interpretation relates to retail investors more directly, at much lower speeds. Consider a given household that seeks exposure to a market risk factor. The default way of getting such exposure is through a mutual fund that can generally be accessed through their default commercial bank. However, mutual funds can only be traded daily. A speed-sensitive investor may instead consider an ETF or an exchange-traded index future. For the latter, the additional cost of speed is represented best by the cost of opening and maintaining a specific brokerage account, the cost of accessing real-time quotes, as well as acquiring the necessary knowledge on how to use a given trading platform. Of course, not all of the related revenues are captured by the trading venues.

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

> Mandating the consolidation of order flow in a single venue would create a monopoly and thereby lose the important benefits of competition among markets. The benefits of such competition include incentives for trading centers to create new products, provide high quality trading services that meet the needs of investors, and keep trading fees low.

[16] The trade through rule is subject to the following features: (i) eligible NMS securities prices are aggregated in the Securities Information Processor (SIP) and then disseminated to market participants. (ii) Prices are quoted gross of trading fees. (iii) The SEC sets an access fee cap of $0.003 per share. (iv) 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 best offer 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. Reg NMS mandates protected prices among organized exchanges, while order prices are not protected, between exchanges and many ATS (especially dark venues that, by definition, do not display prices). For more details See the SEC's Reg NMS documentation. In Canada, the Order Protection Rule (OPR) implemented by the Investment Industry Regulatory Organization of Canada shares the same spirit but aims to protect orders beyond the top level of the order book.

and speed.[17] This approach gives more discretion and less transparency but requires a simpler set of linkages between markets and can promote innovation by not enforcing uniformity.[18] Table A2 lists several international examples of each approach.

In our framework, beyond its effect on execution prices, price protection affects the nature of competition between venues.[19] An additional type of market intervention that we study is the direct regulation of trading speed. We briefly discuss some examples of such policies in Appendix A.3.

3 Trading Model

The structure of our paper is depicted in Figure 2. This section analyzes the trading stage. It provides the explicit micro foundation of how investors value speed in financial markets. We present our trading model and analyze the equilibrium in one venue. The key result of this section is a characterization of value functions as a function of speed and investor characteristics.

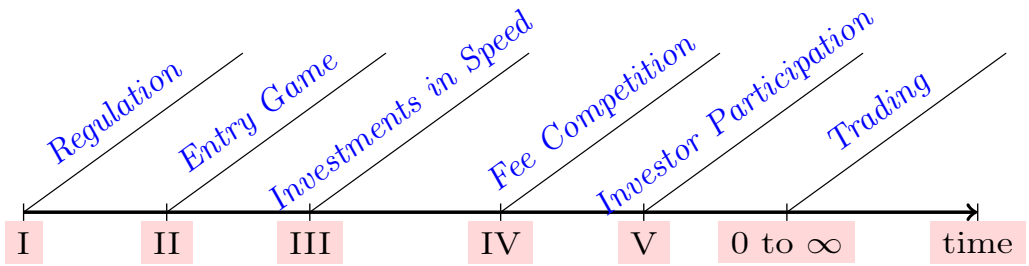

Figure 2. Timing and structure of the model

[17] 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." In Europe, MiFID deregulated price competition in European markets in 2007. Prior to MiFID there was a 'concentration' rule in most European markets. Countries like Spain, Italy or France forced stocks listed in their countries to trade domestically. A "transparency" regime was introduced, but no formal trade-through rule that venues are held responsible for. Both in Europe and Japan, sell-side best execution policies do not mandate to consider or monitor every venue. Monitoring of execution quality is generally left to clients, creating problems when investors have inadequate knowledge of financial markets.

[18] Arbitrageurs and smart routing technologies often work to partially undo price differentials between markets. However, pre MiFID I empirical evidence by Foucault and Menkveld (2008) suggests that this fact does not make price protection rules redundant. These authors study the competition between a London Stock Exchange order book (EuroSETS) and Euronext Amsterdam for Dutch firms and find that, even when there is formal entry barrier to arbitrageurs, the trade-through rate in their sample equals 73%. In the U.S., the SEC estimated that, prior to the implementation of Reg NMS, one out of eleven shares traded in NASDAQ listed stocks was a significant trade-trough (see Regulation NMS Adopting Release, 70 FR at 37502).

[19] This is consistent with the view of Stoll (2006), who argues:

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

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

Time is continuous and we set 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 μ of a perishable non-tradable good. The flow utility that an investor derives from holding a_t units of the asset at time t is

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

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

As explained in the introduction, the ϵ-shocks can capture time-varying liquidity demands, financing costs, hedging demands, specific investment opportunities, and specific requirements such as margins. For instance, a portfolio manager may need to buy and sell to meet inflows and outflows from investors. Traders may need to rebalance their portfolios to track their benchmarks. Corporate investors may need to sell their financial assets to finance real investments. Household may do the same for purchases of a durable good or a house. The parameter σ 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 γ measures the mean reversion of the utility flow process and is assumed, for simplicity, to be the same for all investors.[20]

Our paper focuses on the trading technology for the asset. For clarity, we describe here the case in which all investors trade at the same speed (later, we endogenize speed choices and consider venues with different speeds). The venue where investors trade the asset is characterized by the constant contact rate ρ. Conditional on being in contact, the market is Walrasian and clears at price p_t.[21] Any investor in contact with the venue at time t can trade at the price p_t. Investors who

[20] We introduce heterogeneity in σ and not in γ because the key point in our analysis is the link between gains from trade and speed. It is important to understand that a higher value of γ may imply *lower* gains from trade. Investors with an extreme value of γ are not eager to trade, since they can simply wait for their type to mean-revert. In particular, a high value of γ would not capture, per se, the idea of fleeting trading opportunities. This idea is better captured by a high value of σ.

[21] It would be straightforward to add bargaining with market makers and bid-ask spreads, but this would not

are not in contact simply keep their holdings constant.

Our assumptions about technology and preferences imply that the value function of a class-σ investor with current valuation ϵ_s and current asset holdings a at time t is

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

where the realization of the random type at time $s > t$ is ϵ_s and T denotes the next time the investor makes contact with the venue. Expectations are defined over the random variables T and ϵ_s and are conditional on the current type ϵ_t and the speed of the venue ρ.

3.2 Trading Equilibrium

We show that the asset price remains constant during the trading game. The value functions are thus time independent. Letting $a^*_{\sigma,\epsilon}$ denote the optimal choice of asset holding for type (σ,ϵ), equation (1) becomes simply

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

Following LR09, we define the adjusted holding utility as

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

LR09 (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 ϵ is i.i.d. with mean zero, $\mathbb{E}\left[u_{\sigma,\epsilon'}(a) \mid \epsilon\right] = \mu a$ for any a and any ϵ. This expected utility over ϵ' does not depend on σ or ϵ. This result implies that

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

Recall that G is the ex ante distribution of permanent types. Let $\hat{G}$ be the distribution of types *in the venue*. The total number of traders who join the venue is n and $\hat{G}$ is a cumulative distribution function. If all potential investors join the venue, we simply have $n = 1$ and $\hat{G} = G$. In the generic case, however, we have $n\hat{G} \leq G$ since some investors do not participate. Indeed, we shall see in the multiple-venue model that the support of $\hat{G}$ is typically not connected. We therefore present our results without placing restrictions on $\hat{G}$.

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

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

bring new insights compared to the results of Duffie, Garleanu, and Pedersen (2005) and LR09. For simplicity, we therefore assume competitive trading conditional on being in contact with the venue. A similar market mechanism is considered in the monetary economy of Rocheteau and Wright (2005) (which they label competitive equilibrium).

and the market clearing condition

$$\int_{\sigma} \rho \sum_{\epsilon=\pm 1} \frac{a^*(p;\sigma,\epsilon)}{2} d\hat{G}(\sigma) = \rho \bar{a}. \tag{4}$$

Proof. See Proposition 1 of LR09. The proposition only needs to be adapted to take into account heterogeneity in σ. Note that we assume $\epsilon = \pm 1$ with probability $1/2$. *Q.E.D.*

Equation (4) is the market clearing equation in flows. The supply (per capita) is $\bar{a}$ and a fraction ρ per unit of time is available for trading. Similarly, on the left-hand-side, we have the flow of demand. Note that the same ρ appears on both sides of of the equation and we could therefore drop it. In this example with one venue, market clearing in stock is enough to ensure market clearing in flows. We will, however, encounter a case where ρ and $\bar{a}$ are correlated across markets and where we need to write market clearing in flows.

There is 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-σ types always sell their entire holdings when they contact the venue. Moreover, there is a *marginal buyer*, that is, a type $\tilde{\sigma}$, that is indifferent between buying and not buying when $\epsilon = 1$. This **marginal trading type** is defined by

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

The demand function is therefore $a^* = 1$ when $\epsilon = +1$ and $\sigma \geq \tilde{\sigma}$. It is $a^* = 0$ in all other cases.

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 $\sigma > \tilde{\sigma}$ and nothing if $\sigma < \tilde{\sigma}$. The demand for the asset is $1/2(\hat{G}(\bar{\sigma}) - \hat{G}(\tilde{\sigma}))$. The ex ante supply of the asset (per capita) is $\bar{a}$. The market clearing condition is therefore

$$\frac{1-\hat{G}(\tilde{\sigma})}{2} = \bar{a}. \tag{6}$$

Note that the asset holdings of types $\sigma < \tilde{\sigma}$ are non stationary, since they never purchase the asset. They sell their holding $\bar{a}$ on first contact with the venue and never trade again. The fact that they stop trading is really just a consequence of linear preferences. With curvature in the utility function, low-σ types would trade repeatedly, but in smaller quantities. The only important point is that they trade less than the high-σ types. We call the traders with $\sigma < \tilde{\sigma}$ light traders, and traders with $\sigma \geq \tilde{\sigma}$ repeat or heavy traders.[22]

[22] Alternatively, we could let light traders receive larger shocks from time to time (by making σ itself random in addition to ϵ). Then all investors would trade as long as their σ were large enough, but some would trade more often than others. None of these extensions would change our main results, but they would complicate the (already complicated) analysis of entry and investment in speed. We choose to use the model with linear preferences and constant σ because it facilitates aggregation across heterogeneous types. Heterogeneity among investors is, of course, a key element of our analysis, but it is also a major source of complexity, so we need to make an assumption to keep the analysis tractable.

Over time, the assets move from the low-σ to the high-σ types and then keep circulating among the high-σ types in response to ϵ shocks and trading opportunities. It is easy to see that the price remains constant along the transition path. The gross supply of assets is always $\rho\bar{a}$. The gross demand from high-σ types is always $\rho\left(1-\hat{G}\left(\tilde{\sigma}\right)\right)/2$. From equation (6), the market always clears.[23]

We can now characterize the steady-state distribution among types $\sigma > \tilde{\sigma}$. Let $\alpha_{\sigma,\epsilon}(a)$ be the share of class-σ investors with trading type ϵ 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 venue. Outflows result only from changes of ϵ from +1 to -1, which occurs with intensity $\gamma/2$. There are two sources of inflow: types $(\epsilon = -1, a = 1)$ that switch to $\epsilon = 1$ and types $(\epsilon = +1, a = 0)$ that purchase one unit when they contact the venue. In steady state, outflows must equal inflows:

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

The dynamics for types $(\epsilon = -1, a = 0)$ are similar: $\frac{\gamma}{2}\alpha_{\sigma,-}(0) = \rho\alpha_{\sigma,-}(1) + \frac{\gamma}{2}\alpha_{\sigma,+}(0)$. For types $(\epsilon = +1, a = 0)$ and $(\epsilon = -1, a = 1)$, trade creates outflows, yielding $\left(\frac{\gamma}{2}+\rho\right)\alpha_{\sigma,+}(0) = \frac{\gamma}{2}\alpha_{\sigma,-}(0)$ and $\left(\frac{\gamma}{2}+\rho\right)\alpha_{\sigma,-}(1) = \frac{\gamma}{2}\alpha_{\sigma,+}(1)$. Finally, the shares must add up to one; therefore: $\sum_{\epsilon=\pm,a=0,1}\alpha_{\sigma,\epsilon}(a) = 1$. We summarize our results in the following lemma

Lemma 2. *The trading equilibrium is characterized by the price p and marginal trading type $\tilde{\sigma}$ defined in equations (5) and (6), respectively. The transition dynamics are as follows. The price remains constant while asset holdings shift from low-σ types to high-σ types. Low-σ types $(\sigma < \tilde{\sigma})$ sell their initial holdings $\bar{a}$ and do not purchase the asset again. High-σ types $(\sigma \geq \tilde{\sigma})$ buy when $\epsilon = 1$ and sell when $\epsilon = -1$. The distribution of holdings among high-σ types converges to the steady-state distribution of well-allocated assets $\alpha_{\sigma,+}(1) = \alpha_{\sigma,-}(0) = \frac{1}{4}\frac{2\rho+\gamma}{\gamma+\rho}$ and misallocated assets $\alpha_{\sigma,+}(0) = \alpha_{\sigma,-}(1) = \frac{1}{4}\frac{\gamma}{\gamma+\rho}$.*

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

$$\mathcal{V} = \frac{\rho}{2}\left(\alpha_{\sigma,+}(0) + \alpha_{\sigma,-}(1)\right) \times \left(1-\hat{G}\left(\tilde{\sigma}\right)\right). \tag{8}$$

The right-hand side of equation (8) is given by the product of the contact rate, the proportion of agents with misallocated assets, and the population of steady-state traders. Using equation (6) and the equilibrium expressions for $\alpha_{\sigma,+}(0)$ and $\alpha_{\sigma,-}(1)$ in Lemma 2, one obtains the (free market participation) equilibrium trade volume rate $\mathcal{V} = \frac{\rho}{2}\frac{\gamma}{\gamma+\rho}\overline{a}$.

The best way to understand Lemma 2 is to focus on gains from trade and deviations from the Walrasian allocation, which maximizes the gains from trade. Taking the limit $\rho \to \infty$ in equations (5) and (6), we obtain the following lemma.

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

Lemma 3. *The Walrasian equilibrium with $\bar{a} \leq 1/2$ has a price $p_w = \frac{1}{r}\left[\mu + G^{-1}(1-2\bar{a})\right]$, the instantaneous volume rate equals $\mathcal{V}_w = \frac{\overline{a}\gamma}{2}$, and total gains from trade (welfare) are given by $\frac{1}{2r}\int_{G^{-1}(1-2\overline{a})}^{\overline{\sigma}} \sigma dG(\sigma)$.*

Note again the symmetry around $\bar{a} = 1/2$. When $\bar{a} < 1/2$, the price is higher than the mean value μ/r. In that case, types $\sigma < \tilde{\sigma}$ sell and then do not buy again. For a given distribution of investors $\hat{G}$ in the market, the difference $p - \mu/r$ increases with the speed of trading. When $\bar{a} > 1/2$, the price is lower than μ/r. At the investor level, σ indexes the gains from trade. The general properties are that high speed brings the equilibrium closer to the Walrasian outcome and that that low-σ investors trade less. A faster venue always realizes more gains from trade than a slower one. Since high-σ traders have higher gains from trade, they are more willing to pay for speed. When $\overline{a} < 1/2$ the price is higher in the faster venue and when $\overline{a} > 1/2$ the price is lower in the faster venue. Finally, when $\bar{a} = 1/2$, there is no unique equilibrium price but a range that includes μ/r.

3.3 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 venue. We do so in two steps. We first compute the steady-state value functions for investors that continue trading. We later compute the ex ante values, taking into account the transition dynamics.

Consider the steady-state value functions for types $\sigma > \tilde{\sigma}$. For the types holding the assets we have

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

and $rV_{\sigma,+}(1) = \mu + \sigma + \frac{\gamma}{2}\left[V_{\sigma,-}(1) - V_{\sigma,+}(1)\right]$. Analogous expressions hold for the types not holding the assets, forming a system of equations that can be solved to compute the explicit form of the value functions. Note that the asset price p is pinned down by the marginal value (minimum type in each venue). For now, we use it as a (venue-specific) parameter. Also note that the no-trade outside option of any investor is

$$W_{out} = \frac{\mu\bar{a}}{r}. \tag{9}$$

The following proposition characterizes the ex ante value functions, i.e., those of an investor that knows his permanent type σ but not the temporary preference shock, taking into account the transition dynamics leading up to the steady-state allocations.

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

$$W(\sigma, \tilde{\sigma}, s) - W_{out} = \frac{s\tilde{\sigma}}{r}\bar{a} + \frac{s}{2r}\max(0; \sigma - \tilde{\sigma}), \tag{10}$$

where the marginal trading type $\tilde{\sigma}$, defined in equation (5), increases in p and decreases in ρ and where effective speed s is defined by

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

The net value of participation, $W - W_{out}$, is composed of two parts. One is the option to sell the asset on the exchange: $\frac{s\bar{a}\tilde{\sigma}}{r} = \frac{\rho}{r+\rho}\left(p - \frac{\mu}{r}\right)\bar{a}$. It is independent of σ and is the value that can be achieved by all types $\sigma < \tilde{\sigma}$ with a "sell and leave" strategy. The term $\frac{\rho}{r+\rho}$ is the discount due to expected trading delays. The second part, $\frac{s}{2r}\max(0; \sigma - \tilde{\sigma})$, is the value of trading repeatedly and it depends on the type σ. This part of the value function is super modular in (s, σ). Proposition 1 provides the building block for our analysis of the industrial organization of financial markets.

4 Market Structure and Welfare

This sections provides a formal definition of the market structure equilibrium and characterizes the regulation problem.

4.1 Market Structure Equilibrium

Our market structure game is a sequential game where, taking regulations as a given, venues decide whether to enter, select trading speeds, and post membership fees. Venues make these decisions in Stages II to IV in Figure 2. We introduce a fixed entry cost κ to analyze the entry game (Stage II). Venues face the same increasing and convex investment cost function $C(s)$ when they choose their speeds (Stage III). Venues compete in fees à la Bertrand (Stage IV). Let q_i be the membership fee posted by venue i, and let n_i be the number of investors who join venue i. The total net profits of venue i are therefore $q_i n_i - C(s_i) - \kappa$. Given venues' decisions, investors decide which venue to join at Stage V. Participation decisions are described by a mapping from types σ to venues $\mathcal{P} : [0, \bar{\sigma}] \longrightarrow \{0, 1, ..., I\}$, where $\mathcal{P}(\sigma) = i$ means joining venue i and $\mathcal{P}(\sigma) = 0$ means staying out. If an investor joins venue i, it pays a membership fee q_i and is then allowed to use the trading venue (staying out costs nothing, so formally $q_0 = 0$ and $W = W_{out}$).[24] Recall that we have defined $\hat{G}_i(.)$ as the distribution of types in venue i. Let us now formally define an equilibrium of the game.

Definition 1. *A market structure equilibrium is a set of participation decisions by investors and entry, speed, and fee strategies by trading venues, such that*

- *Venues maximize profits: The sequence of entry, speed, and fee strategies is a Nash equilibrium of each corresponding stage game (Stages II to IV).*
- *Participation decisions are optimal: For all σ and all i, $\mathcal{P}(\sigma) = i$ implies $W(\sigma, \tilde{\sigma}_i, s_i) - q_i \geq W(\sigma, \tilde{\sigma}_j, s_j) - q_j$ for all $j \neq i$; reciprocally, when $W(\sigma, \tilde{\sigma}_i, s_i) - q_i > W(\sigma, \tilde{\sigma}_j, s_j) - q_j$ for all $j \neq i$, then we must have $\mathcal{P}(\sigma) = i$.*

[24] Appendix E also analyzes competition in trading fees as opposed to membership fees.

- *The distribution of types in venue i is consistent with individual participation decisions:* $n_i = \int_{\mathcal{P}(\sigma)=i} dG(\sigma)$ *and* $\hat{g}_i(\sigma) = \frac{g(\sigma)}{n_i}\mathbf{1}_{\mathcal{P}(\sigma)=i}$ *for all* $\sigma \in [0, \bar{\sigma}]$.
- *The investor market clears:* $\sum_{i\in\mathcal{I}} n_i \hat{G}_i(\sigma) = G(\sigma)$ *for all* $\sigma \in [0, \bar{\sigma}]$.
- *Subsequent asset prices and marginal types satisfy equations (5) and (6).*

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

4.2 Welfare and Regulation

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

$$\mathcal{W} \equiv \underbrace{\sum_{i=1:I} n_i \int_{\sigma} (W(\sigma, \tilde{\sigma}_i, s_i) - W_{out}) d\hat{G}_i(\sigma)}_{\text{Total gains from trade}} - \underbrace{\sum_{i=1:I} (\kappa + C(s_i))}_{\text{Entry and speed investment}}. \tag{12}$$

Welfare gains are the sum of investors' expected participation gains minus the fixed entry costs and the costs of investments in speed. Effective speed s_i enters the calculation because it affects allocative efficiency. The following lemma characterizes the welfare function with two venues, taking into account the results of Section 3.

As a benchmark, we can consider the case in which the planner can decide entry, speed, and pricing. This is not realistic, but it will help us build intuition for our results on the regulation of speed and entry in otherwise decentralized markets. The planner faces the same cost structure as the private sector: a setup cost κ for each venue, a default effective speed $\underline{s}$ available at no cost, and a cost function $C(s)$.

Lemma 4. *The unconstrained planner's solution is to operate one venue with full participation and a level of speed satisfying* $\frac{\partial C}{\partial s}(s^*) = \frac{\mathbb{E}[\sigma]}{2r}$.

The proof follows directly from equations (10) and (12). The setup costs are fixed and there is no marginal cost of adding traders to a venue. The unconstrained solution is then clearly to open one fast venue with full participation financed by lump-sum taxes on all agents. Of course, the venue is then a loss-making operation that must be subsidized by lump transfers from the agents. In Appendix G, we also solve a restricted planner problem in which direct subsidies are ruled out and therefore venues must break even. Even in that case, however, we can show that the planner chooses to operate only one venue. It chooses a speed that is lower than that in Lemma 4, but still higher than what a monopoly would choose. We use the constrained planner as a welfare benchmark in Section 8.

Entry and Speed Regulations. In the remainder of the paper, we analyze monopoly and duopoly equilibria and study the regulation of speed and entry by a *regulator*. By regulator, we mean a restricted planner, that is, an authority that can only affect one dimension, such as speed, while taking as a given the structure of the game (entry and pricing).

The regulator fundamentally wants to (i) increase participation, (ii) avoid duplication costs, and (iii) increase speed. A basic virtue of entry is to foster competition and reduce prices. This is the classic case for inter-market competition when liquidity externalities are moderate (e.g., Economides (1996)). With fixed costs, however, there can be excessive entry. This creates a trade-off between (i) and (ii). The fact that speed can be used for differentiation creates a tradeoff between (i) and (iii). These tradeoffs are analyzed in Sections 6 and 7.

Trading Regulation: Price Protection Rule. An important aspect of the market structure game is to study the impact of regulations that affect how asset prices in different venues relate. There are two polar cases of analysis.

Definition 2. *We say that there is* ***segmentation*** *if venues do not execute trades coming from investors of another venue. If instead venues give access to the same market, with a single clearing price, we say there is* ***integration****.*

Definition 2 clarifies what we mean by a venue. In our model, a venue is an access gate to a market where transactions clear at a market price, as described in Figure 3. The market clearing condition is given in equation (4). Segmentation means that the venues give access to different markets and therefore to different market clearing prices. Under segmentation, an investor joins a venue and never buys from–and never sells to–an investor from the other venue. Integration means that the two venues give access to the same market with a single market clearing price. The trades cleared in that market come from both venues. A fast venue simply provides faster access to the market.

The real world is, of course, somewhere in between these two polar cases. Arbitrage is imperfect because of many well-recognized frictions. That is why we assume that integration is enforced by a price protection rule and we refer to perfectly integrated markets as the *protected case* hereafter.[25] We will show how trading regulations affect the expected profits of venues and therefore their entry decisions.

5 Fee Competition and Venue 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

[25] This is our stylized way of capturing access and trade-through rules in the SEC's Reg. NMS. The distinction between top-of-the-book (U.S. version) and full-depth (Canadian version) protection is not material in our model, since we only consider unitary orders.

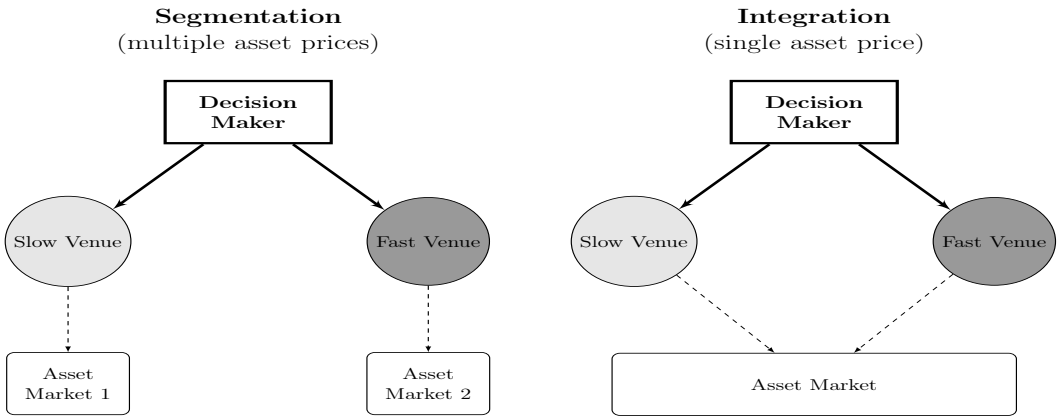

Figure 3. Fragmented markets: Analysis cases.

protection in the trading game affects these equilibrium outcomes. In other words, we analyze how trading regulations affect the ex ante competition among venues. In this section we take the set of venues as a given, as well as their speed. We endogenize speed in Section 6 and entry in Section 7. We define venue 2 as the fast one, so $s_2 > s_1$.

5.1 Monopoly

Consider the case of one venue charging a membership fee q. Recall that we have defined $\tilde{\sigma}$ as the marginal trading type. We now define $\hat{\sigma}$ as the **marginal participating type**, i.e., as the largest solution to $W(\hat{\sigma}, \tilde{\sigma}, s) - W_{out} \equiv q$. The value function (10) is flat for all types below the marginal trading type $\tilde{\sigma}$. In any interior solution, the marginal trading type must also be the marginal participating type: $\hat{\sigma} = \tilde{\sigma}$. The marginal trading type is indifferent between joining the venue and not joining the venue. Thus we have $W(\hat{\sigma}, \hat{\sigma}, s) - W_{out} = q$, which implies

$$q_m = \frac{\bar{a}}{r} s_m \hat{\sigma}_m. \tag{13}$$

All types below $\hat{\sigma}$ are indifferent between joining and staying out. Let δ be the mass of light traders. Market clearing requires $\delta = (1/2\overline{a} - 1)(1 - G(\hat{\sigma}))$. When $\bar{a}$ is less than $1/2$, there are δ light traders who join to sell their asset but do not trade repeatedly.[26] The equilibrium is depicted in Figure 4. We have an interior solution (where some traders do not join) as long as $\delta < G(\hat{\sigma})$, that is, as long as $G(\hat{\sigma}) > 1 - 2\overline{a}$. 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-σ type investors to ensure the existence of interior solutions.

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

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

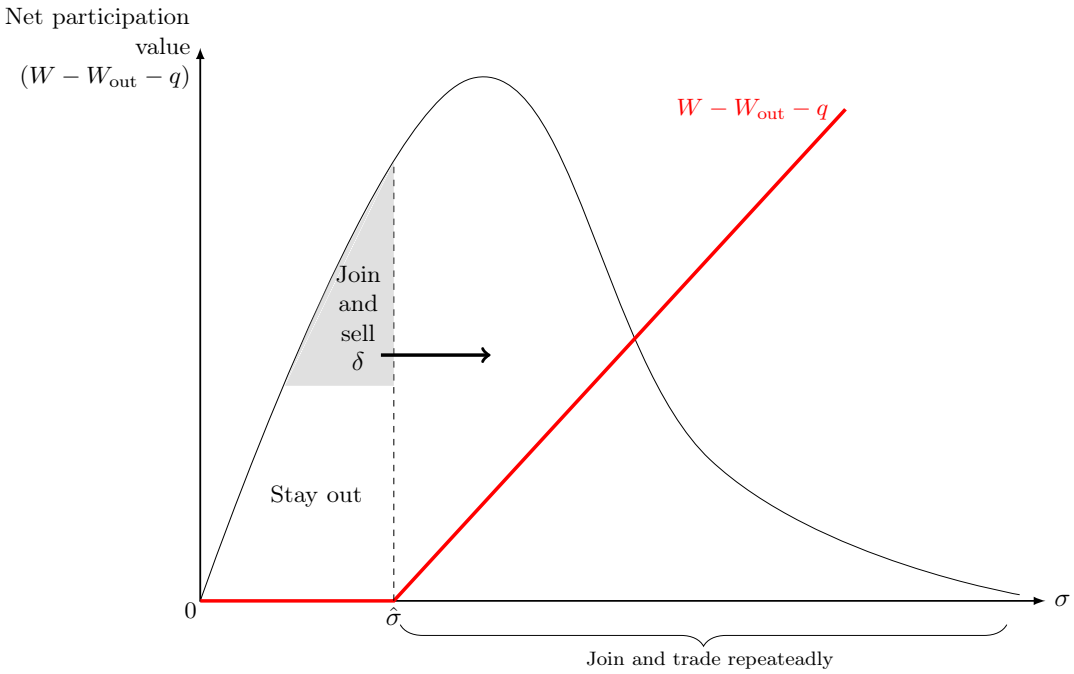

Figure 4. Investor affiliation choices with one trading venue.

clearing:

$$\pi_m = \frac{s_m \hat{\sigma}_m}{2r} \left(1 - G\left(\hat{\sigma}_m\right)\right). \tag{14}$$

The program of the monopolist is simply to maximize (14) with respect to q and subject to (13), which leads to the following lemma.

Lemma 5. *The monopolist chooses a level of participation $\hat{\sigma}_m$ that is independent of its speed and satisfies*

$$1 - G\left(\hat{\sigma}_m\right) = g\left(\hat{\sigma}_m\right) \hat{\sigma}_m. \tag{15}$$

Equation (15) is the first-order condition of the monopoly problem.[27] The result that $\hat{\sigma}_m$ is independent of the speed in the venue comes from our assumption that the marginal cost of adding traders to an existing venue is zero. This assumption allows us to focus on speed choices. The monopoly fee q_m is proportional to the effective speed s.

Let us now consider the duopoly. Since we assume that venues compete in fees à la Bertrand, the equilibrium without differentiation implies zero fees and zero profits. The interesting case arises for differentiation by speed.

[27] First-order conditions are sufficient in this environment. Note that since g is positive and log-concave, it is also quasi-concave. Thus the tail distribution $1 - G$ is quasi-concave as well, which results in the quasi-concavity of $\pi = \sigma\left(1 - G\left(\sigma\right)\right)$. If c were the marginal cost of adding a trader to the venue, profits would be $\pi = (q - c)\frac{1-G(\hat{\sigma})}{2\bar{a}} = \left(\frac{s\bar{a}}{r}\hat{\sigma} - c\right)\left(1 - G\left(\hat{\sigma}_m\right)\right)$ and the first-order condition would be $\left(1 - G\left(\hat{\sigma}_m\right)\right) = g\left(\hat{\sigma}_m\right)\left(\hat{\sigma}_m - \frac{rc}{s\bar{a}}\right)$. In this case $\hat{\sigma}_m$ would depend on s. This effect does not add new insight, so we drop it. Moreover, we think that this kind of marginal costs is less important than fixed and investment costs in infrastructure. These costs are at the heart of our analysis.

5.2 Duopoly with Segmented Prices

Consider first the case in which venues are segmented and thus prices can then be different. The key issue is to understand the affiliation choices of investors. We proceed by backward induction. Investors anticipate that each venue i will be characterized by its speed and price, which together define the marginal trading type $\tilde{\sigma}_i$. Investors can then estimate their value functions W, defined in equation (10). The net value from joining venue $i = 1, 2$ is $W(\sigma, \tilde{\sigma}_i, s_i) - W_{out} - q_i$. These value functions are depicted in the middle panel of Figure 5.

It is important to keep in mind that the value functions are not super-modular for low-σ types. In addition, we know that each venue must attract a mass δ of light traders. Because these types must be indifferent between joining and staying out, we must have $W(\tilde{\sigma}_i, \tilde{\sigma}_i, s_i) - W_{out} - q_i = 0$ in both venues. In other words, as in the case of the monopoly, the marginal trading type $\tilde{\sigma}_i$ must be indifferent between participating in venue i and not. Therefore, we must have

$$q_i = \frac{\bar{a} s_i \tilde{\sigma}_i}{r}, \; i = 1, 2. \tag{16}$$

Note, however, an important difference from the monopoly case. The marginal trader in venue 2, $\tilde{\sigma}_2$, would indeed be indifferent between joining venue 2 and not participating. But it is clear from Figure 5 that $\tilde{\sigma}_2$ in fact joins venue 1. This means that, with two venues, marginal trading types and marginal participating types are not the same. They coincide only for the slowest market: $\hat{\sigma}_1 = \tilde{\sigma}_1$ but $\hat{\sigma}_2 > \tilde{\sigma}_2$. We define a new marginal type, $\hat{\sigma}_2$, that is indifferent between joining venue 1 and joining venue 2. By definition, this type must be such that $W(\hat{\sigma}_2, \tilde{\sigma}_2, s_2) - q_2 = W(\hat{\sigma}_2, \tilde{\sigma}_1, s_1) - q_1$. This implies

$$\frac{s_2 \bar{a} \tilde{\sigma}_2}{r} + \frac{s_2}{2r}(\hat{\sigma}_2 - \tilde{\sigma}_2) - q_2 = \frac{s_1 \bar{a} \tilde{\sigma}_1}{r} + \frac{s_1}{2r}(\hat{\sigma}_2 - \tilde{\sigma}_1) - q_1$$

and, therefore, using equation (16), we obtain

$$\hat{\sigma}_2 = \frac{r}{\bar{a}} \frac{q_2 - q_1}{s_2 - s_1}. \tag{17}$$

Note that $\hat{\sigma}_1 = \tilde{\sigma}_1 < \tilde{\sigma}_2 < \hat{\sigma}_2$. The set of types that join venue 2 cannot be continuous over an interval. It is composed of all the types above $\hat{\sigma}_2$ and some types below $\hat{\sigma}_1$. The affiliation is depicted in the top panel of Figure 5.

Market clearing in venue 2 requires $(1 - G(\hat{\sigma}_2) + \delta_2)\bar{a} = \frac{1-G(\hat{\sigma}_2)}{2}$. The second-stage payoff for the fast venue under segmentation is $\pi_2^{seg} = q_2(1 - G(\hat{\sigma}_2) + \delta_2) = q_2 \frac{1-G(\hat{\sigma}_2)}{2\bar{a}}$. Market clearing for the slow venue requires $(G(\hat{\sigma}_2) - G(\hat{\sigma}_1) + \delta_1)\bar{a} = \frac{G(\hat{\sigma}_2)-G(\hat{\sigma}_1)}{2}$. The payoff for the slow venue is $\pi_1^{seg} = q_1 \frac{G(\hat{\sigma}_2)-G(\hat{\sigma}_1)}{2\bar{a}}$. The affiliation of investors to venues 1 and 2 is given by the marginal types described in (13) and (17), respectively. Venues 1 and 2 simultaneously solve $\max_{q_1} \frac{q_1}{2\bar{a}}(G(\hat{\sigma}_2) - G(\hat{\sigma}_1))$ and $\max_{q_2} \frac{q_2}{2\bar{a}}(1 - G(\hat{\sigma}_2))$, respectively. The first-order conditions from the previous system result in the following lemma.

Lemma 6. *In a segmented duopoly, marginal participating types* $(\hat{\sigma}_1^{seg}, \hat{\sigma}_2^{seg})$ *solve the system*

$$1 - G(\hat{\sigma}_2) = g(\hat{\sigma}_2)\left(\hat{\sigma}_2 + \frac{\hat{\sigma}_1}{s_2/s_1 - 1}\right), \tag{18}$$

$$G(\hat{\sigma}_2) - G(\hat{\sigma}_1) = \left(g(\hat{\sigma}_1) + \frac{g(\hat{\sigma}_2)}{s_2/s_1 - 1}\right)\hat{\sigma}_1. \tag{19}$$

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

The system of equations (18) and (19) shows equilibrium participation depends only on the degree of speed differentiation $s_2/s_1 \in [1, \infty)$.

5.3 Duopoly with Protected Prices

Consider now the case in which both venues provide access to the same market with a single market clearing conditions and a single price p. The analysis of venues with protected prices is complex for several reasons. One reason is that the price is not constant over time: the price must be relatively high initially to ensure market clearing when assets are concentrated in the slow venue.[28] Over time, assets migrate to the fast venue, the trade-able supply increases, and the price decreases to its long run equilibrium. These transition dynamics complicate the value functions without adding new insights so we consider here an approximation where we assume that the price remains constant. We compute the transitions dynamics in an Appendix and we show that the approximation is good as long as r is small relative to ρ, which is clearly the case in practice. Intuitively, assets migrate at the speed ρ and the net present value calculations are is essentially determined by the long run gains from trade.[29]

Venue 1 is still characterized by the indifference condition (16) for the marginal trading type $\tilde{\sigma}_1$. However, this condition does not hold for venue 2 because low-σ types can join venue 1 and effectively sell their assets to investors in venue 2. Instead, the asset price is the same in both venues. From equation (5), this implies the constraint $\left(1 + \frac{\gamma}{r+\rho_1}\right)\tilde{\sigma}_2 = \left(1 + \frac{\gamma}{r+\rho_2}\right)\tilde{\sigma}_1$. This means that $\tilde{\sigma}_2 < \tilde{\sigma}_1$. The indifference condition for $\hat{\sigma}_2$ is still $W(\hat{\sigma}_2, \tilde{\sigma}_2, s_2) - q_2 = W(\hat{\sigma}_1, \tilde{\sigma}_1, s_1) - q_1$. We show in Appendix C that this leads to $\hat{\sigma}_2 = \frac{2r}{s_2 - s_1}\left(q_2 - \frac{z}{2\bar{a}}q_1\right)$, where $z \equiv 1 - \frac{1+r/\rho_1}{1+r/\rho_2}(1 - 2\bar{a})$. The structure of the value functions is still as depicted in the bottom panel of Figure 5. There is now only one market clearing condition. Therefore, the light traders join venue 1, where they can sell at a higher price because they can sell to investors in venue 2. We then have $\delta_2 = 0$ and the market clearing condition is $(1 - G(\hat{\sigma}_1) + \delta_1)\bar{a} = \frac{1}{2}(1 - G(\hat{\sigma}_1))$. The following lemma summarizes the protected price equilibrium.

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

[29] In the Appendix, we show that asset migration is given by $m(t) = (1 - G(\hat{\sigma}_2))\left(\frac{1}{2} - \bar{a}\right)\left(1 - e^{-\rho_2 t}\right)$ and the value function of temporary trader is $W(t) = A + Be^{-\rho_2 t}$, where A is the long run solution that we use in the main text. The important point is that the dynamics happen at speed ρ_2 which is typically much faster that the rate of time discounting r. For a natural case, we find that the degree of approximation of our value function is around 6% for the most affected traders (and much less for the other ones).

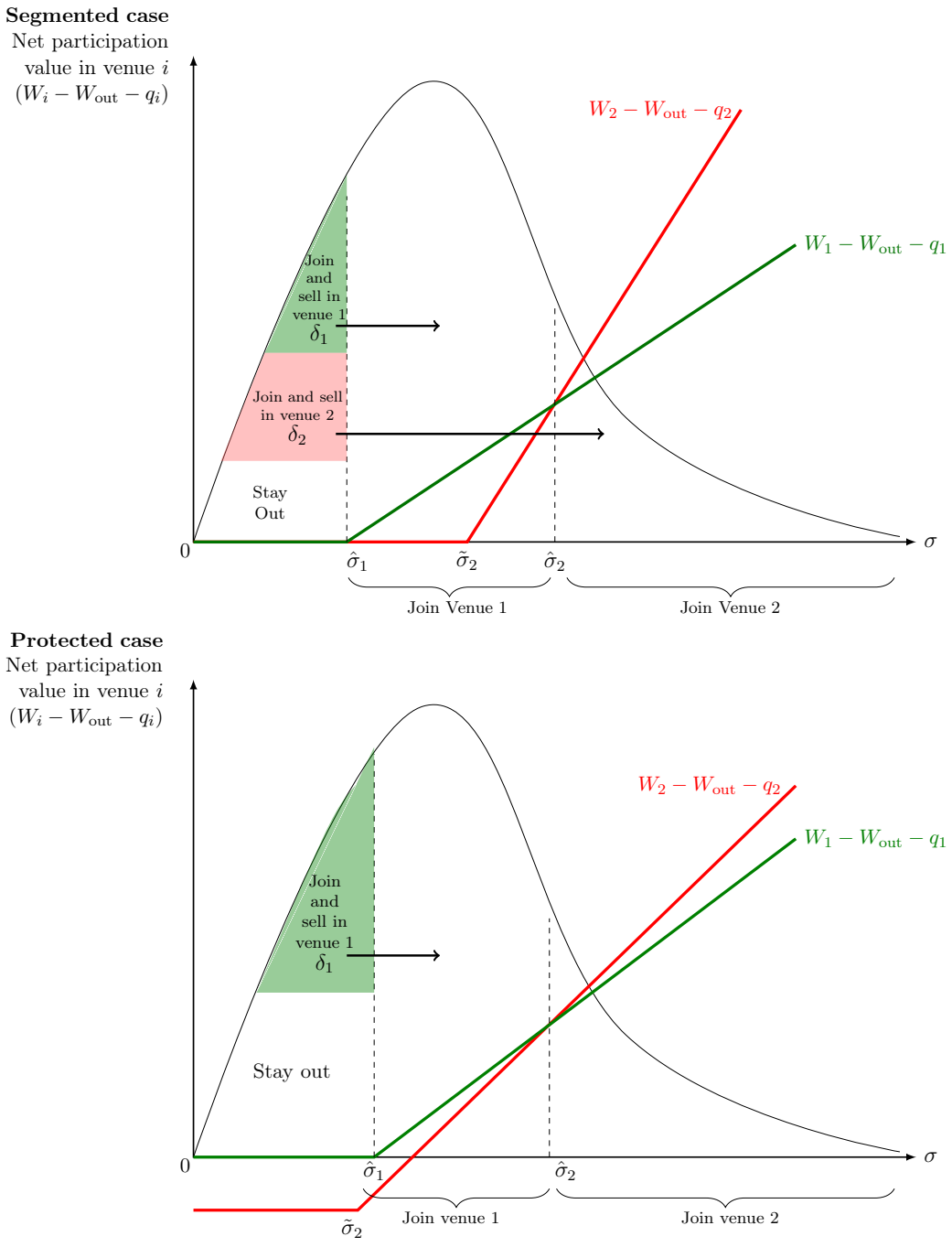

Figure 5. Investor affiliation choice with two trading venues.

Lemma 7. *In a duopoly with protected prices, marginal participating types* $\left(\hat{\sigma}_1^{prot}, \hat{\sigma}_2^{prot}\right)$ *solve the system*

$$\begin{aligned} 1 - G(\hat{\sigma}_2) &= g(\hat{\sigma}_2)\left(\hat{\sigma}_2 + z\frac{\hat{\sigma}_1}{s_2/s_1 - 1}\right), \\ G(\hat{\sigma}_2) - \frac{G(\hat{\sigma}_1)}{2\bar{a}} &= \left(\frac{g(\hat{\sigma}_1)}{2\bar{a}} + z\frac{g(\hat{\sigma}_2)}{s_2/s_1 - 1}\right)\hat{\sigma}_1 + 1 - \frac{1}{2\bar{a}}. \end{aligned}$$

Note that the allocation under protected prices converges to that under segmented markets when $\bar{a} = 1/2$. Recall that in that case we can set the price to μ/r without loss of generality.

5.4 Investor Participation in Different Market Structures

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

Assumption 1. The distribution of types σ is such that, for all σ

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

Assumption 1 is needed to prove a basic yet important result. At the core of our analysis is the idea that vertical differentiation (via investment in trading technology) decreases price competition. We then need to show that, in equilibrium, prices are higher and participation is lower when trading speeds are more differentiated. Assumption 1 is needed to prove this comparative static. Assumption 1 is not restrictive: It holds for all the distributions that we consider in our numerical analysis and many more.[30] Some results, however, can only be proven for specific classes of distributions and we mainly use two such classes.

Definition of Distributions. *To derive analytical results, we consider the exponential distribution* $G(\sigma) = 1 - e^{-\frac{\sigma}{\nu}}$ *and the uniform distribution* $G(\sigma) = \frac{\sigma}{\bar{\sigma}} 1_{\sigma \in [0, \bar{\sigma}]}$.

The following proposition characterizes the equilibrium of the affiliation game.

Proposition 2. *The equilibrium of the affiliation game has the following properties:*
(i) Competition among venues increases participation. *With or without price protection and for a given speed, participation in the fast venue alone is higher than total participation under a monopoly, that is,* $\hat{\sigma}_2 < \hat{\sigma}_m$. *Total participation is even higher, since* $\hat{\sigma}_1 < \hat{\sigma}_2$.

[30] For example, it holds for exponential, normal, log-normal, Pareto, Weibull, inverse Gaussian, gamma, and Kumaraswamy distributions.

(ii) Speed differentiation, defined as s_2/s_1, relaxes price competition. *Under Assumption 1, participation with a duopoly is lower ($\hat{\sigma}_1$ and $\hat{\sigma}_2$ are higher) when speeds are more differentiated.* ***(iii) Price protection increases the profits of the slow venue and decreases total participation****, that is, $\pi_1^{prot} \geq \pi_1^{seg}$ and $\hat{\sigma}_1^{prot} \geq \hat{\sigma}_1^{seg}$. Conditional on speed, price protection has an ambiguous impact on participation in the fast venue. (The proof is analytical for exponential and uniform distributions, and numerical in other cases).*

The intuition for (i) is simply that price competition increases participation. A result that is perhaps less obvious is that participation in the fast venue *alone* is already higher than total participation with a monopoly. Point (ii) helps us understand how speed affects the affiliation game, and more precisely what happens to fees (q) and marginal types ($\hat{\sigma}$) when speed differentiation s_2/s_1 increases. The system is given by equations (18) and (19), which determines the participation as a function of the degree of speed differentiation s_2/s_1. We show in Appendix B that $\frac{\partial \hat{\sigma}_1}{\partial (s_2/s_1)} > 0$ and $\frac{\partial \hat{\sigma}_2}{\partial (s_2/s_1)} > 0$. This result is fundamental since it shows that differentiation decreases competition and therefore decreases participation.

Point (iii) shows that price protection has two main effects. First, it increases the comparative advantage of the slow venue because it allows its investors to trade with investors from fast venue. The per-capita gains from trade are higher in the fast venue because of higher speed and because the fast venue attracts investors with high σ values, as can be seen from equation (10) and the fact that $\frac{s\bar{a}\tilde{\sigma}}{r} = \frac{\rho}{r+\rho}\left(p - \frac{\mu}{r}\right)\bar{a}$. Under price protection, the investors in the slow venue benefit from these gains from trade, which makes them more willing to join the slow venue. The second effect of protection is to soften the price elasticity of the marginal type $\hat{\sigma}_2$ by making the value function steeper. With more demand and less competition, venue 1 can charge a higher price and make higher profits. Interestingly, the low competition effect is strong enough that participation decreases $\hat{\sigma}_1^{prot} \geq \hat{\sigma}_1^{seg}$.[31] On the other hand, participation in the fast venue can go up or down. If venue speeds are relatively similar, then the decrease in the price elasticity leads to large increase in q_1 and an increase in participation in venue 2. If, instead, venue speeds are very different, the change in q_1 is smaller and participation in venue 2 may decrease. In our calibrated asset markets, we typically find that price protection leads to an increase in participation for venue 2 (see Table VII).

Proposition 2 plays an important role 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 following sections.

[31] These results hold irrespective of the value of $\bar{a}$. When $\bar{a} < 1/2$, sellers are on the short side and they benefit from the high price in the fast venue. When $\bar{a} > 1/2$, buyers are on the short side and they are attracted by the low price in the fast venue. When there is excess demand, the price in the fast venue is higher than in the slow venue. The converse holds when there is excess supply. In our simple setup with linear utility and [0,1] holdings, the condition for excess demand is simply $\bar{a} < 1/2$. In a more general case with continuous holdings, the conditions are more complicated but the intuition remains the same. We numerically checked the robustness of the result $\pi_1^{prot} \geq \pi_1^{seg}$ to alternative assumptions about the underlying distribution of σ.

5.5 Extensions

Asset Prices. We can also relate these results to equilibrium asset prices. From equation (5), we know that the equilibrium asset price in venue i is given by $p_i = \frac{\mu}{r} + \frac{\tilde{\sigma}_i}{r} \times \left(\frac{r+\gamma s_i}{r+\gamma}\right)$. The key differences with the benchmark case of Duffie et al. (2005) is that, in our study, both participation decisions among heterogeneous traders and market contact frictions (driven by the venue speed) are endogenously determined. For example, under price protection, $\hat{\sigma}^{prot}$ is given by Lemma 7. Under segmentation, there are two prices reflecting $\hat{\sigma}_1^{seg}$ and $\hat{\sigma}_2^{seg}$. Consequently, regulations, the venue structure, and speed and affiliation choices all affect asset prices. This framework then offers a rich set of empirical predictions on liquidity and asset prices which are developed by Pagnotta (2015).

Trading fees. In our benchmark model, investors pay membership fees and then trade freely. We think that this setup describes modern financial markets, where the relevant costs relate to trading infrastructure and investment in information technology more than marginal costs per trade. Trading fees can still be significant in some markets, however, so we analyze them carefully in Appendix E. We assume that each trade entails a cost ϕ: If the market price is p, a seller effectively receives only $p - \phi$, while a buyer effectively pays $p + \phi$.[32] Proposition 9 in the same appendix summarizes our results. Trading fees induce an additional range of investor types that are characterized by partially inactive trading patterns. They also improve venue price discrimination and can thus increase average profits. Although this extension convey interesting economic intuition about venues pricing strategies, it entails significant complexity and it does affect our main results. In particular, trading fees do not affect results (i)-(ii)-(iii) in Proposition 2.

Multi-venue Participation. We have also analyzed the possibility that some traders may choose to pay both membership fees and trade in both venues (see in Appendix F). The key issue is whether multi-venue traders send both buy and sell orders to both venues. 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 venues. Alternatively, multi-venue traders may prefer to wait for a good deal rather than sell at a low price in the slow venue or buy at a high price in the fast venue. In the context of our model, however, we show that multi-venue traders do not play a quantitatively important role because only investors with extremely large σ would choose to join two venues. This possibility is clearly interesting, especially in its implications on asset prices and arbitrage, but it is left for future research.

6 Trading Speed

This section analyzes investment in trading technology, taking as a given the set of active venues. We study entry in the next section. We focus on the case in which $\bar{a} = 1/2$ to separate this analysis

[32] Another way to introduce trading fees is to use bargaining as in Duffie et al. (2005).

from that of trading regulations in the previous section.[33] Based on the analysis in Sections 4 and 5, we can rewrite equation (12) as the following lemma.

Lemma 8. *Social welfare with one venue is*

$$\mathcal{W}(1) = \frac{s}{2r}\int_{\hat{\sigma}}^{\bar{\sigma}} \sigma dG(\sigma) - C(s) - \kappa \tag{20}$$

and with two venues it is

$$\mathcal{W}(2) = \frac{s_1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_2} \sigma dG(\sigma) + \frac{s_2}{2r}\int_{\hat{\sigma}_2}^{\bar{\sigma}} \sigma dG(\sigma) - \sum_{i=1,2} C(s_i) - 2\kappa. \tag{21}$$

When we want to derive closed-form solution, we assume that the cost of speed ρ is linear, $c\rho$, with $c \geq 0$. Given that $s \equiv \frac{\rho}{r+\gamma+\rho}$, this implies the following cost expression.

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

6.1 Venue Speed Choices

Monopoly. Lemma 5 shows that the participation cutoff $\hat{\sigma}_m$ chosen by a monopolist does not depend on its effective speed s. The monopoly chooses its speed to maximize $s\frac{\hat{\sigma}_m}{2r}(1 - G(\hat{\sigma}_m)) - C(s)$, as in the following proposition.

Proposition 3. *The monopolist chooses a speed level s_m such that $\frac{\partial C}{\partial s}(s_m) = (1 - G(\hat{\sigma}_m))\frac{\hat{\sigma}_m}{2r}$, where $\hat{\sigma}_m$ is given by equation (15).*

Under Assumption 2, we can obtain closed-form solutions for various distribution of types:

- If types are exponentially distributed, the monopolist chooses $s_m = 1 - \sqrt{2rc(r+\gamma)e/\nu}$.
- If the types are uniformly distributed, $s_m = 1 - \sqrt{8rc(r+\gamma)/\overline{\sigma}}$.

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

[33] This is just for simplicity. Recall that when $\bar{a} = 1/2$, price protection does not affect the venues' profit functions. See Appendix C for the general formula.

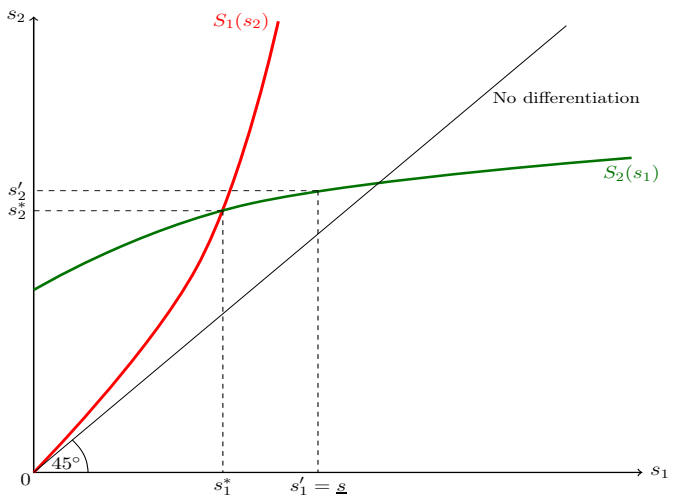

Figure 6. Regulation of speed and venue differentiation. The function $S_i(s_j)$ denotes venue i's best speed response to s_i; s_1^* and s_2^* are the (unregulated) equilibrium speed choices, $\underline{s}$ represents the minimum speed that the regulator may want to impose, and s_1' and s_2' are the optimal speed choices when $\underline{s} > s_1^*$.

Duopoly. In a duopoly, venues have an incentive to offer different speeds to reduce price competition. Recall that the revenue functions for venues 1 and 2 can be expressed as $\pi_1 = q_1 \left(G\left(\hat{\sigma}_2\right) - G\left(\hat{\sigma}_1\right)\right)$ and $\pi_2 = q_2 \left(1 - G\left(\hat{\sigma}_2\right)\right)$, and that fees are given by $q_1 = \frac{1}{2r} s_1 \hat{\sigma}_1$ and $q_2 = \frac{1}{2r}\left(\hat{\sigma}_2 \left(s_2 - s_1\right) + \hat{\sigma}_1 s_1\right)$, and the affiliation equilibrium is given in Lemma 6. Venues 1 and 2 then simultaneously solve $\max_{s_i} \Pi_i(s_i, s_j) = \pi_i\left(s_i, s_j\right) - C\left(s_i\right)$. The following optimality conditions are straightforward to derive.[34]

Lemma 9. *Speed choices* (s_1, s_2) *satisfy*

$$\frac{\partial q_2}{\partial s_2}\left(1 - G\left(\hat{\sigma}_2\right)\right) - \frac{\partial \hat{\sigma}_2}{\partial s_2} g\left(\hat{\sigma}_2\right) q_2 = \frac{\partial C}{\partial s}\left(s_2\right), \tag{22}$$

$$\left(G(\hat{\sigma}_2) - G(\hat{\sigma}_1)\right) \frac{\partial q_1}{\partial s_1} + q_1 \left(g(\hat{\sigma}_2)\frac{\partial \hat{\sigma}_2}{\partial s_1} - g(\hat{\sigma}_1)\frac{\partial \hat{\sigma}_1}{\partial s_1}\right) = \frac{\partial C}{\partial s}\left(s_1\right). \tag{23}$$

The solution to the system of equations (22) and (23) implicitly characterizes a function $S_i\left(s_j\right)$ that represents the best response of venue i to s_j.

Figure 6 displays the speed choices. The 45-degree line represents the case in which there is no product differentiation, which would lead to Bertrand competition and would be inconsistent with entry by both venues for any arbitrarily small entry cost. The actual equilibrium satisfying equations (22) and (23) is at point (s_1^*, s_2^*) where the best response functions intersect. In this equilibrium, there is a fast venue and a slow venue.

Let us now compare the market equilibria under monopoly and duopoly. There is a fundamental tension between profitability and elasticity. On the one hand, the marginal return to speed

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

depends on $\hat{\sigma}(1-G(\hat{\sigma}))$ for both the monopolist and the fast venue. The monopoly chooses $\hat{\sigma}_m$ to maximize precisely this quantity; therefore, we know that $\hat{\sigma}_m(1-G(\hat{\sigma}_m)) > \hat{\sigma}_2(1-G(\hat{\sigma}_2))$. This profitability effect makes the monopolist more willing to invest in speed. On the other hand, competing venues have an incentive to differentiate their services. As s_2 increases, competition is relaxed, q_1 increases, and $\hat{\sigma}_2$ decreases: $\frac{\partial \hat{\sigma}_2}{\partial s_2} < 0$. This leads to greater participation and higher profits for venue 2. Which effect dominates depends on the distribution of types. With a uniform distribution of types, we are able to show in Appendix B that the second effect dominates and, therefore, that the equilibrium speed is higher under a duopoly.

Proposition 4. *When types are uniformly distributed, the fast venue of a duopoly chooses a higher speed than a monopoly does: $s_2 \geq s_m$.*

6.2 Optimal Regulation of Speed

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

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

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

Proposition 5. *When types are uniformly distributed, it is optimal for the regulator to mandate a minimum speed but not a maximum speed, that is, $\underline{s} > s_1$ but $\overline{s} = 1$.*

Consider the monopoly first. The speed chosen by the monopolist is as in Proposition 3. The regulator seeks to maximize social welfare and thus solves $\max_s \frac{s}{2r}\int_{\hat{\sigma}_m}^{\bar{\sigma}} \sigma dG(\sigma) - C(s)$. In this constrained solution the regulator takes $\hat{\sigma}_m$ as a given. Since $\int_{\hat{\sigma}_m}^{\bar{\sigma}} \sigma dG(\sigma) > \hat{\sigma}_m(1-G(\hat{\sigma}_m))$, $s_m^* > s_m$. This is a standard result and it holds for any distribution of types. The planner prefers a higher speed than the monopoly because the planner values the welfare gain for the infra-marginal types ($\sigma > \hat{\sigma}_m$) while the monopoly does not.

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

$$2r\frac{\partial C}{\partial s}(s_2^*) = \int_{\hat{\sigma}_2}^{\bar{\sigma}} \sigma dG(\sigma) - (s_2 - s_1)\hat{\sigma}_2 g(\hat{\sigma}_2)\frac{\partial \hat{\sigma}_2}{\partial s_2} - s_1\hat{\sigma}_1 g(\hat{\sigma}_1)\frac{\partial \hat{\sigma}_1}{\partial s_2}.$$

The term $\int_{\hat{\sigma}_2}^{\bar{\sigma}} \sigma dG(\sigma)$ is the surplus of the high-σ types that the fast venue does not appropriate and therefore does not internalize. Allocation efficiency for types $\sigma > \hat{\sigma}_2$ calls for higher speed. On the other hand, $\frac{\partial \hat{\sigma}_2}{\partial s_2}$ and $\frac{\partial \hat{\sigma}_1}{\partial s_2}$ capture the impact of s_2 on differentiation which softens competition. The link between social welfare and speed depends on the tradeoff between participation and trading efficiency for the high-σ types. We show in Appendix B that the trading efficiency effect dominates

TABLE II
VENUES ENTRY PAYOFFS AND TRADING REGULATION $\intercal \in \{seg; prot\}$

Venue 1 $\downarrow$ and 2 $\rightarrow$	In	Out
In	$\pi_1^{\intercal} - \kappa, \pi_2^{\intercal} - \kappa$	$\pi_1^m - \kappa, 0$
Out	$0, \pi_2^m - \kappa$	$0, 0$

when the types are uniformly distributed. This is particularly interesting, since we have shown in Proposition 4 that the fast venue of a duopoly chooses a higher speed than a monopoly does. Proposition 5 states that this is not enough and the regulator would like an even higher speed. Therefore, the regulator does not find it optimal to impose an upper limit on speed.

On the other hand, it is optimal for the regulator to impose a minimum speed requirement that is higher than that chosen by the slow venue. The intuition is that such a minimum speed increases the welfare of the low-σ types and also increases competition with the fast venue. The equilibrium with a minimum speed requirement is represented by $\left(\underline{s}_1, s_2^{'}\right)$ in Figure 6.[35] We provide calibrated welfare enhancement estimates in Section 8 and further discuss related regulations in Section 9.

7 Entry

This section completes the description of the equilibrium market structure by analyzing venue entry decisions.

7.1 Price Protection and Entry

Let us first study the impact of trading regulations on entry for exogenous speeds. There are two potential entrants, with exogenous speeds s_1 and s_2, respectively, with the convention that $s_1 < s_2$ and we assume for simplicity that $C(s) = 0$ (see below for a discussion of entry with endogenous speeds). We consider first the case where the entry cost κ is the same for both venues. Venue i's net profit is then given by $\pi_i^{\intercal} - \kappa$, where $\intercal \in \{seg; prot\}$ denotes trading regulations.[36] For a given speed, asset supply $\overline{a} \leq 1/2$, and regulatory framework, the profit functions π are as in Section 5. A given venue i finds it optimal to enter whenever net profits are non-negative.

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

[36] 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 involves Brazil, where the incumbent exchange BM&F Bovespa also controls the single national clearinghouse. By denying clearing access to entrants, the incumbent forces new competitors to develop their own clearinghouses.

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

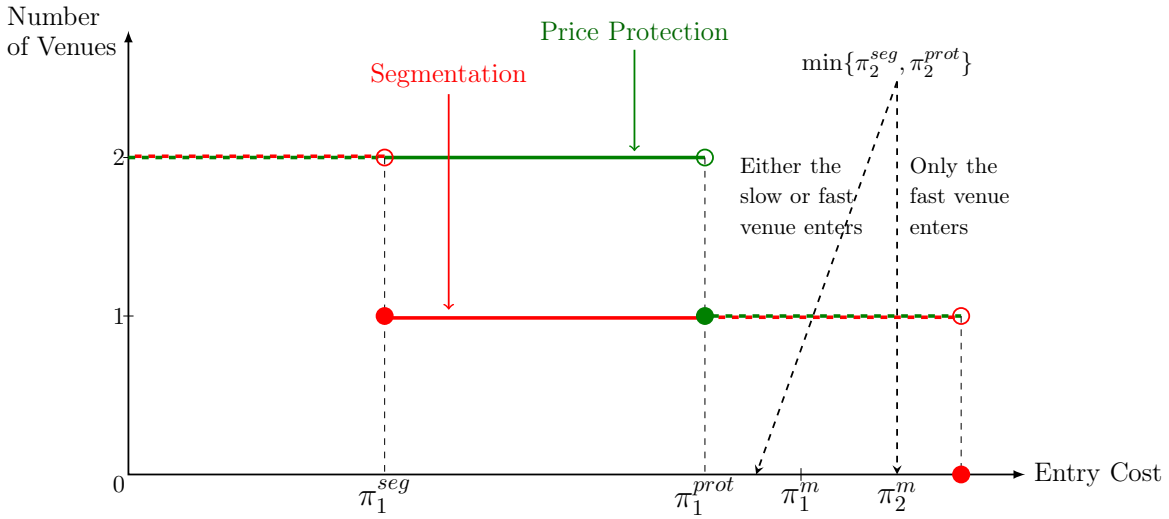

Figure 7. Entry cost, trading regulation, and equilibrium fragmentation. The graph shows the equilibrium number of venues as a function of entry costs κ. Price protection affects the equilibrium number of venues that enter the market when entry costs are between the expected profits of the slow venue under segmentation, π_1^{seg}, and under price protection, π_1^{prot}.

Proposition 6. *Suppose that the distribution of types is exponential. Then price protection at the trading stage helps sustain entry at the initial stage.*[37]

As shown in Figure 7, price protection expands the ex ante number of venues for economies with intermediate entry costs (between π_1^{seg} and π_1^{prot}). The expected level of fragmentation therefore 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 venue enters whenever $\kappa > \pi_1^{prot}$. We characterize the cases with multiple equilibria in the proof of Proposition 6 in Appendix B.

We have also considered the case of heterogenous entry costs, i.e., $\kappa_1 < \kappa_2$. This is a natural setup which captures the choice between a basic package (low entry cost, but low speed) and an advanced package (high speed but high entry cost). Heterogenous entry costs can change our analysis because venue 2 can be the marginal one in terms of entry decisions. Take the limit case where κ_1 goes to zero. Then venue 1 always enters and we only need to worry about venue 2. What can we say, then, about the consequences of price protection? It clearly depends on the comparison between π_2^{seg} and π_2^{prot}. If price protection lowers π_2, it could discourage entry by the fast venue. As

[37] We prove this result analytically for an exponential distribution and numerically for other distributions.

we have explained in the discussion of Proposition 2, the impact of protection on π_2 is ambiguous. On the one hand, protection makes venue 1 more attractive, but on the other hand it induces venue 1 to raise its price. For the parameters derived from our calibration, however, we find a small *increase* in π_2 under protection, i.e., we find $\pi_2^{prot} > \pi_2^{seg}$. For these parameters, we can therefore say that protection increases entry irrespective of the distribution of entry costs.[38] Finally, we have also extended Proposition 6 to the case in which speed choices are endogenous, but this can only be done numerically.

7.2 Regulation of Entry

Let us now consider the problem of a regulator who can decide the number of venues but nothing else. The regulator takes into account that entry affects speed choices, venues' fees, investor participation, and therefore welfare. We study when the regulator wants to encourage or restrict entry.

The traditional argument to restrict entry relies on the existence of thick market externalities, as for Pagano (1989). When externalities are strong enough, the welfare gains from increased competition are not enough to compensate for the loss of matching efficiency between buyers and sellers. As discussed in the introduction, this type of externality is likely to be less relevant in today's markets, which is why it is not included in our model. Absent thick market externalities, the source of excess entry can be related to cost duplication and speed choices. Mankiw and Whinston (1986) identify three general conditions under which excess entry occurs: (i) some form of economies of scale due to fixed costs, (ii) post-entry prices exceeding marginal cost, and (iii) enough business stealing to decrease average firm output. The first two conditions are easily verified in our environment. The third is not, however, since trading services are vertically differentiated. This is a fundamental difference between our model and the Hotelling models of Spence (1976) and Mankiw and Whinston (1986).

The welfare functions under monopoly and duopoly are given in Lemma 8. The welfare gain of moving from a monopoly to a duopoly is therefore

$$\Delta\mathcal{W}_{1\to 2} = \frac{s_1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_2} \sigma dG(\sigma) - C(s_1) + \frac{s_2}{2r}\int_{\hat{\sigma}_2}^{\overline{\sigma}} \sigma dG(\sigma) - C(s_2) - \frac{s_m}{2r}\int_{\hat{\sigma}_m}^{\overline{\sigma}} \sigma dG(\sigma) + C(s_m) - \kappa.$$

On the other hand, entry is profitable for the slow venue if and only if $\pi_1 > \kappa$, that is, if $\frac{s_1\hat{\sigma}_1}{2r}\left(G(\hat{\sigma}_2) - G(\hat{\sigma}_1)\right) > \kappa + C(s_1)$. Excess entry thus occurs if and only if $\pi_1 > \kappa$ and $\Delta\mathcal{W}_{1\to 2} < 0$ hold simultaneously. We can then obtain the following result.

Proposition 7. *For any fixed cost κ, entry of a second venue is never excessive, as long as the cost of speed is low enough.*

[38] On the other hand, Appendix B gives an example where $\pi_2^{prot} < \pi_2^{seg}$. This can happen when $\frac{s_1}{s_2 - s_1}$ is close to zero, i.e., s_1 is very small relative to s_2. In that case there is not much fee competition to start with, so the positive impact on π_2 of the increase in q_1 is limited, and the overall effect of protection is to lower π_2. Notice, however, that this only happens in the case where competition does not matter much, and venue 2 is close to a monopoly in any case. The latter is not a case where we would normally worry about entry by venue 2.

Proposition 7 applies to the transition from a monopoly to a duopoly, which is unlikely to yield excessive entry since monopoly distortions are typically large. The same might not be true when there are already *several* venues and we consider an additional entrant. An interesting analysis of entry therefore requires us to consider more than two venues, which is what we do in the next section.

7.3 General Oligopoly

Affiliation Game. Let I denote the number of active venues and, consistent with our previous notation, let $\hat{\sigma}_i$ be the lowest type that joins venue i. This marginal type is indifferent between venues i and $i-1$; therefore $\hat{\sigma}_i = \frac{r}{\overline{a}} \frac{q_i - q_{i-1}}{s_i - s_{i-1}}$. By repeated substitutions, it is then easy to show that we must have $q_i = \frac{\overline{a}}{r} \sum_{j=1}^{i} \hat{\sigma}_j (s_j - s_{j-1})$, where $s_0 \equiv 0$. Defining $\hat{\sigma}_{I+1} \equiv \overline{\sigma}$, we can write the revenues of any venue $i \in \{1, ..., I\}$ as $\pi_i = q_i \left(G(\hat{\sigma}_{i+1}) - G(\hat{\sigma}_i) \right)$. We consider the affiliation game where venues compete in fees to attract investors. Taking first-order conditions with respect to q_i, we obtain the following proposition.

Proposition 8. ***Equilibrium of the Affiliation Game with I Active Venues.*** *For all for all $i \in \{1, ..., I\}$, the set of marginal participating types $\{\hat{\sigma}_i\}$ satisfies*

$$G(\hat{\sigma}_{i+1}) - G(\hat{\sigma}_i) = \left(\frac{g(\hat{\sigma}_i)}{s_i - s_{i-1}} + \frac{g(\hat{\sigma}_{i+1})}{s_{i+1} - s_i} \right) \sum_{j=1}^{i} \hat{\sigma}_j (s_j - s_{j-1}).$$

Proposition 8 generalizes the results with two venues described in Lemma 6.[39] It allows us to compute the equilibrium of the affiliation game for an arbitrary number of venues.

Entry. We can now study entry with more than two venues. As explained above, moving from one to two venues is likely to be socially efficient since it introduces competition in a market where there is none. It is less clear whether moving from I to $I+1$ is socially efficient when I is already above one.

We analyze the unexpected entry of a third venue in an existing duopoly. The incumbents have already chosen their speeds and paid their fixed entry costs, expecting to be in a duopoly. We then ask if the entry of a third venue would raise welfare. The third venue chooses its speed optimally, given the speeds of the existing duopoly. The considered approach allows us to bypass the issue of entry deterrence,[40] which is beyond the scope of our paper. The model can only be solved numerically and we use the baseline parameters described in the next section. We highlight an example where entry can be inefficient

[39] The proof is a generalization of our result in Lemma 6 and is thus omitted. Generalization to the case of price protection is also straightforward but is not our focus here.

[40] There is, of course, a large literature that studies entry deterrence. For instance, Donnenfeld and Weber (1995) consider a vertically differentiated duopoly facing the threat of entry by a third firm. They show that incumbent firms can deter entry by choosing quality levels that reduce ex post differentiation relative to an unchallenged duopoly. Therefore, entry deterrence may improve welfare, even without actual entry.

Lemma 10. *Entry can reduce welfare when the speed of the new entrant is lower than the speed of the slow incumbent.*

Proof. The proof is numerical. See Appendix I. □

Entry always increases total participation but can also lead to misallocations and this effect can be large if entry takes place at the low end of the range of incumbents' speeds. Let us explain the intuition for this result. To keep our notation simple, we denote the low- and high-speed venues of the existing duopoly by (l, h) and the new entrant by e. In a duopoly, we have the mapping $(l, h) \to (1, 2)$. When we add venue e, the new ordering depends on the relative speed of the entrant. As explained above, here we consider the case in which $s_e < s_l$. In the oligopoly with three venues, the ranking in terms of Proposition 8 is therefore $(e, l, h) \to (1, 2, 3)$. It is clear that entry creates direct competition for venue l. Therefore, venue l is forced to lower its price. The important point is to consider the reaction of venue h. Venue h does not compete directly with venue e, but it competes with l. Venue h reacts to the induced drop in q_l. The optimal pricing condition for venue h is

$$1 - G\left(\hat{\sigma}_h\right) = g\left(\hat{\sigma}_h\right)\left(\hat{\sigma}_h + 2r\frac{s_l}{s_h - s_l}q_l\right), \tag{24}$$

where $\hat{\sigma}_h$ is the marginal type for the fast venue.

Under Assumption 1, the function $\frac{1-G(\sigma)}{g(\sigma)} - \sigma$ is decreasing in σ.[41] Therefore equation (24) implies that $\hat{\sigma}_h$ is a decreasing function of q_l. This result explains the potential inefficiency. Entry by the third venue forces the middle venue to lower its fee, but it is not profitable for the fast venue to fully accommodate this fee change. Therefore, $q_h - q_l$ increases and $\hat{\sigma}_h$ goes up: Investors who used to trade in the fast venue now trade in the middle venue. This is clearly a misallocation since the planner would rather have more investors in the fast venue. Naturally, the private surplus increases for investors who move from h to l venue, but the profit losses of the fast venue are even greater. The analysis in Appendix I shows that, on the other hand, allowing for a third venue can lead to significant welfare gains when entry takes place at the high end of the speed ladder. This case of analysis could capture the (unobserved) welfare consequences of having new venues, such as BATS or Direct Edge, enter and challenge the incumbents (NYSE and NASDAQ) after Reg NMS. The new venues entered with, arguably, better technologies than the incumbents had.We view these results as a first step in the analysis of dynamic entry in the context of financial intermediation.

8 Calibration and Discussion of Regulations

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

[41] For instance, with a uniform distribution, $\frac{1-G(\sigma)}{g(\sigma)} - \sigma = \bar{\sigma} - 2\sigma$.

8.1 Calibration

The baseline parameters are displayed in Table III. Unless otherwise noted, these parameters are held constant across our experiments. We assume a uniform distribution of investor types and we use the functional form in Assumption 2 for the cost function. We set the fixed cost κ to zero and the asset supply to $1/2$ when we do not analyze price protection.[42] The rate r is based on the composite rate of long-term U.S. Treasury securities from January 2007 to December 2013.[43] We set the asset holding cashflow, μ, to match the average S&P500 dividend yield over the same time period (2.13%).[44] Trading days last 6.5 hours, as in U.S. equity markets, and we set the upper bound of the σ-type distribution to $\mu/2$.[45]

We have so far assumed a unit mass of investors. To compare the model-implied per-capita volume in equation (8) with the volume in the data, we thus need to specify the number of investors, N, as an additional parameter. We start from the number of institutional funds in the U.S. as a proxy of the size of the buy side of financial markets. According to Morningstar, at the end of 2007 there were 629 exchange-traded funds, 17,500 mutual funds, and 10,096 hedge funds. These 28,225 funds represent the set of potential investors. This is an upper bound on the number of investors. The most difficult part of the calibration is to determine how many of these investors are active in each market.

Table III
Baseline Parameter Values

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

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

Volume-Implied γ. Let k denote an asset class (bonds, stocks, futures). A critical parameter for our calibration is N_k, defined as the potential number of traders for a typical asset in class k. Participation would be N_k in the first best allocation, $N_k/2$ with a monopoly, and an in-between

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

[43] This rate is the unweighted average of bid yields on all outstanding fixed-coupon bonds neither due nor callable in less than 10 years. Using instead the 10-year T-bond yield yields virtually the same results for the considered period.

[44] With $\mu = 2.75$ and $r = 0.0375$, relative to the Walrasian price, the dividend yield is $\frac{\mu}{p_w} \approx 2.13\%$. The value of μ affects the asset price in the model. Real asset prices obviously also reflect market risk exposure, among other factors, which is not the focus of this paper. This parameter also affects global welfare. However, our analysis focuses on the fraction of welfare that is earned in excess of the autarchy value, that is, $W(\sigma, \cdot) - W_{out}$, as in equation (12). This is the main reason why we do not calibrate μ separately for each asset class (and that we abstract from the fact that the futures contract yields no cashflow).

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

value with a duopoly. Once we calibrate N_k, we can use the observed volume $\mathcal{V}_k$ and speed ρ_k to back out the rate of preference shocks γ_k. For instance, with a single venue, the model-implied transaction rate for a *typical* asset in class k is

$$\mathcal{V}_k = N_k \frac{\gamma_k}{4} \left[\frac{\rho_k}{\gamma_k + \rho_k} \left(1 - G\left(\hat{\sigma}_m\right)\right) \right]. \tag{25}$$

Using a uniform distribution of types, one then obtains $\gamma_k = \frac{8\mathcal{V}_k\rho_k}{N_k\rho_k - 8\mathcal{V}_k}$. The case of multiple venues is more complicated but the principle is the same, as explained in Appendix H. We select different values for N_k in the next section. Equation (25) implies that, for a given observed volume, a high N_k is equivalent to a low γ_k.

Speed-Implied c. We compute the implicit cost parameter c that rationalizes ρ as an optimal speed, given all the other parameters. Inverting the first order condition in Proposition 3 and assuming a uniform distribution of types, we obtain

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

For the duopoly, we use the first-order condition of the fast venue, as explained in Appendix H. Once all model parameters are set, we solve the duopoly program numerically to obtain the sub-game perfect equilibrium values of critical types and speeds $\{\rho_1, \rho_2, \hat{\sigma}_1, \hat{\sigma}_2\}$. The numerical results for investor participation, volume, and welfare are presented in Section 8.3.

8.2 Stylized and Predicted Parameter Values by Asset Class

Table IV presents our calibration of three asset classes: corporate bonds, equities, and S&P500 index futures. Our calibration is consistent with the common wisdom about the relative efficiency of these three markets and, in particular, we have $\left(\frac{\rho}{\gamma}\right)_{Futures} > \left(\frac{\rho}{\gamma}\right)_{Stocks} > \left(\frac{\rho}{\gamma}\right)_{Bonds}$. We test the sensitivity of our results to values of N_k that are one-third lower or higher than our benchmark.

S&P500 Index Futures. The Chicago Mercantile Exchange (CME) has a monopoly over its E-mini futures contracts and we calibrate the corresponding parameters using the monopoly formulas in Section 8.1. There is one contract and, as a benchmark, we consider that most investors are active in this market, so $N_{Emini} = 3/4N$. We use the average number of daily trades on May 6, 2010, as reported by Kirilenko, Kyle, Samadi, and Tuzun (2015).[46] The stylized speed considered is equivalent to an average delay of 200ms. This allows for round-trip communication at near light speed to any location within the U.S.[47] The implied γ means that there are 390 shocks per trading

[46] This date corresponds to the so-called flash crash and displays both a large volume and a large number of investors trading. Using instead the reported values for May 3 to May 5, 2010, yields a lower value for γ. In our calibration, participation in the monopoly then equals $N_{Emini}/2 \approx 10,584$. Kirilenko et al. (2015) report the number of daily active traders to be between 11,875 and 15,422 in their CME sample.

[47] With a normalized trading day of 6.5 hours, there are 23,400 seconds in a trading day. Thus, a contact rate equal to five times this value, 117,000, is equivalent to an average contact delay of 200ms.

day and per investor, or approximately one shock every minute. When $\pm 1/3 N_{Emini}$, there is one shock every 40 second or 80 seconds instead.

We choose c to match the contact rate. Using this value, we estimate that, if the market were a duopoly, our model would imply trading speeds for the slow and fast venues such that average delays would be 45 seconds and 185ms, respectively.[48]

TABLE IV
STYLIZED, IMPLIED, AND PREDICTED PARAMETER VALUES

Panel I: Stylized values									
	Corporate Bonds			**Equities**			**S&P500 Futures**		
Volume			1.97			3,023.4			1,030,204
Number of assets			21,723			2,805			1
Stylized ρ_m			-			-			117,000
Stylized ρ_1			1			195			-
Stylized ρ_2			39			23,400			-
Panel II: Implied and Predicted Values									
	Corporate Bonds			**Equities**			**S&P500 Futures**		
N_k	8	13	17	61	92	115	14,113	21,169	28,225
Implied γ	1.378	0.834	0.588	299.72	182.95	139.64	586.93	390.63	292.73
Implied c ($\times 10^{-3}$)	36.444	36.201	36.415	0.1605	0.1570	0.1564	2.7457	2.7503	2.7526
Predicted ρ_m	37.377	36.211	35.220	22,616	21,986	21,551	117,000	117,000	117,000
Predicted ρ_1	1.644	1.044	0.750	386.50	239.13	183.38	773.57	516.93	388.13
Predicted ρ_2	40.367	38.132	38.065	24,442	23,758	23,286	126,414	126,402	126,396

All rates are daily. For corporate bonds, the benchmark calibration is a duopoly where the low speed is $\rho_1 = 1$, the high speed is $\rho_2 = 39$, and there are 13 active traders per bond. The implied γ means that an institutional investor in this market is subject to 0.834 preference shock per trading day on average. For equity, there are 183 shocks per trading day, or one shock every 128 seconds.

Corporate Bonds. All trades for 2013Q4 are collected from the Trade Reporting and Compliance Engine (TRACE) data set. The average daily number of trades for each of these bonds is 1.97, reflecting the fact that most corporate bonds trade infrequently. Our sample contains 21,723 bonds. The non-transparent nature of the corporate bond market makes it difficult to estimate the participants. According to the *Investment Company Fact Book*, out of 8,000 mutual funds surveyed in 2007, about 800 are bond funds, so we assume that 10% of investors are active in corporate bonds. As a starting point, we assume that each investor is active in 100 individual bonds, which is consistent with anecdotal evidence. This gives us $N_{Bonds} = 0.1 * 100 * 28,225/21,723 = 13$. We perform robustness checks with $N_{Bonds} = 8$ and $N_{Bonds} = 17$, as explained above. We calibrate the corporate bond market as a duopoly. Corporate bonds trade in traditional phone-based OTC broker networks, stylized here as the first (slow) venue, or in modern electronic platforms, the second (fast) venue. The stylized contact rates are one and 39, for the first and second venues. The first

[48] One may interpret the competing slow venue here as a broker–dealer firm offering an asset with identical features, or a traditional trading pit. In this regard, a delay of 45 seconds is consistent with human intervention.

is equivalent to an average trading delay of a day, a value that captures the delay in a traditional voice-based OTC network. The latter value represents an average delay of 10 minutes, closer that for an electronic platform based on the request for quote (RFQ) protocol. The implied γ means that there is 0.834 preference shock per trading day per investor in this market.

We choose c to match (approximately) the two contact rates. Interestingly, the model has no difficulty in explaining the wide range of speeds observed in practice, and the predicted speeds are close to the stylized values.

Equities. We calibrate the model to 2007, because that was when Reg NMS was implemented. According to data from the NYSE Group (www.nyxdata.com), the average number of daily trades for a typical NYSE-listed stock in 2007 was equal to $3,023$. The number of listed stocks in 2007 was $2,805$. According to the *Investment Company Fact Book*, 46% of funds are US equity funds. There is no simple way to estimate N_{Stocks} because many equity-related trades are index trades, not trades on individual stocks. We choose the typical number of actively traded stocks to capture the intuitive idea that the stock market lies somewhere in between the bonds market and the futures market in terms of efficiency and number of traders per asset. Assuming that a trader is active in 20 individual stocks, we obtain $N_{Stocks} = 92$. We perform robustness checks with values that are higher and lower, as explained above. We calibrate the equity market as a duopoly, given the prevalence of the NYSE and the NASDAQ at the time of Reg NMS implementation.[49] To calibrate the stylized contact rate parameters, we consider SEC Rule 605 data for the NYSE for 2007, before the full implementation of Reg NMS. The value for the fast venue matches the average execution delay of 1 second in 2007 for small automated orders. The value for the slow venue represents a human broker–dealer round-trip delay of 1 minute and is consistent with the SEC data presented in Figure A2 in the Appendix A. The implied γ means that there is one shock per investor every 128 seconds in this market.[50]

We choose c to match the contact rates and, as in the case of bonds, we find that the model correctly predicts the relative speeds of the two venues.

8.3 Welfare Analysis

Table V shows the main equilibrium outcomes in the benchmark case with $\overline{a} = 1/2$. All the values in the table are relative to the constrained first best, which represents a fictional planner that

[49] According to the SEC (2010), the NYSE executed as much as four-fifths of the volume of NYSE-listed stocks just before Reg NMS.

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

TABLE V
CALIBRATION OUTCOMES (PLANNER CASE =100)

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

Each cell is normalized relative to the constrained planner outcome.

is subject to the same technological frictions and operate under a break-even budget constraint (see the analysis in Appendix G). Panels I to III of Table V, respectively, display the outcomes corresponding to the parameters implied by the low, medium, and high values of N_k, as in Table IV.

Participation. Participation under monopoly and uniform distribution of investors is slightly above one-half that of the planner. Participation increases dramatically to around 88% when two venues compete. We verify numerically that participation in the second venue alone is always greater than in the monopoly case, as predicted by the theory. Participation levels are similar across asset classes because the degree of *relative* differentiation s_2/s_1 is similar across asset classes.[51]

Trading Volume. Even in markets with high speeds, the duopoly fails to realize an important fraction of the potential trades as volume represents about 75% of the planner's total. This fact reflects lack of full participation in the duopoly relative to the Walrasian setting. But, importantly, it also reflects the inefficient allocation of the asset across investors in the first venue due to speed differentiation. Thus, the slow venue volume share is roughly one-half of its relative investor participation for all asset classes.

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

Welfare. Although the monopoly only attracts half of all investors, it achieves a much higher fraction of the maximum attainable welfare. This is chiefly because those investors who choose to participate benefit from high gains from trade (high σ values) and, as reflected in Table IV, the monopolist optimally offers a relatively high trading speed. The calibration suggests high social gains from encouraging entry. Welfare typically increases by at least 15 percentage points when transitioning from one to two venues. Naturally, the gains from trade are disproportionately distributed across the slow and fast venue. For example, in the benchmark case for equities, the slow venue only contributes 10% of the total gains from trade.

The welfare associated with a particular market structure in the model is driven by both technology and market power frictions. To further understand their relative importance in each case, we decompose the welfare gap between the market outcome and the constrained planner. For the monopoly, the decomposition is as follows.

$$\mathcal{W}_P - \mathcal{W}_m = \underbrace{\frac{s_P}{2r}\int_{\hat{\sigma}_P}^{\hat{\sigma}_m} \sigma dG(\sigma)}_{\substack{\text{Participation}\\ \text{Loss}}} + \underbrace{\frac{s_P - s_m}{2r}\int_{\hat{\sigma}_m}^{\bar{\sigma}} \sigma dG(\sigma)}_{\substack{\text{Speed}\\ \text{Loss}}} + \underbrace{C(s_P) - C(s_m)}_{\substack{\text{Speed cost}\\ \text{differential}}} \tag{26}$$

Analogous expressions can be derived for two or more venues. The decomposition is fairly intuitive. Keeping the efficient technology s_P constant, imperfect competition distorts the lowest active type from $\hat{\sigma}_P$ to $\hat{\sigma}_m$. This generates a *limited participation* welfare loss. In turn, for any given level of market participation, the market equilibrium is less efficient in re-allocating the asset from low to high types using the suboptimal technology level $s_m < s_P$. This generates a *trading speed* welfare loss. The total effect of the distortion in technology choices also depends on speeds costs. Furthermore, the participation and speed losses can be computed for the constrained planner relative to Walrasian frictionless benchmark, using the same calibration values, in analogous fashion.[52]

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

Let us now consider speed losses. The misallocation loss for the planner, which reflects the cost of the trading technology, is the only meaningful loss of welfare. Its value ranges from a maximum of 2.53% of the frictionless gains from trade for corporate bonds, to only 0.177% in the case of the large index futures markets (for which c is low). The misallocation loss of the monopoly is fairly modest in value. Interestingly, however, the latter represents the bulk of the welfare losses when there is competition among venues. The speed loss is between four and five times greater than the participation loss even with only two venues. This is due to the fact that the duopoly equilibrium

[52] The planner–Walrasian welfare gap is given by

$$\underbrace{\frac{1}{2r}\int_{\hat{\sigma}_W}^{\hat{\sigma}_P} \sigma dG(\sigma)}_{\text{Participation loss}} + \underbrace{\frac{1 - s_P}{2r}\int_{\hat{\sigma}_P}^{\bar{\sigma}} \sigma dG(\sigma)}_{\text{Speed loss}} - \underbrace{C(s_P)}_{\text{Speed cost differential}}$$

TABLE VI
SOURCES OF WELFARE LOSS

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

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

forces venues to decrease access fees significantly, allowing near-efficient total participation, but it allocates nearly one third of the active investors to the slow speed venue (see Table V). Note that the speed cost differential has different signs for the planner and market outcomes. The positive sign for the planner is simply a consequence of considering a Walrasian benchmark for which costs are zero. The negative sign in the case of the monopoly is a consequence of $s_m < s_P$ and for the duopoly is a feature of the calibration outcomes displaying $\sum_{i=1,2} C(s_i) < C(s_P)$.

8.4 Is Price Protection Socially Desirable?

Trading regulations $\intercal \in \{seg, prot\}$ affect not only trading prices but competition among venues and, ultimately, welfare. In particular, the welfare consequences depend on how the regulation affects venue entry and investor affiliation. There are three different cases to consider, but two have already been analyzed. When entry costs κ are larger than π_1^{prot}, only one venue can enter and trading regulations are irrelevant. When $\pi_1^{seg} < \kappa \leq \pi_1^{prot}$, price protection increases the number of venues (see Proposition 6) and we have seen in Section 8.3 that this has a large positive effect on welfare. Our goal in this section is to quantify the potential consequences of price protection when $\kappa \leq \pi_1^{seg}$ so that it does not affect the entry game, but it distorts competition among venues.

Table VII presents our estimates of the welfare cost of these distortions setting $\overline{a} = 0.45$ and, as in Table V, displaying values that are relative to the constrained efficient allocation. To facilitate the connections with the propositions in Section 5, we keep the same speeds regardless of trading regulations.[53] Participation at time 0 includes all the traders. Over time, the light traders drop

[53] Endogenizing speeds in the first stage has a second-order effect on outcomes relative to the fee distortions in the

TABLE VII
SEGMENTED AND PROTECTED EQUILIBRIA OUTCOMES (PLANNER CASE=100)

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

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

out and at time ∞ only the heavy traders remain. This process, however, is the same in the constrained efficient allocation so, although fewer traders are active in the market solution, the participation ratios reported in the Segmented panel of Table VII do not change.[54] For the bond market, for instance, participation is always 96.3% that of the constrained efficient participation. Price protection, in turn, affects both total participation and the distribution of investors across venues. We observe that $\hat{\sigma}_1^{prot} > \hat{\sigma}_1^{seg}$ as in Proposition 2 and that the relative affiliation to venue 1 increases under protection, but $\hat{\sigma}_2^{prot} < \hat{\sigma}_2^{seg}$ so affiliation to the fast venue decreases. However, as assets migrate from the slow venue to the fast venue, as described in Appendix D, the proportion of active investors in the fast venue increases and in the steady state is nearly 10% higher than in under segmentation. The total drop in investor participation is about 1.45%. The effects of price protection on equilibrium profits of the slow venue are even larger: For the three asset classes considered, we find that its profits increase by more than 6.5%.

Table VII also shows that the welfare impact of price protection (around 1.8%) is low relative to the welfare impact of market structure changes in Table V.[55] The key point is then whether protection affects the number of venues, as in proposition 6. This is an important insight in light of the debate regarding the impact of the SEC's trade-through rule on market quality.[56] According

second stage of the game.

[54] Participation at time $t=0$ represents total investor affiliation (which drives venue revenue). For the protected case is $(2\bar{a}G(\hat{\sigma}_2) - G(\hat{\sigma}_1) + 1 - 2\bar{a})$ for venue 1 and $1 - G(\hat{\sigma}_2)$ for venue 2. At time $t=\infty$, participation in venue 1 is $G(\hat{\sigma}_2) - G(\hat{\sigma}_1)$. Participation at time $t=0$ in the segmented case is $\frac{1}{2\bar{a}}(G(\hat{\sigma}_2) - G(\hat{\sigma}_1))$ and $\frac{1}{2\bar{a}}(1 - G(\hat{\sigma}_2))$ for venues 1 and 2. At time $t=\infty$, participation is $1 - G(\hat{\sigma}_2)$ and $G(\hat{\sigma}_2) - G(\hat{\sigma}_1)$ for venues 1 and 2. These terms represent the same fraction of the constrained planner case for all t. The welfare expressions for the segmented case are given in Lemma 8 and those for the protected case are given in Appendix C.

[55] The social value of protection technically depends on the value of $\overline{a}$ (0.45 here), but welfare differences remain small in any case. Of course, we do not capture welfare gains from mitigating execution price uncertainty, given that our traders are risk neutral.

[56] For example, the SEC has recently requested its Equity Market Structure Advisory Committee to assess the

to our model, price protection had a significant impact on welfare in U.S. markets because it likely encouraged entry. On the other hand, for markets that are already fragmented, as in most of Europe, price protection may not increase the number of venues and thus only bring a modest welfare change. In fact, if as considered in Section 5 and Appendix B, $\hat{\sigma}_2^{prot} > \hat{\sigma}_2^{seg}$, the welfare impact of price protection can be negative.

8.5 Speed and Welfare

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

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

Overall, the positive predictions of the model seem to fit the experience of the past 20 years in terms of speed, entry and fragmentation. The normative analysis, on the other hand, suggests limited welfare gains from purely technological improvements and highlights the importance of regulations.

9 Concluding Remarks

We have provided an equilibrium analysis of entry, investment in speed, and competition among trading venues. Let us briefly summarize our main conclusions.

Normative Results. On the normative side, our model clarifies the circumstances under which competition, fragmentation, and speed improve or reduce welfare.

- *Entry.* we find that it is optimal to challenge monopolies and that welfare losses are still significant in a duopoly. In addition, we find that entry by a fast venue is likely to increase

effectiveness of the trade-through rule (Rule 611 of SEC Regulation NMS).

TABLE VIII

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

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

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

welfare, while entry by a slow venue might not. These results are relevant for several regulatory initiatives around the world, such as MiFID, Reg ATS, and Reg NMS.

- *Speed.* We find that, barring front-running issues, it is not optimal to limit venue speed. This does not mean, however, that there is much to expect from purely technological improvements in trading speeds. Perhaps one of the most striking results of our quantitative analysis is that a reduction is the cost of speed leads to a vast increase in speed by the fast venue, almost no increase by the slow venue, and, as a result, very limited welfare gains. On the other hand, we find that it can be optimal to increase both speed and competition by pushing the slow venues to upgrade their technology. Slow and inefficient markets, such as that for corporate bonds, could benefit from such a rule. This finding is also consistent with the effect that the automation mandate of Reg NMS had on the NYSE (see Figure A3 in Appendix A), and the likely effect of recent regulatory efforts, such as Dodd–Frank and European Market Infrastructure Regulation/MiFID II, that push for more electronic trading for OTC derivatives.

- *Price protection.* We find it to be efficient when it encourages entry. The effect is ambiguous though when the active number of venues is not affected, as one would expect in markets that are highly fragmented ex ante. These findings are relevant for Reg NMS in the U.S. and the order protection rule in Canada, and informative for similar regulations elsewhere.

The empirical implementation of the model in Section 8 also permits a quantitative assessment of such diverse policy interventions.

Positive Results. On the positive side, our model provides an explanation for the joint evolution of trading regulations, fragmentation, and speed. The recent sharp increase in market fragmentation in developed countries has encouraged a new wave of empirical studies whose results appear to be consistent with the predictions of our model. Foucault and Menkveld (2008), O'Hara and Ye (2011), and Degryse, De Jong, and van Kervel (2014), among others, find that an increase in trading fragmentation is associated with lower costs and faster execution speeds in a given asset class. Our result that price protection increases entry and thus fragmentation helps to rationalize (i) the sharp increase in fragmentation experienced in U.S. equity markets after 2007 and (ii) the fact that fragmentation levels since Reg NMS are the highest in the world. One advantage of our model is that we can estimate the welfare consequences of these evolutions.

It has been argued that moving away from continuous trading toward periodic auctions would eliminate the speed investment frenzy Budish, Cramton, and Shim (e.g., 2016). Our results suggest that these reforms could mitigate but are unlikely to stop this phenomenon. As the cost of speed decreases, trading speeds increase and become more differentiated. This is likely to encourage entry by fast venues, further increasing the market average speed.

Let us conclude with some caveats and ideas for future work. One limitation of our analysis is that it misses important sources of differentiation among exchanges, such as between lit and dark trading venues. One could also consider more sophisticated trading protocols to better describe

particular asset classes. The framework we have developed can, however, be generalized in such directions. Another interesting extension would be to endogenize the contracts between trading venues and liquidity providers, such as high-frequency firms. By introducing such features, one could study the performance of incentive schemes for liquidity creation and stability, a key concern in the debate over designated market makers in today's electronic markets.

Another limitation of our analysis is that we do not take into account asymmetric information. One can argue that the desire to take advantage of information is one reason behind the observed increase in speed. It should be noted, however, that the joint increase in average speed, differentiation, and fragmentation is not a new phenomenon, that it has been observed in virtually every asset class, and that even investors who are not interested in any sort of front running still decide to trade in fast exchanges, such as IEX, or fast venues with RFQ protocols. Current models of front running in a single venue can explain why some traders have an incentive to become *individually* faster, but they do not account for the relative participation of investors across trading venues. Nothing prevents the formation of a relatively slow and cheap venue. If uninformed traders choose to join fast venues, they must value speed; otherwise, they would all join the slow venue, depriving the fast venue of liquidity. The idea that speed is provided exclusively to satisfy a fraction of informed traders seems to be inconsistent with free entry.

Our model, on the other hand, captures a *fundamental part of the demand for speed* that would be present in any model, with or without front running or other information-based demand for speed. We argue that speed-sensitive gains from trade are required to explain investment in speed and we therefore think that information-based models are a complement, rather than a substitute, to the model that we have presented. Some participants may use speed to take advantage of other investors and 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

Ait-Sahalia, Y., Y. Aït-sahalia, and M. Saglam (2013): "High Frequency Traders: Taking Advantage of Speed," *mimeo*, October, 1–52.

Angel, J. J., L. E. Harris, and C. S. Spatt (2011): "Equity Trading in the 21st Century," *Quarterly Journal of Finance*, 01, 1–53.

Baruch, S., A. G. Karolyi, and M. L. Lemmon (2007): "Multimarket Trading and Liquidity: Theory and Evidence," *Journal of Finance*, LXII, 2169–2200.

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

Biais, B., T. Foucault, and S. Moinas (2015): "Equilibrium Fast Trading," *Journal of Financial Economics*, forthcomin.

Biais, B., J. Hombert, and P.-O. Weill (2014): "Equilibrium Pricing and Trading Volume under Preference Uncertainty," *The Review of Economic Studies*, 81, 1401–1437.

Boehmer, E. (2005): "Dimensions of execution quality : Recent evidence for US equity markets," *Journal of Financial Economics*, 78, 553–582.

Budish, E., P. Cramton, and J. Shim (2016): "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *Quarterly Journal of Economics*.

Champsaur, P. and J.-c. Rochet (1989): "Multiproduct Duopolists," *Econometrica*, 57, 533–557.

Chao, Y., C. Yao, and M. Ye (2016): "What Drives Dispersion and Market Fragmentation Across U.S. Stock Exchanges?" *mimeo*, 1–60.

Chowdhry, B. and V. Nanda (1991): "Multimarket Trading and Market Liquidity," *Review of Financial Studies*.

Colliard, J.-E. and T. Foucault (2012): "Trading Fees and Efficiency in Limit Order Markets," *Review of Financial Studies*, 25, 3389–3421.

Degryse, H., F. De Jong, and V. van Kervel (2014): "The Impact of Dark Trading and Visible Fragmentation on Market Quality," *Review of Finance*, 1–36.

Donnenfeld, S. and S. Weber (1995): "Limit qualities and entry deterrence," *Journal of Economics*, 26, 113–130.

Duffie, D., N. Garleanu, and L. H. Pedersen (2005): "Over-the-counter markets," *Econometrica*, 73, 1815–1847.

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

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

Foucault, T., J. Hombert, and I. Rosu (2016): "News Trading and Speed," *Journal of Finance*, 71, 335–382.

Foucault, T. and A. J. Menkveld (2008): "Competition for Order Flow and Smart Order Routing Systems," *Journal of Finance*, LXIII, 119–158.

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

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

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

Gârleanu, N. (2009): "Portfolio choice and pricing in illiquid markets," *Journal of Economic Theory*, 144, 532–564.

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

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

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

Kirilenko, A., A. S. Kyle, M. Samadi, and T. Tuzun (2015): "The flash crash: The impact of high frequency trading on an electronic market," *Journal of Finance*.

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

Lagos, R., G. Rocheteau, and R. Wright (2015): "Liquidity : A New Monetarist Perspective," *mimeo*.

Lewis, M. (2014): *Flash Boys: A Wall Street Revolt*, New York: Allen Lane.

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

Mankiw, N. G. and M. D. Whinston (1986): "Free Entry and Social Inefficiency," *The RAND Journal of Economics*, 17, 48–58.

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

Mussa, M. and S. Rosen (1978): "Monopoly and Product Quality," *Journal of Economic Theory*, 18, 301–317.

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

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

Pagnotta, E. S. (2015): "Speed, Fragmentation, and Asset Prices," *mimeo*.

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

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

Ronnen, U. (1991): "Minimum Quality Standards, Fixed Costs, and Competition," *The RAND Journal of Economics*, 22, 490–504.

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

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

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

Shaked, A. and J. Sutton (1982): "Relaxing Price Competition Through Product Differentiation," *Review of Economic Studies*, 44, 3–13.

——— (1983): "Natural Oligopolies," *Econometrica*, 51, 1469–1483.

Spence, M. (1976): "Product Differentiation and Welfare," *American Economic Review*, 66, 407–414.

Stoll, H. R. (2006): "Electronic Trading in Stock Markets," *Journal of Economic Perspectives*, 20, 153–174.

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

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

Weill, P.-o. (2008): "Liquidity premia in dynamic bargaining markets," *Journal of Economic Theory*, 140, 66–96.

Appendices

Supplement to "Competing on Speed"

Emiliano Pagnotta and Thomas Philippon

Imperial College London and NYU Stern School of Business

Contents

Appendix A Remarks on the Security Exchange Industry

A.1 Supplement to Section 2: Tables and Figures

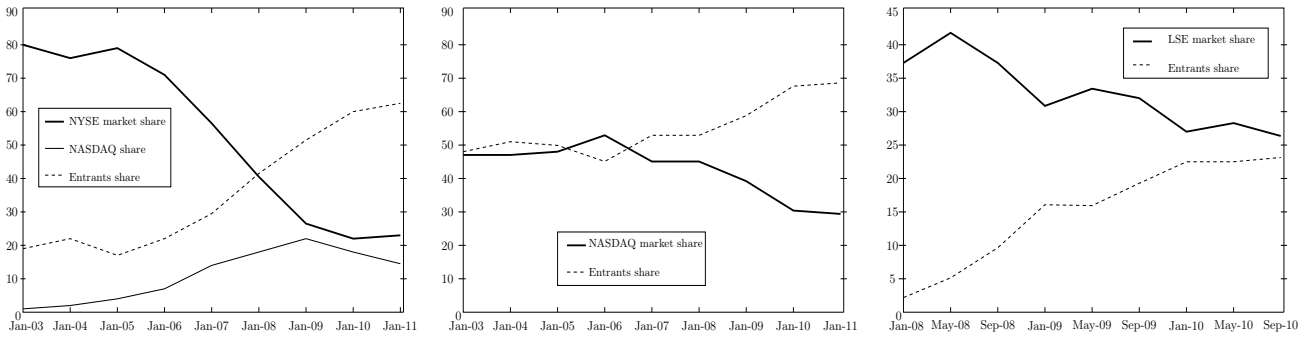

Figure A1. Equity market volume fragmentation by listing venue: NYSE, NASDAQ, and the London Stock Exchange (LSE). Source: Barclays Capital Equity Research.

Table A1
Selected Speed Investments by World Exchanges (2008-2012)

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

Source: Hand collected list from various sources.

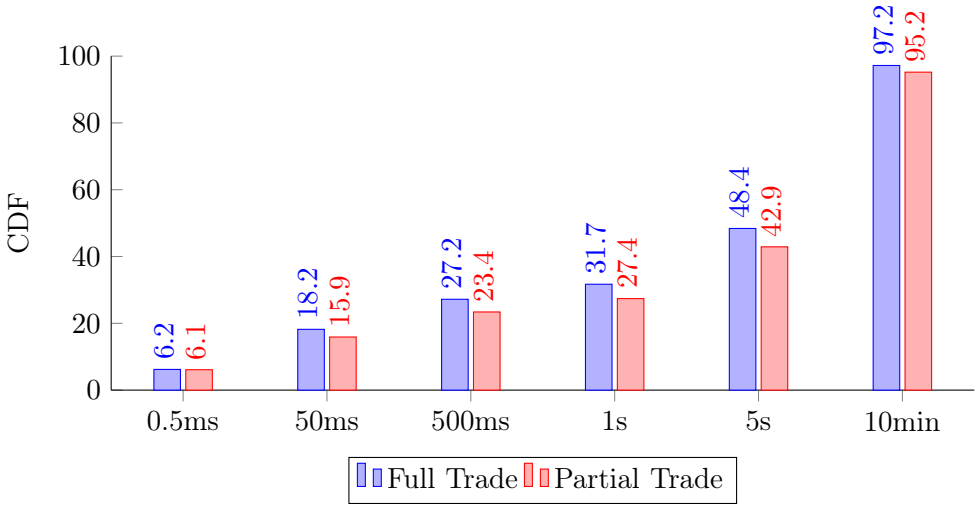

Figure A2. Distribution of trading speed in U.S. equity markets (Source: SEC, October 2013).

TABLE A2
VENUE COMPETITION AND INVESTOR PROTECTION IN SELECTED COUNTRIES

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

Source: www.fidessa.com and regulating agencies' websites

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

Corporate Bonds. The corporate bond market has traditionally operated in a decentralized fashion and over the phone (Duffie et al., 2005). Since the last financial crisis, institutional investors have begun migrating some of their orders execution away from voice and towards the electronic request for quote protocol (eRFQ).[57] The eRFQ represented an evolutionary step toward efficiency, not a market structure change. Similar to picking up a phone and calling a handful of dealers, it allows investors to gather a pool of potential liquidity providers. This mechanism can then be seen as the electronic version of the status quo. More recently, there has been a proliferation of electronic trading venues, some of them operating a central limit order book as well (CLOB). We list some of them below.

- Slow venue: Voice trading using traditional dealer banks.
- Fast venue: (mainly eRFQ). Bank-sponsored electronic bond trading networks: GSessions (Goldman Sachs), Bond Pool (Morgan Stanley), Price Improvement Network (UBS), Aladdin

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

Trading Network (BlackRock's), BondPoint (Knight). Bond trading platforms: Bloomberg, MarketAxxes, Tradeweb, Bonds.com.

- Faster venues. (mainly CLOB): ICAP's BrokerTec, GFI, NYSE Bonds.

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

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

- Slow Venue: Traditional banks/trading desks acting as voice brokers/dealers, trading at human speeds.

- Fast Venue. Multiple venues operating with different technologies. Inter-dealer electronic brokers platforms (EBS, Reuters, in London); ECNs (such as Currenex), 10-15 single-banks platforms. Trading speed is sub-second.

Despite the rapid growth of electronic venues in the FX market, by the end of 2012 only 60% of the global trading volume was electronic[58] (from 51% in 2010).

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

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

- Slow Venue: OTC broker–dealer (such as UBS, Credit Suisse and Morgan Stanley) trading over the phone, or traditional RFQ.

[58] Reported in Greenwich Associates' Global Foreign Exchange Services Study, 2012.

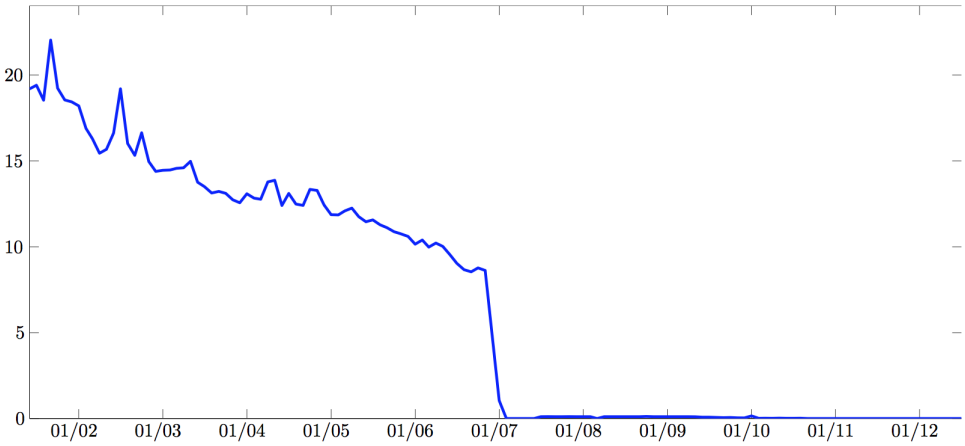

Figure A3. Average small order execution time for the NYSE, executed at the top of the book (seconds). Source: SEC Rule 605 reports.

- Fast Venue: Inter-dealer electronic platforms as ICAP i-Swaps, Tradition and BCG for IRS, and Bloomberg's BSEF for credit default swaps; several electronic Swap Execution Facilities (SEF).

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

A.3 Regulation of Speed

Minimum Speed. Regulation NMS differentiates market centers based on manual or automated quotations. A manual quotation is one that is considered not immediately and automatically accessible and which does not receive protection against trade-throughs. The SEC defines "immediate" as the fastest response possible without any programmed delay.[59] Given that only automated quotes are protected, the NYSE was deeply affected by Reg NMS's forced adoption of automation, a de-facto minimum speed requirement. Figure A3 illustrates this fact: At the time of full implementation in 2007, the average execution delay sharply declined from human to machine-driven speeds. A similar increase in the speed of the slow venues is to be expected in non-equity products due to the Dodd–Frank Act and MiFID II/EMIR push for electronic trading of OTC derivatives.

Maximum Speed. Several market interventions have been proposed to reduce trading speed, chiefly in equity markets. The motivation is typically to mitigate the front-running of institutional orders. Although most large financial institutions are now sophisticated enough so as to avoid simple

[59] For an exchange like IEX, the response time is not a result of a programmed delay; rather, the response time is merely a result of the coiled cable inside the "magic shoebox" to get to the matching engine where all participants are afforded equal access. This is essentially identical to a colocation center equally measuring connectivity cable to matching engines. In this regard, IEX would qualify as an automated market center under Reg NMS.

detection algorithms by 'predatory' high-frequency traders, continuous-time markets do not eliminate this possibility entirely Budish, Cramton, and Shim (2016). The current regulatory framework for equities in the U.S. and Europe does not impose explicit speed limits. Despite more intense scrutiny over algorithmic trading, speed limits are also absent in the current MiFID II proposals. Non regulatory speed limits, however, are sometime adopted. In foreign exchange market, for example, inter-dealer brokerage platforms such as Electronic Broking Services and Reuters have minimum quote life or minimum fill ratios that act effectively as a limit on maximum trading speed.

Appendix B Proofs of Propositions

B.1 Proof of Proposition 1

Consider the steady-state value functions for types $\sigma > \tilde{\sigma}$. They solve the following system: For the types holding the assets,

$$rV_{\sigma,+}(1) = \mu + \sigma + \frac{\gamma}{2}\left[V_{\sigma,-}(1) - V_{\sigma,+}(1)\right], \tag{27}$$

$$rV_{\sigma,-}(1) = \mu - \sigma + \frac{\gamma}{2}\left[V_{\sigma,+}(1) - V_{\sigma,-}(1)\right] + \rho\left(p + V_{\sigma,-}(0) - V_{\sigma,-}(1)\right), \tag{28}$$

and, for the types not holding the assets,

$$rV_{\sigma,-}(0) = \frac{\gamma}{2}\left[V_{\sigma,+}(0) - V_{\sigma,-}(0)\right], \tag{29}$$

$$rV_{\sigma,+}(0) = \frac{\gamma}{2}\left[V_{\sigma,-}(0) - V_{\sigma,+}(0)\right] + \rho\left(V_{\sigma,+}(1) - V_{\sigma,+}(0) - p\right). \tag{30}$$

Define $I_{\sigma,\epsilon} \equiv V_{\sigma,\epsilon}(1) - V_{\sigma,\epsilon}(0)$ as the value of owning the asset for type (σ,ϵ). Then, taking the differences of equations (27) to (30), we obtain

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

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

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

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

$$\bar{V}_\sigma(0) = \frac{\rho}{2r}(I_{\sigma,+} - p)$$
$$\bar{V}_\sigma(1) = \frac{\mu}{r} + \frac{\rho}{2r}(p - I_{\sigma,-})$$

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 $\sigma < \tilde{\sigma}$. They join the venue to sell at price p, and then do not trade again. Averaging over types $\epsilon = \pm 1$, the ex ante value function $\tilde{W}$ solves the Bellman equation $r\tilde{W} = \mu\bar{a} + \rho\left(p\bar{a} - \tilde{W}\right)$, and thus $\tilde{W} = \frac{\mu+\rho p}{r+\rho}\bar{a}$. Since $\mu + \rho p = \frac{\mu}{r}(r+\rho) + \rho\left(p - \frac{\mu}{r}\right)$ we can rewrite

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

From the definition of $\hat{\sigma}$ we also now that $\frac{\rho}{r+\rho}(rp-\mu) = s(\rho)\tilde{\sigma}$, with $s(\rho) \equiv \frac{\rho}{r+\gamma+\rho}$, therefore $\tilde{W} = \frac{\mu\bar{a}}{r} + s\frac{\bar{a}}{r}\tilde{\sigma}$. The marginal type $\tilde{\sigma}(p,\rho)$ is defined in (5), is increasing in p and decreasing in ρ. The key point is that $\hat{W}$ does not depend on the type σ, but only on the price and speed of the venue. Of course we also have $\tilde{W} = \bar{a}\bar{V}_{\tilde{\sigma}}(1)$.

Let us now consider the steady state types, $\sigma > \tilde{\sigma}$. 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

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

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

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

Q.E.D.

B.2 Proof of Proposition 2

Let us introduce some notations to simplify the exposition:

$$\begin{aligned} \alpha &\equiv 2\bar{a} \\ k &\equiv \frac{s_1}{s_2 - s_1} \\ \nu(\hat{\sigma}) &\equiv \frac{1 - G(\hat{\sigma})}{g(\hat{\sigma})} \end{aligned} \tag{31}$$

The monopoly allocation $\hat{\sigma}_m$ is the solution to $\hat{\sigma}_m = \nu(\hat{\sigma}_m)$. Rearranging the first order conditions, the segmentation allocation $(\hat{\sigma}_1^{seg}, \hat{\sigma}_2^{seg})$ is the solution to

$$\begin{aligned} \hat{\sigma}_2 &= \nu(\hat{\sigma}_2) - k\hat{\sigma}_1 \\ \hat{\sigma}_1\left(\frac{g(\hat{\sigma}_1)}{g(\hat{\sigma}_2)} + k\right) &= \frac{g(\hat{\sigma}_1)}{g(\hat{\sigma}_2)}\nu(\hat{\sigma}_1) - \nu(\hat{\sigma}_2) \end{aligned}$$

The price protection allocation $\left(\hat{\sigma}_1^{prot}, \hat{\sigma}_2^{prot}\right)$ is the solution to

$$\begin{aligned} \hat{\sigma}_{12} &= \nu(\hat{\sigma}_2) - z(\alpha)k\hat{\sigma}_1 \\ \hat{\sigma}_1\left(\frac{g(\hat{\sigma}_1)}{g(\hat{\sigma}_2)} + \alpha z(\alpha)k\right) &= \frac{g(\hat{\sigma}_1)}{g(\hat{\sigma}_2)}\nu(\hat{\sigma}_1) - \alpha\nu(\hat{\sigma}_2) \end{aligned}$$

where we highlight in red the differences to help the comparison. Remember that $z(\alpha)$ is increasing. Notice first that $\hat{\sigma}_2 < \hat{\sigma}_m$ irrespective of whether prices are free or protected. Participation in venue 2 alone is higher than under monopoly.

Proof of Point (ii). We use the following notations to simplify the algebra

$$x \equiv \hat{\sigma}_1, \ y \equiv \hat{\sigma}_2.$$

The duopoly system with segmented prices (or with $\bar{a} = 1/2$) is

$$\begin{aligned} G(y) - G(x) &= (g(x) + kg(y))\, x \\ 1 - G(y) &= g(y)(y + kx) \end{aligned}$$

We can differentiate the system with respect to k:

$$\begin{aligned} g(y)\,dy - g(x)\,dx &= g(x)\,dx + kg(y)\,dx + x\left(g'(x)\,dx + g(y)\,dk + kg'(y)\,dy\right) \\ -g(y)\,dy &= g'(y)\,dy\,(y + kx) + g(y)\,(dy + d[kx]) \end{aligned}$$

After some manipulations and simplifications, we get

$$\begin{aligned} dx\left(2g\left(x\right)+xg'\left(x\right)+kg\left(y\right)\theta\right) &= -xg\left(y\right)\theta dk \\ dy &= -g\left(y\right)\frac{d\left[kx\right]}{2g\left(y\right)+g'\left(y\right)\left(y+kx\right)} \end{aligned}$$

where

$$\theta \equiv \frac{3g\left(y\right)+yg'\left(y\right)}{2g\left(y\right)+g'\left(y\right)\left(y+kx\right)}$$

We can then prove that under 1, we have:

$$\frac{\partial y}{\partial k}<0 \text{ and } \frac{\partial x}{\partial k}<0.$$

First note that the second duopoly FOC implies that $\frac{1-G(y)}{g(y)}=y+kx$. Therefore, under 1, we have

$$2g\left(y\right)+g'\left(y\right)\left(y+kx\right)\geq 0.$$

This shows that the denominator in θ is strictly positive. Let's study the numerator and show that it is also strictly positive. Either $g'\left(y\right)>0$ and then $3g\left(y\right)+yg'\left(y\right)>0$. Or $g'\left(y\right)<0$ but then, since $kx>0$,

$$3g\left(y\right)+yg'\left(y\right)>2g\left(y\right)+yg'\left(y\right)>2g\left(y\right)+\left(y+kx\right)g'\left(y\right)>0.$$

Therefore $\theta>0$. It is then easy to see that

$$\frac{\partial x}{\partial k}<0.$$

For $\frac{\partial y}{\partial k}$, we need to check that . This is true since

$$\partial x\left(\frac{2g\left(x\right)+xg'\left(x\right)}{g\left(y\right)\theta}+k\right)+x\partial k=0,$$

and $x>0$, and $2g\left(x\right)+xg'\left(x\right)>0$ under 1). Therefore

$$\frac{\partial y}{\partial k}<0.$$

QED.

Point (iii) with Uniform Distribution Here it is convenient to define differentiation as

$$d \equiv {}^{s_2}/_{s_1}.$$

With a uniform distribution over $[0,\overline{\sigma}]$ the system is

$$\begin{aligned} y &= \overline{\sigma} - y - \frac{z(\alpha)\, x}{d-1} \\ x\left(1 + \frac{\alpha z(\alpha)}{d-1}\right) &= \overline{\sigma} - x - \alpha(\overline{\sigma} - y) \end{aligned}$$

so

$$\begin{aligned} 2y &= \overline{\sigma} - \frac{z(\alpha)\, x}{d-1} \\ x\left(2 + \frac{\alpha z(\alpha)}{d-1}\right) &= (1-\alpha)\,\overline{\sigma} + \alpha y \end{aligned}$$

so

$$\begin{aligned} y &= \frac{\overline{\sigma}(d-1) - z(\alpha)x}{2(d-1)} \\ x &= \frac{\left(1 - \frac{\alpha}{2}\right)(d-1)}{2(d-1) + \frac{3}{2}\alpha z(\alpha)}\overline{\sigma} \end{aligned}$$

The solution for x is clear: x is decreasing in α. **Price protection means α goes down, so x goes up**. The impact on y is ambiguous since

$$y = \left(1 - z(\alpha)\frac{1 - \frac{\alpha}{2}}{2(d-1) + \frac{3}{2}\alpha z(\alpha)}\right)\frac{\overline{\sigma}}{2}$$

and $z(\alpha) = 1 - \frac{1+\frac{r}{\rho_1}}{1+\frac{r}{\rho_2}}(1-\alpha)$. Clearly, if α is small, then y decreases with α. But if α and d are both close to 1, this can be reversed.

Point (iii) with Exponential Distribution. Under exponential distribution, we have $G(\sigma) = 1 - e^{-\sigma/\nu}$ and therefore $\nu(\hat{\sigma}) = \nu$ and the system is

$$\begin{aligned} \frac{\hat{\sigma}_2}{\nu} &= 1 - z(\alpha)\, k \frac{\hat{\sigma}_1}{\nu} \\ \frac{\hat{\sigma}_1}{\nu}\left(e^{\frac{\hat{\sigma}_2 - \hat{\sigma}_1}{\nu}} + \alpha z(\alpha)\, k\right) &= e^{\frac{\hat{\sigma}_2 - \hat{\sigma}_1}{\nu}} - \alpha \end{aligned}$$

It is convenient to define

$$\begin{aligned} \Delta &\equiv \frac{\hat{\sigma}_2 - \hat{\sigma}_1}{\nu} \\ x &\equiv \frac{\hat{\sigma}_1}{\nu} \end{aligned} \tag{32}$$

and , so that we can write the system in (x, Δ):

$$(1 + z(\alpha) k) x = 1 - \Delta \quad (33)$$

$$e^{\Delta} - \alpha = \left(e^{\Delta} + \alpha z(\alpha) k\right) x \quad (34)$$

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

The second equation of the system is

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

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

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

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 $\frac{\alpha(1+kz)}{e^{\Delta}+\alpha z k}$ is increasing in α. Therefore the equilibrium condition $e^{\Delta} - \alpha = \left(e^{\Delta} + \alpha k z\right) x$ implies a schedule x increasing in Δ and decreasing in α. The first equilibrium condition $(1 + z(\alpha) k) x = 1 - \Delta$ gives a schedule x decreasing in Δ and decreasing in α. Straightforward analysis then shows **that x must be decreasing in α.** The segmented price model corresponds to $\alpha = 1$, and the protected price model to $\alpha = 2a < 1$. **Therefore, x and $\hat{\sigma}_1 = \nu x$ are higher under protection.**

Impact of protection on $\hat{\sigma}_2$

The analysis of $\hat{\sigma}_2$ is ambiguous. It is clear that when $k \to 0$ we have $\hat{\sigma}_2 \to \nu$, which is the monopoly solution. Define $y = \frac{\hat{\sigma}_2}{\nu} = x + \Delta$, and get the system

$$(1 + kz) y = 1 + kz\Delta$$

$$1 - y = kz \frac{e^{\Delta} - \alpha}{e^{\Delta} + \alpha k z}$$

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

$$(1 + k\alpha) y = 1 + k\alpha\Delta$$

$$1 - y = k\alpha \frac{e^{\Delta} - \alpha}{e^{\Delta} + k\alpha^2}$$

We study the case where α is close to one. The free price solution is

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

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

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

The second one gives

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

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 Δ 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 ϵ: σ_2 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 ϵ: σ_2 is then lower, and participation in the fast venue 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

$$F_1(t_1, t_2, \alpha) = t_1(1 - \alpha + \alpha G(\hat{\sigma}_2) - G(t_1))$$
$$F_2(t_1, t_2, \alpha) = t_2(1 - G(\hat{\sigma}_2))$$

and we have $\hat{\sigma}_{12} = k(t_2 - z(\alpha) t_1)$ and $\hat{\sigma}_1 = t_1$.

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 segmentation case corresponds to $\alpha = 1$ and $z = 1$. We can always return to the system in σ 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

$$1 - \alpha + \alpha G(\hat{\sigma}_2) - G(\hat{\sigma}_1) = t_1(\alpha z(\alpha) k g(\hat{\sigma}_2) + g(\hat{\sigma}_1))$$
$$1 - G(\hat{\sigma}_2) = t_2 k g(\hat{\sigma}_2)$$

With **exponential distributions** we have that t_2 does not depend on α: $t_2 = \frac{\nu}{k}$. 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

$$\frac{\partial F_1}{\partial \alpha} = t_1\left(-1 + G(\hat{\sigma}_{12}) + \alpha g(\hat{\sigma}_{12})\frac{\partial \hat{\sigma}_2}{\partial \alpha}\right) = t_1\left(-1 + G(\hat{\sigma}_2) - \alpha g(\hat{\sigma}_2) k t_1 z'(\alpha)\right)$$

Since $z'(\alpha) > 0$, we see that $\frac{\partial F_1}{\partial \alpha} < 0$: price protection increases the profits of the slow venue. The economic intuition is simple. The term $-1 + G(\hat{\sigma}_2)$ 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 $\hat{\sigma}_2$.

With a **uniform distribution**, on the other hand, we have

$$y = \frac{m - z(\alpha) k x}{2}$$
$$x = \frac{1 - \frac{\alpha}{2}}{2 + \frac{3}{2}\alpha z(\alpha) k} m$$

so

$$F_1 = x(1 - \alpha + \alpha y/m - x/m)$$
$$F_2 = \left(\frac{y}{k} + zx\right)(1 - y/m)$$

and

$$F_1 = \frac{1}{2+\frac{3}{2}\alpha z k}\left(1-\frac{\alpha}{2}\right)^2\left(1-\frac{1}{2}\frac{2+\alpha z k}{2+\frac{3}{2}\alpha z k}\right)m$$
$$F_2 = \left(\frac{1}{k}+\frac{z}{2}\frac{1-\frac{\alpha}{2}}{2+\frac{3}{2}\alpha z\left(\alpha\right)k}\right)\left(\frac{1}{2}+\frac{z\left(\alpha\right)k}{2}\frac{1-\frac{\alpha}{2}}{2+\frac{3}{2}\alpha z\left(\alpha\right)k}\right)m$$

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

B.3 Proof of Proposition 3.

The interior solution FOC for speed is $C'\left(s\right) = \left(1-G\left(\hat{\sigma}_m\right)\right)\frac{\hat{\sigma}_m}{2r}$. Using Assumption 2 we have $C'\left(s\right) = \frac{c(r+\gamma)}{(1-s)^2}$. Under exponential distribution of types we have $\hat{\sigma}_m = \nu$ and thus $\left(1-G\left(\hat{\sigma}_m\right)\right)\frac{\hat{\sigma}_m}{2r} = \frac{\nu}{2er}$. Combining these expressions yields $s_m = 1-\left(2rc(\gamma+r)e/\nu\right)g^{1/2}$. Under uniform distribution of types, we have $\hat{\sigma}_m = \overline{\sigma}/2$ and $\left(1-G\left(\hat{\sigma}_m\right)\right)\frac{\hat{\sigma}_m}{2r} = \frac{\overline{\sigma}}{8r}$. Thus, $s_m = 1-\left(8rc(r+\gamma)/\overline{\sigma}\right)^{\frac{1}{2}}$. *Q.E.D.*

B.4 Proof of Proposition 4.

Consider the venue 2's program

$$\max_{s_2}\frac{1}{2r}\left(\hat{\sigma}_2\left(s_2-s_1\right)+\hat{\sigma}_1 s_1\right)\left(1-G\left(\hat{\sigma}_2\right)\right)-C\left(s_2\right). \tag{35}$$

It is immediate that this program converges to the monopolist's when $s_1 \to 0$. We then have $\lim_{s_1\to 0} S_2\left(s_1\right) = s_m$. Thus, to show that $s_2 > s_m$, it suffices to show that $S_2'\left(s_1\right) > 0$.

Differentiating the FOC of equation 35 with respect to s_1, and re-arranging, yields

$$S_2'\left(s_1\right) = \frac{\partial^2\pi_2}{\partial s_2\partial s_1}\left(C''\left(s_2\right)-\frac{\partial^2\pi_2}{\partial s_2^2}\right)^{-1}. \tag{36}$$

We now use the uniform distribution to sign the terms in the RHS of equation 36. The revenue functions are given by

$$\pi_2\left(s_1,s_2\right) = \frac{1}{2r}\left(\hat{\sigma}_2\left(s_2-s_1\right)+\hat{\sigma}_1 s_1\right)\left(1-\frac{\hat{\sigma}_2}{\overline{\sigma}}\right) \tag{37}$$
$$\pi_1\left(s_1,s_2\right) = \frac{s_1}{2r}\hat{\sigma}_1\frac{\hat{\sigma}_2-\hat{\sigma}_1}{\overline{\sigma}}. \tag{38}$$

Using equations 18 and 19 and $d \equiv \frac{s_2}{s_1}$ we have

$$\hat{\sigma}_1 = \overline{\sigma}\frac{d-1}{4d-1};\quad \hat{\sigma}_2 = \overline{\sigma}\frac{2d-1}{4d-1}. \tag{39}$$

or

$$\hat{\sigma}_1 = \overline{\sigma}\frac{s_2-s_1}{4s_2-s_1};\quad \hat{\sigma}_2 = \overline{\sigma}\frac{2s_2-s_1}{4s_2-s_1}$$

Replacing $\hat{\sigma}_1$ and $\hat{\sigma}_2$ in the revenue functions by the expressions in 39, yields

$$\pi_1(s_1, s_2) = \frac{\overline{\sigma}}{2r} \frac{s_2 - s_1}{(4s_2 - s_1)^2} s_1 s_2$$
$$\pi_2(s_1, s_2) = \frac{\overline{\sigma}}{2r} \frac{s_2 - s_1}{(4s_2 - s_1)^2} (2s_2)^2$$

Notice that $\pi_2 = 4\frac{s_2}{s_1}\pi_1$. After some algebra, one can show that

$$\frac{\partial^2 \pi_2}{\partial s_2^2} = -\frac{4\overline{\sigma}}{r} \frac{s_1^2 (s_1 + 5s_2)}{(4s_2 - s_1)^4} < 0$$
$$\frac{\partial^2 \pi_2}{\partial s_2 \partial s_1} = \frac{4\overline{\sigma}}{r} \frac{s_1 s_2 (s_1 + 5s_2)}{(4s_2 - s_1)^4} > 0.$$

These inequalities, together with convexity of C, yield $S_2'(s_1) > 0$.

Q.E.D.

B.5 Proof of Proposition 5

Consider the maximum speed first. Holding market shares constant, the regulator seeks to maximize social welfare in each venue i. For venue 1, the optimality condition is $\frac{1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_2} \sigma dG(\sigma) = C'(s_1)$. Using equation 39 we can compute $\frac{1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_2} \sigma dG(\sigma) = \frac{\overline{\sigma}}{4r}\frac{d(3d-2)}{(4d-1)^2}$, where $d \equiv \frac{s_2}{s_1}$. The marginal cost must equal the venue's marginal revenue in an interior solution, which from equation 23 is given by $\frac{\overline{\sigma}}{2r}\frac{d^2(4d-7)}{(4d-1)^3}$. Straightforward calculations then show that $\frac{1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_2} \sigma dG(\sigma) > C'(s_1)$, implying under provision of speed at the market equilibrium for venue 1. Similarly, the regulator optimality condition for venue two is $\frac{1}{2r}\int_{\hat{\sigma}_2}^{\overline{\sigma}} \sigma dG(\sigma) = C'(s_2)$. Using equation 39 we can compute $\frac{1}{2r}\int_{\hat{\sigma}_2}^{\overline{\sigma}} \sigma dG(\sigma) = \frac{\overline{\sigma}}{r}\frac{d(3d-1)}{(4d-1)^2}$. From equation 23 we have that the marginal cost must equal at an market solution $\frac{2\overline{\sigma}}{r}\frac{d^2(2d+1)}{(4d-1)^3}$. Straightforward calculations then show that $\frac{1}{2r}\int_{\hat{\sigma}_2}^{\overline{\sigma}} \sigma dG(\sigma) > C'(s_2)$, implying under provision of speed at the market equilibrium for venue 2.

Consider now the minimum speed. We start by computing the total derivative of the welfare function 12 with respect to s_1 at the market equilibrium.

$$\frac{\mathrm{d}\mathcal{W}}{\mathrm{d}s_1} = \left(\frac{1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_2} \sigma dG(\sigma) - C'(s_1)\right) + \left(\frac{1}{2r}\int_{\hat{\sigma}_2}^{\overline{\sigma}} \sigma dG(\sigma) - C'(s_2)\right) S_2'(s_1)$$
$$- \frac{1}{2r}\frac{\mathrm{d}d}{\mathrm{d}s_1}\left(\hat{\sigma}_1 s_1 \frac{\partial \hat{\sigma}_1}{\partial d} + \hat{\sigma}_2 (s_2 - s_1)\frac{\partial \hat{\sigma}_2}{\partial d}\right). \tag{40}$$

We have shown above that bracketed expressions in the first two terms of equation 40 are positive, and also that $S_2'(s_1) > 0$, thus the sum of the first two terms is positive. Consider the third term of the RHS of equation 40. It is immediate from Proposition 2 that $\frac{\partial \hat{\sigma}_1}{\partial d}$ and $\frac{\partial \hat{\sigma}_2}{\partial d}$ are positive under Assumption 1. Also notice that $\frac{\mathrm{d}d}{\mathrm{d}s_1} = \frac{S_2'(s_1) - d}{s_1}$. The revenue function for each venue is homogeneous of degree one, implying that the marginal revenue functions are homogeneous of degree zero. By

Euler's theorem then $d = -\frac{\partial^2 \pi_2}{\partial s_2 \partial s_1} / \frac{\partial^2 \pi_2}{\partial s_2^2}$, and $S_2'(s_1) < d$ given equation 36, which yields $\frac{\mathrm{d}d}{\mathrm{d}s_1} < 0$. We conclude that around the duopoly's equilibrium $\frac{\mathrm{d}\mathcal{W}}{\mathrm{d}s_1} > 0$.

B.6 Proof of Proposition 6.

The relationship between entry costs κ 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 $\underline{\pi}_i \equiv \min\left\{\pi_i^{prot}, \pi_i^{seg}\right\}$ with $i \in \{1, 2\}$. We analyze below the existence of Nash equilibrium (NE) in pure strategies of the normal-form game shown in figure 7.

- *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 π_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 Nash equilibria 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.*

B.7 Proof of Proposition 7.

The welfare functions under monopoly and duopoly are given in Lemma 8. The welfare gain of moving from a monopoly to a duopoly is therefore

$$\Delta\mathcal{W}_{1\to 2} = \frac{s_1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_2} \sigma dG(\sigma) - C(s_1) + \frac{s_2}{2r}\int_{\hat{\sigma}_2}^{\overline{\sigma}} \sigma dG(\sigma) - C(s_2) - \frac{s_m}{2r}\int_{\hat{\sigma}_m}^{\overline{\sigma}} \sigma dG(\sigma) + C(s_m) - \kappa.$$

On the other hand, entry is profitable for the slow venue if and only if $\pi_1 > \kappa$:

$$\frac{s_1\hat{\sigma}_1}{2r}\left(G(\hat{\sigma}_2) - G(\hat{\sigma}_1)\right) > \kappa + C(s_1).$$

Excess entry happens if and only if $\pi_1 > \kappa$ and $\Delta\mathcal{W}_{1\to 2} < 0$ hold simultaneously. A *necessary* condition for excess entry is therefore

$$\frac{s_1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_2}\sigma dG(\sigma)+\frac{s_2}{2r}\int_{\hat{\sigma}_2}^{\overline{\sigma}}\sigma dG(\sigma)-C(s_2)-\frac{s_m}{2r}\int_{\hat{\sigma}_m}^{\overline{\sigma}}\sigma dG(\sigma)+C(s_m)<\frac{s_1\hat{\sigma}_1}{2r}\left(G(\hat{\sigma}_2)-G(\hat{\sigma}_1)\right)$$

which we can rewrite as

$$\frac{s_1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_2}(\sigma-\hat{\sigma}_1)dG(\sigma)+\frac{s_2}{2r}\int_{\hat{\sigma}_2}^{\overline{\sigma}}\sigma dG(\sigma)-\frac{s_m}{2r}\int_{\hat{\sigma}_m}^{\overline{\sigma}}\sigma dG(\sigma)<C(s_2)-C(s_m)$$

The first term is strictly positive. We know from Proposition 2 that $\hat{\sigma}_2 < \hat{\sigma}_m$ and from Proposition 4 that $s_2 \geq s_m$, therefore the second term is also strictly positive. Finally, the term on the right hand side goes to zero when the costs are small.

Q.E.D.

Appendix C Proofs of Lemmas

C.1 Proof of Lemma 2

To see the steady-state allocations, add (7) and $\left(\frac{\gamma}{2}+\rho\right)\alpha_{\sigma,-}(1)=\frac{\gamma}{2}\alpha_{\sigma,+}(1)$ 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 asset holders is simply $\alpha_{\sigma,+}(1)+\alpha_{\sigma,-}(1)=1/2$. *Q.E.D.*

C.2 Proof of Lemma 3

In a frictionless competitive market we have maximum investor participation. Thus, the marginal type is given by

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

and thus $\hat{\sigma}_w=G^{-1}(1-2\overline{a})$. Using this fact and $s_w=1$ yields $p_w=\frac{1}{r}\left[\mu+G^{-1}(1-2\overline{a})\right]$. Moreover, the instantaneous transaction rate becomes

$$\mathcal{V}_w=\frac{\gamma}{4}(1-G(\hat{\sigma}_w))=\frac{\gamma}{2}\overline{a}.$$

The expression $\mathcal{W}_w=\frac{1}{2r}\int_{G^{-1}(1-2\overline{a})}^{\overline{\sigma}}\sigma dG(\sigma)$ for welfare is obtained directly as a particular case of Lemma 8. *Q.E.D.*

C.3 Proof of Lemma 7

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

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

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

$$\hat{\sigma}_2 = \frac{2r}{s_2 - s_1}\left(q_2 - \frac{z}{2\bar{a}} q_1\right)$$

where

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

Note that z is an increasing function of $\bar{a}$ that satisfies $z \leq 1$. When $a \approx 0.5$, we have $z \approx 1$, and $z \approx 2\bar{a}$ when r/ρ is small (the realistic case). The profits of venue 1 are

$$\pi_1^{prot} = q_1 (G(\hat{\sigma}_2) - G(\hat{\sigma}_1) + \delta_1)$$

In the protected price equilibrium, firms therefore maximize

$$\begin{aligned}
\max_{q_2} \pi_2^{prot} &= q_2 (1 - G(\hat{\sigma}_2)) \\
\max_{q_1} \pi_1^{prot} &= \frac{q_1}{2\bar{a}} (1 - 2\bar{a} + 2\bar{a} G(\hat{\sigma}_2) - G(\hat{\sigma}_1))
\end{aligned}$$

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

$$\begin{aligned}
1 - G(\hat{\sigma}_2) &= g(\hat{\sigma}_2)\left(\hat{\sigma}_2 + z\frac{s_1}{s_2 - s_1}\hat{\sigma}_1\right) \\
1 - 2\bar{a} + 2\bar{a} G(\hat{\sigma}_2) - G(\hat{\sigma}_1) &= \left(g(\hat{\sigma}_1) + 2\bar{a} z \frac{s_1}{s_2 - s_1} g(\hat{\sigma}_2)\right)\hat{\sigma}_1,
\end{aligned}$$

which, after using the definition of d, yields the system in Lemma 7. *Q.E.D.*

C.4 Proof of Lemma 8

The welfare formula in equation 12 reflects the joint trading surplus of investors. Transfers from investors to venue owners do not represent net social gains.

Consider the segmented case first. The welfare of type σ joining venue i is

$$W_i(\sigma) - W_{out} = \frac{s_i \tilde{\sigma}(p_i, \rho_i)}{r}\bar{a} + \frac{s_i}{2r}\max(0; \sigma - \tilde{\sigma}(p_i, \rho_i)),$$

where the function $\tilde{\sigma}$ is defined by

$$\tilde{\sigma}(p,\rho) \equiv \frac{r+\rho+\gamma}{r+\rho}(rp-\mu).$$

The net value of participation, $W-W_{out}$, is composed of two parts. One is the option to sell the asset on the exchange: $\frac{s\bar{a}\tilde{\sigma}}{r} = \frac{\rho}{r+\rho}\left(p-\frac{\mu}{r}\right)\bar{a}$. It is independent of σ. It is the value that can be achieved by all types $\sigma < \tilde{\sigma}$ with the "sell and leave" strategy. The term $\frac{\rho}{r+\rho}$ captures the expected trading delay . The second part, $\frac{s}{2r}\max(0;\sigma-\tilde{\sigma})$, is the value of trading repeatedly and it depends on the type σ. This part of the value function is super-modular in (s,σ). As explained in Proposition 1, when a "sell and leave" investor joins venue i her welfare gains $W(\sigma,\hat{\sigma}_i,s_i)-W_{out}$ are independent of her type and amount to $s_i\bar{a}\tilde{\sigma}_i/r$. The mass of these investors equals $\left(\frac{1}{2\bar{a}}-1\right)(G(\hat{\sigma}_2)-G(\hat{\sigma}_1))$ and $\left(\frac{1}{2\bar{a}}-1\right)(1-G(\hat{\sigma}_2))$ in venues 1 and 2, respectively. Thus, total social gains for this group are given by

$$\frac{1}{r}\left(\frac{1}{2}-\bar{a}\right)\left(\left(G(\hat{\sigma}_2)-G(\hat{\sigma}_1)\right)s_1\tilde{\sigma}_1+\left(1-G(\hat{\sigma}_2)\right)s_2\tilde{\sigma}_2\right) \tag{41}$$

The welfare gains of repeat traders investors are

$$s_1\int_{\hat{\sigma}_1}^{\hat{\sigma}_2}\left(\frac{\bar{a}\tilde{\sigma}_1}{r}+\frac{\sigma-\tilde{\sigma}_1}{2r}\right)dG(\sigma) = \int_{\hat{\sigma}_1}^{\hat{\sigma}_2}\frac{s_1\sigma}{2r}dG(\sigma)-\left(G(\hat{\sigma}_2)-G(\hat{\sigma}_1)\right)\frac{s_1\tilde{\sigma}_1}{r}\left(\frac{1}{2}-\bar{a}\right) \tag{42}$$

$$s_2\int_{\hat{\sigma}_2}^{\bar{\sigma}}\left(\frac{\bar{a}\tilde{\sigma}_2}{r}-\frac{\sigma-\tilde{\sigma}_2}{2r}\right)dG(\sigma) = \int_{\hat{\sigma}_2}^{\bar{\sigma}}\frac{s_2\sigma}{2r}dG(\sigma)-\left(1-G(\hat{\sigma}_2)\right)\frac{s_2\tilde{\sigma}_2}{r}\left(\frac{1}{2}-\bar{a}\right) \tag{43}$$

where for short we defined

$$\tilde{\sigma}_i \equiv \tilde{\sigma}(p_i,\rho_i).$$

Adding 41, 42 and 43, and subtracting speed investment costs yields total gains from trade:

$$\mathcal{W}(2) = \frac{s_1}{2r}\int_{\hat{\sigma}_1}^{\hat{\sigma}_2}\sigma dG(\sigma)+\frac{s_2}{2r}\int_{\hat{\sigma}_2}^{\bar{\sigma}}\sigma dG(\sigma)-\sum_{i=1,2}C(s_i)-2\kappa.$$

The single venue expression in equation 20 is a particular case.

Consider now the case of price protection. In that case there is only one price so

$$W_i(\sigma)-W_{out} = \frac{s_i\tilde{\sigma}(p,\rho_i)}{r}\bar{a}+\frac{s_i}{2r}\max(0;\sigma-\tilde{\sigma}(p,\rho_i)).$$

All the temporary traders join market 1, therefore their utility is $\left(\frac{1}{2}-\bar{a}\right)(1-G(\hat{\sigma}_1))\frac{s_1\tilde{\sigma}_1}{r}$. For the repeat traders, we have as before

$$s_1\int_{\hat{\sigma}_1}^{\hat{\sigma}_2}\left(\frac{\bar{a}\tilde{\sigma}_1}{r}+\frac{\sigma-\tilde{\sigma}_1}{2r}\right)dG(\sigma) = \int_{\hat{\sigma}_1}^{\hat{\sigma}_2}\frac{s_1\sigma}{2r}dG(\sigma)-\left(G(\hat{\sigma}_2)-G(\hat{\sigma}_1)\right)\frac{s_1\tilde{\sigma}_1}{r}\left(\frac{1}{2}-\bar{a}\right) \tag{44}$$

$$s_2\int_{\hat{\sigma}_2}^{\bar{\sigma}}\left(\frac{\bar{a}\tilde{\sigma}_2}{r}-\frac{\sigma-\tilde{\sigma}_2}{2r}\right)dG(\sigma) = \int_{\hat{\sigma}_2}^{\bar{\sigma}}\frac{s_2\sigma}{2r}dG(\sigma)-\left(1-G(\hat{\sigma}_2)\right)\frac{s_2\tilde{\sigma}_2}{r}\left(\frac{1}{2}-\bar{a}\right) \tag{45}$$

Adding up, we get

$$\int_{\hat{\sigma}_1}^{\hat{\sigma}_2} \frac{s_1\sigma}{2r} dG(\sigma) + \int_{\hat{\sigma}_2}^{\bar{\sigma}} \frac{s_2\sigma}{2r} dG(\sigma) - \left(\frac{1}{2} - \bar{a}\right)(1 - G(\hat{\sigma}_2)) \left(\frac{s_2\tilde{\sigma}_2}{r} - \frac{s_1\tilde{\sigma}_1}{r}\right)$$

So the welfare

$$\mathcal{W}^{prot}(2) = \mathcal{W}(2) - \left(\frac{1}{2} - \bar{a}\right)(1 - G(\hat{\sigma}_2)) \left(\frac{s_2\tilde{\sigma}_2}{r} - \frac{s_1\tilde{\sigma}_1}{r}\right),$$

and since $\frac{\rho}{r+\rho}(rp - \mu) = s(\rho)\tilde{\sigma}$, we can also write

$$\mathcal{W}^{prot}(2) = \mathcal{W}(2) - \left(\frac{1}{2} - \bar{a}\right)(1 - G(\hat{\sigma}_2)) \left(p - \frac{\mu}{r}\right) \left(\frac{\rho_2}{r+\rho_2} - \frac{\rho_1}{r+\rho_1}\right).$$

The welfare loss relative to the segmented case comes from the fact that temporary traders liquidate their holdings more slowly since they only access the market via venue 1.

Q.E.D.

Appendix D Transition Dynamics

In this section we compute the transition dynamics when the price is not constant. This happens when the market clearing in stocks is not enough to ensure market clearing in flows. To understand the issue, consider one market with two venues. Let N_1 be the number of agents in venue 1, and N_2 in venue 2. In our model this corresponds to

$$\begin{aligned} N_1 &= \delta_1 + G(\hat{\sigma}_2) - G(\hat{\sigma}_1) \\ N_2 &= \delta_2 + 1 - G(\hat{\sigma}_2) \end{aligned}$$

The total quantity of assets is $(N_1 + N_2)\bar{a}$. Let the average holding of agents in venue i be $\bar{a}_{i,t}$. All assets must be held, therefore we must have at any point in time

$$N_1\bar{a}_{1,t} + N_2\bar{a}_{2,t} = (N_1 + N_2)\bar{a}$$

D.1 Segmented Markets

Let us show that the flow market clearing are satisfied with segmented markets. Let $\bar{a}^*_{i,t}$ the average optimal demand in a market. The flow market clearing is

$$N_i\rho_i\bar{a} = N_i\rho_i\bar{a}^*_{i,t}$$

which boils down to

$$\bar{a} = \bar{a}^*_{i,t}$$

so we can find a time invariant demand. The average demand of light traders is zero. The average demand of repeat traders is $1/2$, so

$$\bar{a}_1^* = \frac{\frac{1}{2}\left(G\left(\hat{\sigma}_2\right) - G\left(\hat{\sigma}_1\right)\right)}{\delta_1 + G\left(\hat{\sigma}_2\right) - G\left(\hat{\sigma}_1\right)}$$

and similarly for venue 2

$$\bar{a}_2^* = \frac{\frac{1}{2}\left(1 - G\left(\hat{\sigma}_2\right)\right)}{\delta_2 + 1 - G\left(\hat{\sigma}_2\right)}$$

The flow market clearing conditions are therefore

$$\left(\delta_1 + G\left(\hat{\sigma}_2\right) - G\left(\hat{\sigma}_1\right)\right)\bar{a} = \frac{G\left(\hat{\sigma}_2\right) - G\left(\hat{\sigma}_1\right)}{2}$$

and

$$\left(\delta_2 + 1 - G\left(\hat{\sigma}_2\right)\right)\bar{a} = \frac{1 - G\left(\hat{\sigma}_2\right)}{2}$$

which are the conditions derived in the paper.

D.2 Integrated Markets

The flow market clearing is now

$$N_1\rho_1\bar{a}_{1,t} + N_2\rho_2\bar{a}_{2,t} = N_1\rho_1\bar{a}_{1,t}^* + N_2\rho_2\bar{a}_{2,t}^*$$

and this does not simplify as before. Let us show that there is excess demand in the short term.

D.2.1 Excess Short Term Demand

Short Run In the short run, consider the model at time 0^+ when everyone still has the same holding of $\bar{a}$. We take a random sample so the supply is

$$\left(N_1\rho_1 + N_2\rho_2\right)\bar{a}$$

Suppose that the demand is given by the long term time invariant functions. The demand is then

$$\frac{1}{2}\rho_1\left(N_1 - \delta_1\right) + \frac{1}{2}\rho_2 N_2$$

and the flow market clearing condition in the short run is

$$\frac{\rho_1\left(N_1 - \delta_1\right) + \rho_2 N_2}{2} = \left(N_1\rho_1 + N_2\rho_2\right)\bar{a} \tag{46}$$

Long Run In the long run, we have the same dynamics of α as in the one venue case. In venue one we have fraction $\alpha_-^1\left(1\right)$ who want to sell and fraction $\alpha_+^1\left(0\right)$ who want to buy and they are

equal:

$$\alpha_{-}^{1}(1) = \alpha_{+}^{1}(0) = \frac{1}{4}\frac{\gamma}{\gamma+\rho_1}$$

So in fact we have clearing in both markets, and we can imagine in the long run that slow buyers buy only from slow sellers (and fast from fast). This is another way of saying that assets do not migrate from the slow to the fast venue anymore. The assets held in the fast venue are

$$N_2\left(\alpha_{-}^{2}(1) + \alpha_{-}^{2}(1)\right) = \frac{N_2}{4}\left(\frac{\gamma}{\gamma+\rho_2} + \frac{2\rho_2+\gamma}{\gamma+\rho_2}\right) = \frac{N_2}{2}$$

and assets held in the slow venue are

$$N_2\left(\alpha_{-}^{1}(1) + \alpha_{-}^{1}(1)\right) = \frac{N_1-\delta_1}{2}$$

The market clearing in stock in the long run is

$$\frac{N_1-\delta_1+N_2}{2} = (N_1+N_2)\,\bar{a}, \tag{47}$$

which is identical to the condition used in the paper: $\frac{1-G(\hat{\sigma}_1)}{2} = (\delta_1 + 1 - G(\hat{\sigma}_1))\,\bar{a}$.

Excess Demand The point is that (46) and (47) are inconsistent with each other. More precisely, suppose, as we do in the paper, that (47) holds. Then you can compute the excess short run demand as

$$\Delta = \frac{1}{2}\rho_1(N_1-\delta_1) + \frac{1}{2}\rho_2 N_2 - (N_1\rho_1 + N_2\rho_2)\,\bar{a}$$

which we can write as

$$\Delta = \frac{1}{2}\rho_1(N_1-\delta_1+N_2) + \frac{1}{2}(\rho_2-\rho_1)N_2 - ((N_1+N_2)\rho_1 + N_2(\rho_2-\rho_1))\,\bar{a}$$

and using (47) we get

$$\Delta = \left(\frac{1}{2} - \bar{a}\right)(\rho_2-\rho_1)N_2$$

as a measure of short run excess demand conditional on long run market clearing. We can see why there is a gap simply by rewriting (46) and (47) as

$$\begin{aligned} \frac{1}{2} - \frac{1}{N_1 + \frac{\rho_2}{\rho_1}N_2}\frac{\delta_1}{2} &= \bar{a} \\ \frac{1}{2} - \frac{1}{N_1+N_2}\frac{\delta_1}{2} &= \bar{a} \end{aligned}$$

When $\frac{\rho_2}{\rho_1} > 1$, we over-sample the fast venue where demand is high and supply is low. Let us now compute the equilibrium transition dynamics

D.3 Computing the transition dynamics with price protection

The total supply of assets to the market is

$$(N_1 + N_2)\,\bar{a}$$

which consists of

- $N_2 = 1 - G(\hat{\sigma}_2)$ permanent traders in fast venue, with long run flow-demand is $\rho_2 \frac{N_2}{2}$
- $N_1 = \delta_1 + G(\hat{\sigma}_2) - G(\hat{\sigma}_1)$ temporary and permanent traders in the slow venue, with long run demand $\rho_1 \frac{N_1 - \delta_1}{2}$

But we know that there is excess demand at time 0, so **some of the permanent traders in the slow venue do not buy at time 0**. **Let $\tilde{\sigma}_{1,t} \geq \hat{\sigma}_1$ be the time-varying marginal buyer**. We assume here that there is an interior solution for $\tilde{\sigma}_{1,t}$. The analysis easily extend to the case of a corner solution.

The flow-demand from investors in the slow venue is then $\rho_1 \frac{\tilde{N}_{1,t}}{2}$ where

$$\tilde{N}_{1,t} \equiv G(\hat{\sigma}_2) - G(\tilde{\sigma}_{1,t})$$

On the other hand, we know that assets migrate over time from the slow to the fast venue. **Let m_t measure the net stock of migrated assets**, so that total assets held in venue 2 are

$$N_2 \bar{a} + m_t,$$

and $N_1 \bar{a} - m_t$ in venue 1.

Market clearing then requires that the flow demand equals the flow supply:

$$\begin{aligned} \rho_2 \frac{N_2}{2} + \rho_1 \frac{\tilde{N}_{1,t}}{2} &= \rho_1 (N_1 \bar{a} - m_t) + \rho_2 (N_2 \bar{a} + m_t) \\ &= \rho_1 N_1 \bar{a} + \rho_2 N_2 \bar{a} + m_t (\rho_2 - \rho_1) \end{aligned}$$

Next we need to find the low of motion for m_t. To do so, consider the flows from and to venue 2. Venue 2 traders sell (in gross terms) $\rho_2 (N_2 \bar{a} + m_t)$ and they buy (in gross terms) $\rho_2 \frac{N_2}{2}$ so the net migration is

$$\begin{aligned} \frac{dm}{dt} &= \rho_2 \frac{N_2}{2} - \rho_2 (N_2 \bar{a} + m_t) \\ &= \rho_2 N_2 \left(\frac{1}{2} - \bar{a} \right) - \rho_2 m_t \end{aligned}$$

Given the initial condition $m_0 = 0$, the solution of this differential equation is

$$m(t) = N_2 \left(\frac{1}{2} - \bar{a} \right) \left(1 - e^{-\rho_2 t} \right)$$

We can then use the market clearing condition to compute $\tilde{N}_{1,t}$:

$$\frac{\tilde{N}_{1,t}}{2} = (N_1 + N_2)\,\bar{a} - \frac{1}{2}N_2 - N_2\left(\frac{1}{2} - \bar{a}\right)\frac{\rho_2 - \rho_1}{\rho_1}e^{-\rho_2 t}$$

which using $\frac{N_1 - \delta_1 + N_2}{2} = (N_1 + N_2)\,\bar{a}$, we can also write as

$$\tilde{N}_{1,t} = N_1 - \delta_1 - N_2\,(1 - 2\bar{a})\,\frac{\rho_2 - \rho_1}{\rho_1}e^{-\rho_2 t}$$

Note that it satisfies the correct conditions:

- in the long run we have $\tilde{N}_{1,\infty} = N_1 - \delta_1 = G\,(\hat{\sigma}_2) - G\,(\hat{\sigma}_1)$, or equivalently

$$\tilde{\sigma}_{1,\infty} = \hat{\sigma}_1$$

- in the short run $\tilde{N}_{1,t} = N_1 - \delta_1 - N_2\,(1 - 2\bar{a})\,\frac{\rho_2 - \rho_1}{\rho_1}$ which shows the short run imbalance described earlier.

In terms of the time-varying marginal trading type we have

$$G\,(\tilde{\sigma}_{1,t}) = G\,(\hat{\sigma}_1) + N_2\,(1 - 2\bar{a})\,\frac{\rho_2 - \rho_1}{\rho_1}e^{-\rho_2 t}$$

which gives us the path of convergence of $\tilde{\sigma}_{1,t}$ to $\hat{\sigma}_1$. We can then back out the price from the indifference condition

$$\bar{u}\,(a; \tilde{\sigma}_{1,t}, 1) = r p_t a$$

or

$$p_t = \frac{\mu}{r} + \frac{\tilde{\sigma}_{1,t}}{r}\frac{r + \rho_1}{r + \gamma + \rho_1}$$

Uniform distribution Assuming a uniform distribution of types, we have

$$\tilde{\sigma}_{1,t} = \hat{\sigma}_1 + \bar{\sigma} N_2\,(1 - 2\bar{a})\,\frac{\rho_2 - \rho_1}{\rho_1}e^{-\rho_2 t}$$

and

$$p\,(t) = \bar{p} + \kappa e^{-\rho_2 t}$$

where

$$\bar{p} = \frac{\mu}{r} + \frac{\hat{\sigma}_1}{r}\frac{r + \rho_1}{r + \gamma + \rho_1}$$

and

$$\kappa = \bar{\sigma} N_2\,(1 - 2\bar{a})\,\frac{\rho_2 - \rho_1}{\rho_1 r}\frac{r + \rho_1}{r + \gamma + \rho_1}$$

Consider now the value functions for the temporary traders given price p

$$rW_t = \mu a + \rho_1 \left(p_t a - W_t\right) + \frac{\partial W}{\partial t}$$

It is of the form

$$W\left(t\right) = A + Be^{-\rho_2 t}$$

where

$$r\left(A + Be^{-\rho_2 t}\right) = \mu a + \rho_1 \left(\bar{p}a + a\kappa e^{-\rho_2 t} - A - Be^{-\rho_2 t}\right) - \rho_2 Be^{-\rho_2 t}$$

or

$$\begin{aligned} A &= \frac{\mu + \rho_1 \bar{p}}{r + \rho_1} \bar{a} \\ B &= \frac{\rho_1}{r + \rho_1 + \rho_2} \kappa \bar{a} \end{aligned}$$

We can see that A is just like $\tilde{W}$ in our original formula (appendix proof of prop 1). The time varying part $Be^{-\rho_2 t}$ is the new

$$\begin{aligned} B &= \frac{\rho_1}{r + \rho_1 + \rho_2} \kappa \bar{a} \\ &= \frac{\bar{\sigma}}{r} N_2 \left(1 - 2\bar{a}\right) \frac{r + \rho_1}{r + \gamma + \rho_1} \frac{\rho_2 - \rho_1}{r + \rho_1 + \rho_2} \bar{a} \end{aligned}$$

We can compare the exact and approximate solutions

$$\frac{W^{exact}\left(0\right) - W^{approx}}{W^{approx} - W^{autarky}} = \frac{B}{A - \frac{\mu}{r}} = \frac{\bar{\sigma}}{\hat{\sigma}_1} N_2 \left(1 - 2\bar{a}\right) \frac{r + \rho_1}{\rho_1} \frac{\rho_2 - \rho_1}{r + \rho_1 + \rho_2}$$

Take the case where $\frac{1}{2}\rho_2 = \rho_1 \gg r$ (note that we have assumed an interior solution for $\tilde{\sigma}_{1,t}$ so we cannot take ρ_2/ρ_1 to be too large, or we would need to use the formula for the corner solution which is less interesting, and in any case does not change the main point). Then we make an approximation of the order

$$\frac{1}{3} \frac{N_2 \left(1 - 2\bar{a}\right)}{N_1 - \delta_1}$$

So with long run participation of 60 in the fast venue, 30 in the slow one, and $\bar{a} = 0.45$, we get an approximation of $2 * 0.1/3 = 6.66\%$ in our value functions.

Appendix E Trading Fees

In this section we derive the equilibrium when venues charge a trading fee ϕ per unit of trading. We consider fees paid at execution (initiation/canceling fees yield similar results). 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.

Trading fees create a new class of agents, between those who trade repeatedly and those who sell and drop out. Formally, we have three regions:

1. If $\sigma \geq \tilde{\sigma}$, repeat trades
2. If $\sigma \in (\check{\sigma}, \tilde{\sigma})$, sell and drop out if $\epsilon = -1$, but keep if $\epsilon = +1$. This is the new "wait-to-sell" strategy
3. If $\sigma \leq \check{\sigma}$, sell irrespective of ϵ. This is the strategy in the case $\phi = 0$ that gives flat value functions.

E.1 Value Functions

Heavy Traders

Consider the steady state value functions for any type $\sigma > \tilde{\sigma}$, i.e. for types that trade repeatedly despite the fees. The value of the marginal type $\tilde{\sigma}$ is affected by the trading fee, as explained below. As before, define the value of ownership as $I_\sigma^\epsilon \equiv V_{\sigma,\epsilon}(1) - V_{\sigma,\epsilon}(0)$. We have

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

The equations for $V_{\sigma,-}(0)$ and $V_{\sigma,+}(1)$ 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 σ in venue ρ become:

$$I_\sigma^+ - I_\sigma^- = 2\frac{\sigma + \rho\phi}{r + \gamma + \rho}. \tag{48}$$

Then we can solve

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

The average value (across ϵ) for types $\sigma \geq \tilde{\sigma}$ is therefore

$$\begin{aligned} \bar{V}_\sigma(0) &= \frac{\rho}{2r}\left(I_\sigma^+ - p - \phi\right) \\ \bar{V}_\sigma(1) &= \frac{\mu}{r} + \frac{\rho}{2r}\left(p - \phi - I_\sigma^-\right) \end{aligned} \tag{49}$$

Marginal Type and ex-ante Value for Heavy Trades We define the marginal type $\tilde{\sigma}$ 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 $I_{\tilde{\sigma}}^{+} = p + \phi$, therefore:

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

Let us now compute the ex ante value functions. For the marginal type $\tilde{\sigma}$, we have the ex-ante value function $\tilde{W}$ that solves:

$$\tilde{W} = \bar{a}\bar{V}_{\tilde{\sigma}}(1) + (1-\bar{a})\bar{V}_{\tilde{\sigma}}(0) \tag{50}$$

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

$$\tilde{W}(\phi) = \frac{\mu\bar{a}}{r} + \frac{\bar{a}s}{r}(\tilde{\sigma} - (r+\gamma)\phi).$$

For all types $\sigma > \tilde{\sigma}$, we obtain, taking 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

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

Notice that the value depends on the fee via $-\frac{\rho}{2r}\frac{r+\gamma}{r+\gamma+\rho}\phi$, which is the NPV of the fees paid by the repeat traders. A convenient way to express the value function is

$$W(\sigma,\rho,\phi) = \tilde{W}(\phi) + \frac{s}{2r}(\sigma - \tilde{\sigma}(\phi)) \quad \text{for all}\, \sigma \geq \tilde{\sigma}.$$

Infra-marginal types

Let us now consider types $\sigma < \tilde{\sigma}$. As before, they join the venue to sell but the key difference is that they do not necessarily sell all the time. In fact, some sell only when $\epsilon = -1$ and keep the asset when $\epsilon = +1$. That did not happen without trading fees, since in that case $I_{\tilde{\sigma}}^{+} = p$ implied $V_{\tilde{\sigma},+,1} - V_{\tilde{\sigma},+,0} = p$ so type $(\tilde{\sigma}, +)$ was indifferent to buying starting from $a = 0$ *and* to selling starting from $a = 1$. With trading fees, we have $I_{\tilde{\sigma}}^{+} = p + \phi$, so type type $(\tilde{\sigma}, +)$ is indifferent to buying, but strictly prefers to keep the asset instead of selling it at price $p - \phi$. This is the key point of trading fees: there is now a difference between *keeping* the asset and *buying* the asset. Since the types $\sigma < \tilde{\sigma}$ never want to buy, we have $V(0) = 0$.

One complication that arises here is that we cannot guarantee market clearing with a constant

price without introducing a market maker. More precisely, if n investors join the exchange at time 0, the gross demand for the asset is still $n\frac{1-G(\tilde{\sigma})}{2}$ but the gross supply is $n\left(\bar{a}-\breve{\alpha}_t\right)$ where $\breve{\alpha}_t$ is the number of traders with $a_t = 1$, $\epsilon_t = 1$ and $\sigma \in (\breve{\sigma}, \tilde{\sigma})$. We know that initially $\breve{\alpha}_0 = \frac{n\bar{a}}{2}\left(G\left(\tilde{\sigma}\right) - G\left(\breve{\sigma}\right)\right)$. But over time they sell and eventually $\breve{\alpha}_t \to 0$. The long run market clearing is $\frac{1-G(\tilde{\sigma})}{2} = \bar{a}$ but during the transition there is excess demand. Studying price dynamics in that case is clearly beyond the scope of this paper, so we simply assume the presence of competitive market makers who undo the temporary imbalance. Note that this is risk-less since the evolution of the market imbalance is perfectly predictable.

Traders who sell irrespective of ϵ. The value functions in this case are

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

so

$$V_{\sigma,+}\left(1\right) - V_{\sigma,-}\left(1\right) = \frac{2\sigma}{r+\gamma+\rho}$$

and

$$V_{\sigma,+}\left(1\right) = \frac{\mu + \rho\left(p-\phi\right)}{r+\rho} + \frac{\sigma}{r+\gamma+\rho}$$

and

$$V_{\sigma,-}\left(1\right) = \frac{\mu + \rho\left(p-\phi\right)}{r+\rho} - \frac{\sigma}{r+\gamma+\rho}$$

Define the marginal type $\breve{\sigma}$ who is indifferent to selling when $\epsilon = +1$. It solves $V_{\breve{\sigma},+}\left(1\right) = p - \phi$ so

$$\breve{\sigma}\left(\phi\right) = \left(1 + \frac{\gamma}{r+\rho}\right)\left(r\left(p-\phi\right) - \mu\right)$$

So that's a curve in the (σ, p) space that is parallel to the $\tilde{\sigma}$ curve, but higher by ϕ.

Types below $\breve{\sigma}$ sell irrespective of their types. On average this yields

$$\frac{V_{\sigma,+}\left(1\right) + V_{\sigma,-}\left(1\right)}{2} = \frac{\mu + \rho\left(p-\phi\right)}{r+\rho}$$

so since $V\left(0\right) = 0$, we have for $\sigma \leq \breve{\sigma}$:

$$\begin{aligned} \breve{W}\left(\phi\right) = \bar{a}\bar{V}_{\sigma}\left(1\right) &= \bar{a}\frac{\mu + \rho\left(p-\phi\right)}{r+\rho} \\ &= \bar{a}\frac{\mu}{r} + \bar{a}\frac{\rho}{r+\rho}\left(p - \phi - \frac{\mu}{r}\right) \\ &= \bar{a}\frac{\mu}{r} + \bar{a}s\frac{\breve{\sigma}\left(\phi\right)}{r}. \end{aligned}$$

Note that this is a flat value function. It does not depend on σ.

Wait-to-sell The types $\sigma \in (\breve{\sigma}, \tilde{\sigma})$, sell if $\epsilon = -1$, but keep if $\epsilon = +1$. This is the new "wait to sell" strategy. The Bellman equations are

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

Let us define the value functions relative to the previous ones

$$\tilde{V}_{\sigma,\epsilon} = V_{\sigma,\epsilon} + x_\epsilon$$

Then we have

$$\begin{aligned} rx_- &= +\frac{\gamma}{2}(x_+ - x_-) - \rho x_- \\ rx_+ &= -\frac{\gamma}{2}(x_+ - x_-) + \rho(V_{\sigma,+} - p + \phi) \end{aligned}$$

so

$$x_+ = \frac{\frac{\gamma}{2}x_- + \rho(V_{\sigma,+} - p + \phi)}{r + \frac{\gamma}{2}}$$

and

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

or

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

so

$$\begin{aligned} rx_- + rx_+ &= \rho(V_{\sigma,+} - p + \phi - x_-) \\ &= \rho\left(V_{\sigma,+} - p + \phi - \frac{\frac{\gamma}{2}\rho(V_{\sigma,+} - p + \phi)}{r\left(r + \gamma + \rho + \rho\frac{\gamma}{2r}\right)}\right) \\ \frac{x_- + x_+}{2} &= \frac{\rho}{2r}\frac{r + \gamma + \rho}{r + \gamma + \rho + \rho\frac{\gamma}{2r}}(V_{\sigma,+} - p + \phi). \end{aligned}$$

Therefore

$$\begin{aligned} W(\sigma, \rho, \phi) &= \bar{a}\mathbb{E}\tilde{V}_{\sigma,\epsilon}(1) \\ &= \breve{W} + \bar{a}\frac{\rho}{2r}\frac{r + \gamma + \rho}{r + \gamma + \rho + \rho\frac{\gamma}{2r}}(V_{\sigma,+} - p + \phi) \\ &= \breve{W} + \bar{a}\frac{\rho}{2r}\frac{r + \gamma + \rho}{r + \gamma + \rho + \rho\frac{\gamma}{2r}}(V_{\sigma,+} - p + \phi) \end{aligned}$$

Recall the definitions $\breve{\sigma}(\phi) = \frac{r+\gamma+\rho}{r+\rho}(r(p-\phi) - \mu)$ and $V_{\sigma,+} = \frac{\mu+\rho(p-\phi)}{r+\rho} + \frac{\sigma}{r+\gamma+\rho}$. Thus we have

$$V_{\sigma,+} - p + \phi = \frac{\sigma}{r+\gamma+\rho} + \frac{\mu - r(p-\phi)}{r+\rho} = \frac{\sigma - \breve{\sigma}(\phi)}{r+\gamma+\rho}.$$

And this leads to:

$$W(\sigma, \rho, \phi) = \breve{W}(\phi) + \frac{\bar{a}\rho}{2r} \frac{\sigma - \breve{\sigma}(\phi)}{r+\gamma+\rho+\frac{\gamma\rho}{2r}}.$$

We can also summarize the dependence on ϕ as

$$\frac{\bar{a}\rho}{2r} \frac{r+\gamma}{r+\gamma+\rho+\frac{\gamma\rho}{2r}} \phi$$

Summary

The model with execution fee is characterized by:

1. Low types $\sigma \leq \breve{\sigma}$ have average value

$$\breve{W}(\breve{\sigma}) = \frac{\bar{a}\mu}{r} + \frac{\bar{a}s}{r}\breve{\sigma}(\phi)$$

where the marginal selling type is defined by $V_{\breve{\sigma},+}(1) = p - \phi$ as

$$\breve{\sigma}(p, \phi) = \left(1 + \frac{\gamma}{r+\rho}\right)(rp - \mu - r\phi)$$

2. Wait to sell types $\sigma \in [\breve{\sigma}, \tilde{\sigma}]$ have value

$$W(\sigma, \breve{\sigma}) = \breve{W}(\breve{\sigma}) + \bar{a}\frac{s}{2r+\gamma s}(\sigma - \breve{\sigma})$$

3. Repeat traders $\sigma \geq \tilde{\sigma}$ where the marginal buying type is defined by $I^{+}_{\tilde{\sigma}} = p + \phi$ as

$$\tilde{\sigma}(p, \phi) = \left(1 + \frac{\gamma}{r+\rho}\right)(rp - \mu) + (r+\gamma)\phi$$

with

$$\tilde{W}(p) = \frac{\bar{a}\mu}{r} + \frac{\bar{a}s}{r}(\tilde{\sigma} - (r+\gamma)\phi) = \frac{\bar{a}\mu}{r} + \frac{\bar{a}}{r}\frac{\rho}{r+\rho}(rp - \mu)$$

have value

$$W(\sigma, \tilde{\sigma}, \phi) = \tilde{W}(p) + \frac{s}{2r}(\sigma - \tilde{\sigma}(\phi))$$

So the overall value function is

$$W(\sigma, \breve{\sigma}, \tilde{\sigma}, \phi) = \breve{W}(\breve{\sigma}) + \bar{a}\frac{s}{2r+\gamma s}(\sigma - \breve{\sigma})\mathbb{I}_{(\breve{\sigma},\tilde{\sigma})} + \left\{\frac{\bar{a}s}{r}(\tilde{\sigma} - \breve{\sigma} - (r+\gamma)\phi) + \frac{s}{2r}(\sigma - \tilde{\sigma})\right\}\mathbb{I}_{\sigma \geq \tilde{\sigma}}$$

and note that the marginal types are both pinned down by the surplus price $p - \frac{\mu}{r}$, so they are related by:

$$\breve{\sigma} = \tilde{\sigma} - (r+\gamma)\,\phi - r\left(1 + \frac{\gamma}{r+\rho}\right)\phi$$
$$\breve{\sigma} = \tilde{\sigma} - 2r\frac{\rho}{r+\rho}\left(\frac{1}{s} + \frac{\gamma}{2r}\right)\phi$$

As expected the gap between the marginal-repeat-trader type $\tilde{\sigma}$ and the marginal-wait-to-sell type $\breve{\sigma}$ increases with the trading fee ϕ. Without trading fees, we recover our benchmark case where there is only one marginal trading type since $\breve{\sigma} = \tilde{\sigma}$ when $\phi = 0$.

E.2 Venue's Program

Suppose there is only one venue. The sequence of events and the structure of the equilibrium are as follows:

1. The venue sets (q, ϕ) and traders decide to join, or not. Traders above some cutoff $\hat{\sigma}$ enter. So the mass of traders is $1 - G(\hat{\sigma})$.

2. Trading takes place. The assets migrate towards the high σ types, while the low σ one drop out. There is a marginal trading type $\tilde{\sigma}$ which must satisfy the market clearing condition

$$1 - G(\tilde{\sigma}) = 2\bar{a}\,(1 - G(\hat{\sigma}))\,.$$

   This implies that, in equilibrium, $\tilde{\sigma}$ is determined by $\hat{\sigma}$. The price p adjusts to ensure that $\tilde{\sigma}$ is indeed the marginal trading type:

$$\tilde{\sigma} = \left(1 + \frac{\gamma}{r+\rho}\right)(rp - \mu) + (r+\gamma)\,\phi$$

3. Given the price, we can solve for all the value functions. The free entry condition is then

$$W(\hat{\sigma}, \rho, \phi) = W_{out} + q$$

Profits Investors above some cutoff $\hat{\sigma}$ decide to participate. They split between δ temporary traders and $1 - G(\tilde{\sigma})$ repeat traders with

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

Total profits of the exchange are

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

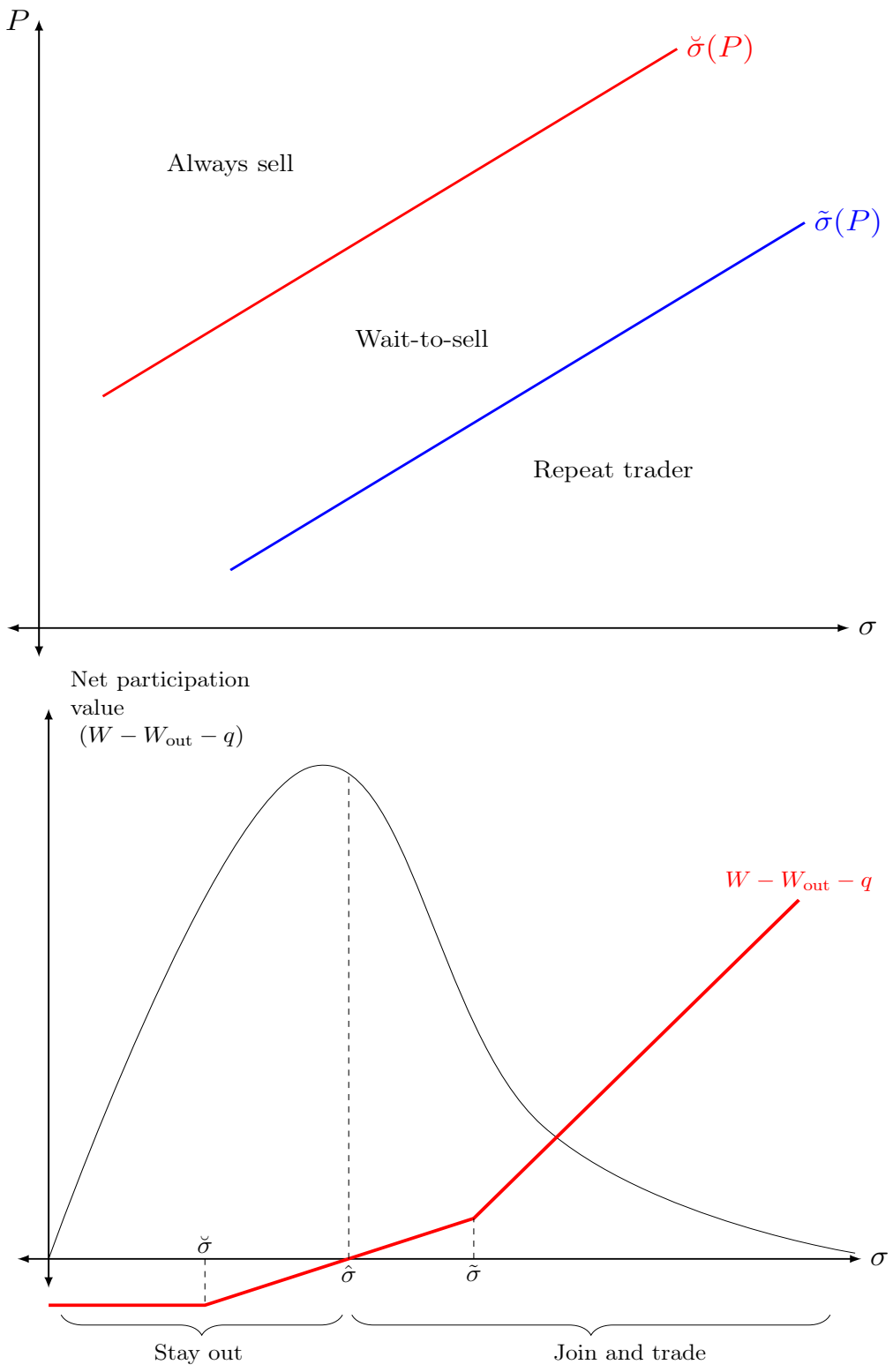

Figure E4. Trading strategies and value functions with trading fees.

where π^{ϕ} denote the value of trading fees. These fees come from temporary and permanent traders. There are δ investors who only trade when their type is low. Let π^{ϵ} be the value for the exchange of trading fees paid by type ϵ. We have

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

Therefore

$$r\left(\pi^{+} + \pi^{-}\right) = \rho\left(\phi - \pi^{-}\right)$$

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

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

which corresponds to the NPV $\frac{\bar{a}\rho}{2r}\frac{r+\gamma}{r+\gamma+\rho+\frac{\gamma\rho}{2r}}\phi$ paid by these wait-to-sell traders. Similarly, for the permanent investors, we can see the value of the fees from the value function: $\frac{\rho}{2r}\frac{r+\gamma}{r+\gamma+\rho}\phi$. Therefore the NPV of trading fees is

$$\pi^{\phi} = \left(1 - G\left(\tilde{\sigma}\right)\right)\frac{\rho}{2r}\frac{r+\gamma}{r+\gamma+\rho}\phi + \pi^{\phi}_{\delta}$$

and from the definition of δ, we have

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

Total profits are then

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

The indifference/participation condition for the marginal type $\hat{\sigma}$, assuming that $\hat{\sigma} \in [\breve{\sigma}, \tilde{\sigma}]$ is

$$\breve{W}\left(\breve{\sigma}\right) + \bar{a}\frac{s}{2r+\gamma s}\left(\hat{\sigma} - \breve{\sigma}\right) = W_{out} + q$$

which implies

$$q = \frac{\bar{a}s}{r}\breve{\sigma} + \bar{a}\frac{s}{2r+\gamma s}\left(\sigma^{e} - \breve{\sigma}\right)$$

Program The venue solves the following program:

$$\max_{q,\phi}\pi^{TOT} \equiv \left(1 - G\left(\tilde{\sigma}\right)\right)\left(\frac{q}{2\bar{a}} + \frac{\phi}{2r}\left(r+\gamma\right)\left(s + \left(\frac{1}{2} - \bar{a}\right)\frac{s}{1+\frac{\gamma s}{2r}}\right)\right)$$

subject to

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

Let us first understand the tradeoff between q and ϕ. Note that we can rewrite the last two constraints as

$$\breve{\sigma} = \tilde{\sigma}-\frac{2r}{s}\frac{\rho}{r+\rho}\left(1+\frac{\gamma s}{2r}\right)\phi$$

and

$$q = \frac{\bar{a}s}{2r}\frac{\hat{\sigma}}{1+\frac{\gamma s}{2r}}+\frac{\bar{a}s}{2r}\left(\frac{1+\frac{\gamma s}{r}}{1+\frac{\gamma s}{2r}}\right)\breve{\sigma}$$

and combine them to get

$$\frac{q}{\bar{a}} = \frac{s}{r}\frac{\hat{\sigma}+\left(1+\frac{\gamma s}{r}\right)\tilde{\sigma}}{2+\frac{\gamma s}{r}}-\left(1+\frac{\gamma s}{r}\right)\frac{\rho}{r+\rho}\phi$$

This then implies that

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

This shows that, holding $\tilde{\sigma}$ and $\hat{\sigma}$ constant, profits increase when ϕ increases relative to q. In other words, if we consider an iso-participation change in the composition of fees. Formally, keep $\hat{\sigma}$ constant. From the first constraint, this means holding $\tilde{\sigma}$ constant. Then consider how ϕ and q must change.

$$\partial\breve{\sigma} = -\frac{2r}{s}\frac{\rho}{r+\rho}\left(1+\frac{\gamma s}{2r}\right)\partial\phi$$

and

$$\begin{aligned}\partial q &= \frac{\bar{a}s}{2r}\left(2-\frac{1}{1+\frac{\gamma s}{2r}}\right)\partial\breve{\sigma}\\ &= -\bar{a}\left(1+\frac{\gamma s}{r}\right)\frac{\rho}{r+\rho}\partial\phi\end{aligned}$$

Then profits actually increase if ϕ goes up while q goes down. The intuition is that ϕ allows for better *price discrimination*. The repeat traders pay the fee repeatedly, while the temporary traders pay the fee only once. Since the repeat traders value participation more, the trading fee is better able to extract the surplus from the traders than the participation fee.

However, notice two important points:

- This would not change our result that price protection increases the profits of the slow venue.
- The differences between ϕ and q vanishes when $\bar{a}$ is close to $1/2$. Therefore our analysis of endogenous speed is unchanged.

Summary. The following proposition summarizes the results on the trading fees extension of the model.

Proposition 9. *In a market with trading fees, the following equilibrium conditions hold:*
(i) There is still a marginal type $\tilde{\sigma}$ above which traders trade repeatedly and another marginal type $\breve{\sigma} < \tilde{\sigma}$ below which traders sell, irrespective of their trading type ϵ. Trading fees, however, create a region of partial inaction for types $\sigma \in (\breve{\sigma}, \tilde{\sigma})$: They sell when $\epsilon = -1$ but, when $\epsilon = +1$, they wait until their trading type switches to $\epsilon = -1$.
(ii) Trading fees improve price discrimination and can increase average profits of venues, but they do not affect our results (i)-(ii)-(iii) in Proposition 2.

Appendix F Multi-venue Traders

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

Bellman equations.

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

$$\begin{aligned} rV^{MV}_{\sigma,+}(0) &= \frac{\gamma}{2}\left[V^{MV}_{\sigma,-}(0) - V^{MV}_{\sigma,+}(0)\right] + \rho_1\left(I^{MV}_{\sigma,+} - p_1\right) + \rho_2\left(I^{MV}_{\sigma,+} - p_2\right) \\ rV^{MV}_{\sigma,-}(1) &= \mu - \sigma + \frac{\gamma}{2}\left[V^{MV}_{\sigma,+}(1) - V^{MV}_{\sigma,-}(1)\right] + \rho_1\left(p_1 - I^{MV}_{\sigma,-}\right) + \rho_2\left(p_2 - I^{MV}_{\sigma,-}\right) \end{aligned}$$

and

$$\begin{aligned} rV^{MV}_{\sigma,-}(0) &= \frac{\gamma}{2}\left[V^{MV}_{\sigma,+}(0) - V^{MV}_{\sigma,-}(0)\right] \\ rV^{MV}_{\sigma,+}(1) &= \mu + \sigma + \frac{\gamma}{2}\left[V^{MV}_{\sigma,-}(1) - V^{MV}_{\sigma,+}(1)\right] \end{aligned}$$

The key issue is whether MVTs send both buy and sell orders in both venues. This happens if and only if $I^{MV}_{\sigma,+} > p_2$ and $p_1 > I^{MV}_{\sigma,-}$, where the values of ownership are defined as before, and the

Bellman equations for MVT are equivalent to one venue 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

$$I^{MV}_{\sigma,+} - I^{MV}_{\sigma,-} = \frac{2\sigma}{r + \gamma + \rho_1 + \rho_2}.$$

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

The second key point is that we must check that MVTs 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 $I^{MV}_{\sigma,+} > p_2$ and $p_1 > I^{MV}_{\sigma,-}$. These conditions are equivalent to $\sigma > \sigma^{MV}_{buy}$ and $\sigma > \sigma^{MV}_{sell}$, where we define two marginal types

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

and

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

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

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

This implies the following Lemma.

Lemma 11. *The price difference cannot be higher than the gains from trade of the lowest MVTs.*

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

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

which, using the equilibrium conditions, we can write as

$$\hat{\sigma}_{MV} \equiv \frac{\frac{\rho_1}{r+\rho_1}}{\frac{\rho_1+\rho_2}{r+\gamma+\rho_1+\rho_2} - \frac{\rho_2}{r+\gamma+\rho_2}} \left(r p_1 - \mu\right)$$

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

The possibility of MVTs 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. Using our benchmark calibration, either for uniform or exponential distribution of types, we verify that, for any $\sigma' \in \left\{\hat{\sigma}_{MV}, \sigma^{MV}_{buy}, \sigma^{MV}_{sell}\right\}$, we have $1 - G\left(\sigma'\right) \approx 0$. Therefore, in our model, the equilibrium does not change if we allow for MV trading.

Appendix G Analysis of a Constrained Planner

We have seen in Section 4 the solution to the unconstrained planner problem. A more interesting case is when external subsidies are ruled out. The planner can still decide entry, speed, and pricing but we require that trading venues break even.

Let us first perform the analysis for a given value of s_1. The program is

$$\max_{s_2,q_1,q_2} \mathcal{W},$$

subject to the break-even constraint

$$q_2 \left(1 - G\left(\hat{\sigma}_2\right)\right) \geq C\left(s_2\right).$$

Interestingly, we find that the planner still chooses a single venue.

Lemma 12. *The Planner subject to break-even constraints chooses one venue with higher participation than under monopoly.*[60]

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. To understand the intuition, it is better to think of $\hat{\sigma}_1$ and $\hat{\sigma}_2$ as control variables instead of q_1 and q_2. We show in the appendix that the Lagrangian of the planner's problem is

$$\begin{aligned}\mathcal{L}\left(s\right) = s_1 \int_{\hat{\sigma}_1}^{\overline{\sigma}} \sigma dG\left(\sigma\right) + \left(s - s_1\right) \int_{\hat{\sigma}_2}^{\overline{\sigma}} \sigma dG\left(\sigma\right) - 2rC\left(s\right) \\ + \lambda \left\{\left(\left(s - s_1\right)\hat{\sigma}_2 + s_1\sigma_1\right)\left(1 - G\left(\hat{\sigma}_2\right)\right) - 2rC\left(s\right)\right\}\end{aligned}$$

where λ is the multiplier of the budget constraint of the fast venue, and we have replaced $q_2 = \left(s - s_1\right)\hat{\sigma}_2 + s_1\hat{\sigma}_1$. The welfare cost of raising $\hat{\sigma}_1$ is $s_1\hat{\sigma}_1 g\left(\hat{\sigma}_1\right)$, and the financing gain is $\lambda s_1\left(1 - G\left(\hat{\sigma}_2\right)\right)$. 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 $\hat{\sigma}_2$. In other words, the slow venue 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 $\kappa = 0$).

Obviously the result is the same if the planner also chooses s_1, and it extends to the case where prices in the venues can be consolidated.

[60] When the break-even constraint binds, it is possible for the planner to choose a lower trading speed than the unconstrained monopolist (Example 1). Intuitively, when the distribution of permanent types has a fat right tail, the monopolist might choose to target 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. Note that in this case the planner trades off speed against participation. For a given participation, the planner always favors higher speed.

Proof. In general, the objective function of the planner is

$$\max_{s_2,q_1,q_2} \frac{s_1}{2r}\int_{\sigma_1}^{\hat{\sigma}_2} \sigma dG(\sigma) + \frac{s_2}{2r}\int_{\hat{\sigma}_2}^{\overline{\sigma}} \sigma dG(\sigma) - C(s_2)$$

and the marginal types are given by 16 and 17, so we have

$$q_1 = s_1\frac{\hat{\sigma}_1}{2r},$$
$$q_2 = (s_2 - s_1)\frac{\hat{\sigma}_2}{2r} + q_1.$$

The break-even constraint is $q_2(1 - G(\hat{\sigma}_2)) \geq C(s_2)$, so the Lagrangian (scaled by $2r$) is

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

and the FOCs of the planner problem are

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

Optimal speed satisfies

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

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

This result can be extended to the case where the planner operates the two venues with one budget constraint. In this case, the constraint is $(G(\hat{\sigma}_2) - G(\hat{\sigma}_1))q_1 + (1 - G(\hat{\sigma}_2))q_2 > C(s_2)$ and the Lagrangian is

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

and the FOCs for affiliations are

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

Optimal speed satisfies the same equation as before. In this case, we see that $\sigma_1^* = \sigma_2^*$, venue 1 is still inactive.

With one venue, the Lagrangian of the planner is

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

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

$$2rC'(s^*) = \frac{1}{1+\lambda}\int_{\sigma^*}^{\overline{\sigma}} \sigma dG(\sigma) + \frac{\lambda}{1+\lambda}\left(1 - G(\sigma^*)\right)\sigma^*,$$

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, as shown in the following example.

Example 1. Case where $s_m > s^*$

We provide a simple example to show that it is indeed possible for the monopoly to choose a higher speed that the planner. 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 $W = \rho\bar{\sigma}\left((1-n)\alpha + n\right) - c(s)$. 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. *Q.E.D.*

Appendix H Details on the Calibration

H.1 Implied γ and c in the duopoly case

We describe in this section the methodology to compute the values of γ_k and c_k in the duopoly case. Under a uniform distribution of types, and scaling the number of transactions by the number of investors N_k, the duopoly volume formula becomes

$$\mathcal{V}_k = N_k \times \frac{\gamma_k}{4}\left[\frac{\rho_{1,k}}{\gamma_k+\rho_{1,k}}\left(\frac{\hat{\sigma}_{2,k}-\hat{\sigma}_{1,k}}{\overline{\sigma}}\right)+\frac{\rho_{2,k}}{\gamma_k+\rho_{2,k}}\left(1-\frac{\hat{\sigma}_{2,k}}{\overline{\sigma}}\right)\right]. \tag{51}$$

Moreover, the results in Section 5 imply that $\hat{\sigma}_1 = \overline{\sigma}\frac{d-1}{4d-1}, \hat{\sigma}_2 = \overline{\sigma}\frac{2d-1}{4d-1}$. Inserting these expressions in equation 51 yields

$$\mathcal{V}_k = N_k \times \frac{\gamma_k}{4} \times \frac{s_{2,k}/s_{1,k}}{4s_{2,k}/s_{1,k}-1}\left[\frac{\rho_{1,k}}{\gamma_k+\rho_{1,k}}+\frac{2\rho_{2,k}}{\gamma_k+\rho_{2,k}}\right].$$

At this point we need to reintroduce the contact rates in order to back out γ_k. Let $\tilde{\rho}_{i,k}$ denote the *stylized values* for speed of venues $i = 1, 2$ in market k. We can then write

$$\tilde{\mathcal{V}}(\gamma_k) = N_k \times \frac{\gamma_k}{4} \times \frac{\tilde{\rho}_{2,k}(\tilde{\rho}_{1,k}+\gamma_k+r)}{4\tilde{\rho}_{2,k}(\tilde{\rho}_{1,k}+\gamma_k+r)-\tilde{\rho}_{1,k}(\tilde{\rho}_{2,k}+\gamma_k+r)}\left[\frac{\tilde{\rho}_{1,k}}{\gamma_k+\tilde{\rho}_{1,k}}+\frac{2\tilde{\rho}_{2,k}}{\gamma_k+\tilde{\rho}_{2,k}}\right]. \tag{52}$$

Equation 52 makes it clear that, given a set of values for $\{r, N_k, \tilde{\rho}_{1,k}, \tilde{\rho}_{2,k}\}$, there is a mapping from γ_k to trading volume $\tilde{\mathcal{V}}$. Given an empirical volume observation $\mathcal{V}_k$, the value of γ_k is found by solving $\tilde{\mathcal{V}}(\gamma_k) = \mathcal{V}_k$.

The next question, of course, is whether the model can predict the correct values of $\tilde{\rho}_{1,k}$ and $\tilde{\rho}_{2,k}$. To calibrate c_k we simply use the first order conditions for speed. In the monopoly case there is only one condition. In the duopoly case we use the fast venue FOC because it is by far the most important in terms of welfare. It is also likely to be more precisely measured in the data. We can then compare the predicted ρ with the stylized $\tilde{\rho}$. This is what we do in Table IV. Notice that ρ_2 and ρ_m are quite similar.

H.2 Comparing Calibration Approaches

Let us compare our calibration approach, PP, to that of DGP (2005; 2007). First and most obvious, there are several parameters that are unique to our model. PP consider heterogeneous agents, adding a distribution for investor types, and endogenize the market structure, thus considering speed cost, entry costs, and other parameters related to the modeling of competing venues and regulations.

One part of the calibration that is common to both papers is that of the rate of preference shocks, γ, which is of course related to trade needs and volume. Volume depends on the number of transaction and on the average transaction size. Trade size is normalized to one both in our paper and in DGP, so it is natural to focus on the *number of trades* per unit of time (a day in our calibration) as opposed to total volume.[61] Consider a standard trading day and a single trading venue. Let ρ denote the daily market contact rate, which can also be interpreted as the bilateral contact rate in DGP. Let $\mathcal{V}$ denote the total number of trades and v the per capita number of trades. Assume there are N traders in the venue. A (steady state) fraction F of them are willing to trade given their holdings, preferences, and the prevailing market price. We can then write

$$\begin{aligned} \mathcal{V} &= \rho \times N \times F(\gamma, \cdot), \\ v &= \frac{\mathcal{V}}{N} = \rho \times F(\gamma, \cdot). \end{aligned} \tag{53}$$

F structurally depends on the rate of preference change γ . Let $\overline{a}$ denote the per capita asset supply, so the total supply is $N\overline{a}$. We can express traders' turnover as

$$\mathcal{T} = \frac{\mathcal{V}}{N\overline{a}} = \frac{v}{\overline{a}} = \frac{\rho F(\gamma)}{\overline{a}}.$$

The calibration strategies can be then summarized as follows:

- DGP start from $\mathcal{T}$ and ρ and then compute

$$\gamma_{DGP} = F_{GDP}^{-1}\left(\frac{\overline{a}\mathcal{T}}{\rho}\right) \tag{54}$$

- PP start from $\mathcal{V}$ and ρ and then compute

$$\gamma_{PP} = F_{PP}^{-1}\left(\frac{\mathcal{V}}{\rho N}\right) \tag{55}$$

In words, the main empirical difference is that DGP consider a stylized trader turnover figure and fix an arbitrary per capita asset supply $\overline{a} \in [0, 1]$ to derive γ. Instead, PP look at the aggregate number of trades for a particular asset and calibrate the number of active institutions to recover a representative institution's γ. We use volume mainly because we find that statistic more readily

[61] See the discussion in Section 8 about order splitting. Needless to say, this simplification will better represent the data when trade size is rather homogeneous.

available across asset classes and trading instruments. A second advantage of using volume is de-emphasizing the role of the ad-hoc parameter $\overline{a}$.

The γ parameter we compute is larger than that in DGP. Let us use the corporate bond market data points to do a numerical comparison. DGP choose asymmetric preference shock rates. Using their notation, in terms of yearly rates, $\lambda_u = 5$ and $\lambda_d = 0.5$, implying that an investor spends an average of 2 years in the high type and 0.2 years as a low type. Since, on average, each investor spends nearly 91% of the time as a high type, the weighted average yearly rate of preference change is $0.91 \times .5 + 0.09 \times 5 \approx 0.91$. The equivalent daily rate is $\gamma_{DGP} = \frac{0.91}{252} = 0.0036$. PP's comparable rate of preference change is $\frac{\gamma_{PP}}{2} = 0.417$ ($\frac{1}{2}$ is the conditional probability of type change given the arrival of a shock). Roughly speaking, there are two order of magnitude difference between the models. This difference chiefly stems from the difference in volume figure. Our TRACE data sample shows that active corporate bonds trade 1.97*252=496.44 times a year. Based on DGP's figures, instead, annual volume is turnover×asset supply=0.5×0.8=0.4 trades a year. Actively traded corporate bonds then trade 1,000 times more than what is predicted in DGP calibration. Despite the fact that a different trading volume will naturally map into a different γ, and different potential gains from trade, the qualitative conclusions of the calibration exercise remain the same. In particular, the main economic interpretations that arise from comparing market outcomes to the solution of the constrained efficient problem are not affected by the specific value of γ.

Appendix I Example of Excess Entry with Three Venues

This appendix analyzes the possibility of excess entry in a market with three venues. Following the notation of Section 7.3, l and h denote the slow and fast incumbents, respectively, and e denotes the entrant. We use the baseline calibration for equities (with $N_k = 92$) in our simulations. We report $\hat{\sigma}_i/\overline{\sigma}$ instead of $\hat{\sigma}_i$ because it is easier to interpret. We also normalize $\Pi_l(\rho_l, \rho_h) = 100$ in the duopoly equilibrium. Welfare is defined relative to the Walrasian outcome, as in Table V.

With two incumbents, the optimal speeds are $\rho_l = 239.13$ and $\rho_h = 23,758.2$, as displayed in Panel II of Table IV. The marginal types are $\hat{\sigma}_l/\overline{\sigma} = 0.125$ and $\hat{\sigma}_h/\overline{\sigma} = 0.417$. Profits are $\Pi_l(\rho_l, \rho_h) = 100$ and $\Pi_h(\rho_l, \rho_h) = 690.69$ and aggregate welfare is $\mathcal{W} = 90.51$ (again, as in Table IV).

Consider now a slow entrant $\rho_e \leq \rho_l$. The entrant optimally chooses $\rho_e = 127.70$ and the marginal types are $\hat{\sigma}_e/\overline{\sigma} = 0.031$, $\hat{\sigma}_l/\overline{\sigma} = 0.146$, and $\hat{\sigma}_h/\overline{\sigma} = 0.458$. There is a significant decrease in participation in the fast venue, as predicted by Lemma 10. The profits are $\Pi_e(\rho_e, \rho_l, \rho_h) = 7.09$, $\Pi_l(\rho_e, \rho_l, \rho_h) = 53.71$, and $\Pi_h(\rho_e, \rho_l, \rho_h) = 594.50$. The profits of the slow incumbent are almost halved by competition from the entrant. Aggregate welfare decreases to $\mathcal{W} = 89.48$ mostly because welfare generated by the fast venue decreases from 81.56 to $\mathcal{W}_h = 77.96$.

The outcome is very different if the entrant has a high speed $\rho_e \geq \rho_l$. In that case, the entrant would optimally choose $\rho_e = 29,319$. The marginal types become $\hat{\sigma}_e/\overline{\sigma} = 0.334$, $\hat{\sigma}_l/\overline{\sigma} = 0.241 \times 10^{-3}$, and $\hat{\sigma}_h/\overline{\sigma} = 0.802 \times 10^{-3}$. The venues generate welfare $\mathcal{W}_l \approx 0$, $\mathcal{W}_h = 10.63$, and $\mathcal{W}_e = 87.75$. Aggregate welfare increases to $\mathcal{W} = 98.38$. We also find (in untabulated results) that welfare

increases when the entrant has an intermediate speed $\rho_l \leq \rho_e \leq \rho_h$.

To summarize our numerical results, we find that entry can reduce welfare when the entrant has a low speed relative to that of the incumbent. The reason is that increased participation by low-σ types is not enough to compensate for the misallocation of high-σ types. On the other hand, when entry takes place at the high end of the speed ladder, we find that it improves welfare.